UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC

CURSO DE CIÊNCIA DA COMPUTAÇÃO

MAICON BASTOS PALHANO

MEDIDAS DE QUALIDADE APLICADAS AOS ALGORITMOS DE CLUSTERIZAÇÃO DA SHELL ORION DATA MINING ENGINE

CRICIÚMA

2013

MAICON BASTOS PALHANO

MEDIDAS DE QUALIDADE APLICADAS AOS ALGORITMOS DE CLUSTERIZAÇÃO DA SHELL ORION DATA MINING ENGINE

Trabalho de Conclusão de Curso, apresentado para obtenção do Grau de Bacharel no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientadora: Profª. MSc. Merisandra Côrtes de Mattos Garcia

CRICIÚMA

2013

MAICON BASTOS PALHANO

MEDIDAS DE QUALIDADE APLICADAS AOS ALGORITMOS DE CLUSTERIZAÇÃO DA SHELL ORION DATA MINING ENGINE

Trabalho de Conclusão de Curso, apresentado para obtenção do grau de Bacharel no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Criciúma, 27 de novembro de 2013.

BANCA EXAMINADORA

Profa. Merisandra Côrtes de Mattos Garcia - MSc. - (UNESC) - Orientador

Prof. Kristian Madeira –MSc. - (UNESC)

Prof. Cristian Cechinel – Dr. - (UFPEL)

A minha querida família e amigos por todo o apoio e consideração.

“Não importa o quão bom seja o seu índice, há um conjunto de dados lá fora esperando para enganar ele (e você).”

Nikhil R. Pal e James C. Bezdek

RESUMO

Nos últimos anos cada vez mais se torna visível o tanto de informações que são armazenadas diariamente em bases de dados, porém quanto maior a quantidade de dados, maiores são as dificuldades para analisar esses dados. O *data mining* surge para analisar e processar essas grandes quantidades de dados armazenadas em bases de dados de forma automatizada. Tendo como principal objetivo a busca por informações relevantes e implícitas em meio aos dados, o que justifica o nome mineração de dados, pois traz só o que é mais relevante ao usuário. O *data mining* é composto por tarefas e métodos, onde uma dessas tarefas é a de clusterização que gera *clusters* de dados, por meio da semelhança entre elementos. Porém como a clusterização é um processo não-supervisionado, onde o particionamento dos dados é definido pelo algoritmo, devem ser aplicadas métricas de avaliação denominadas neste trabalho por medidas de qualidade na clusterização para avaliar e qualificar as clusterizações obtidas pelos algoritmos. Como ferramenta para a análise de algoritmos e medidas de qualidade optou-se pela Shell Orion Data Mining Engine, que é uma ferramenta de data mining que vem sendo desenvolvida pelo curso de Ciência da Computação da UNESC. No trabalho foram aplicados sete algoritmos de clusterização da Shell Orion em quatro bases de dados diferentes e avaliaram-se os resultados dos particionamentos por meio de cinco medidas de qualidade sendo estas: Coeficiente de Partição, Coeficiente de Partição Entrópica, Xie-Beni, Dunn e C-Index. Foi realizada também uma revisão de medidas existentes e da aplicabilidade das medidas de qualidade na literatura e compararam-se os resultados dos algoritmos da Shell Orion e das bases de dados com o de outros autores da área de medidas de qualidade em clusterização. Ao término da aplicação das medidas de qualidade foi possível observar que alguns algoritmos não conseguiram particionar corretamente algumas bases de dados, o que inviabilizou a coleta dos valores de algumas medidas, geralmente em bases de dados maiores como no caso da *Wine* e das bacias hidrográficas. Apesar disso, os resultados obtidos pelas medidas de qualidade foram bem próximos dos resultados alcançados por outros autores em outros trabalhos, o que mostra o funcionamento correto da maioria dos algoritmos e das cinco medidas empregadas.

Palavras-chave: Data Mining. Clusterização. Medidas de Qualidade. Shell Orion.

ABSTRACT

In latest years becomes increasingly visible the quantity of informations which are stored daily in databases, however, the higher the amount of data, bigger the difficulty to analyze these data, the data mining arises for analyze and process these large amounts of data stored in databases an automated form. Having as main objective the search for relevant informations and implied in middle to data, that justified the name Data Mining, it brings only what is most relevant to the user. The Data Mining is composed by tasks and methods, where one of these tasks is the clustering that generating clusters of data by means of the similarity between elements. However, as clustering is an unsupervised process, where the partitioning of data is defined by the algorithm, should be applied metric evaluation denominated in this work by quality measures in order to evaluate and qualify clusterings obtained by clustering algorithms. As a tool for the analyse the algorithms and measures of quality, it was decided by Shell Orion Data Mining Engine, which is a tool of data mining, that has been developed by the course of Computer Science of UNESC. In this work were applied seven algorithms of clustering of Shell Orion in four different databases and valued the results of partitioning by five quality measures, these being: Partition Coefficient, Entropy Partition Coefficient, Xie-Beni, Dunn e C-Index. Was realized also a review of existing measures and the applicability of the quality measures in the literature and comparing the results of the algorithms of Shell Orion Data Mining and databases with other authors in the area of quality measures in clustering. To end of the application of quality measures was observed that some algorithms failed to correctly partition some databases, it was not possible to collect the values of some measures, generally in larger databases such as the case of Wine and watersheds. Nevertheless, the results obtained by the measures of quality were very close to the results obtained by other authors in other work, which shows the correct functioning of most of the algorithms and the five measures employed.

Palavras-chave: Data Mining. Clustering. Quality Measures. Shell Orion.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

A²CA	Adaptative Ant-Clustering Algorithm
ABC	Ant Based Clustering
ACM	Association for Computing Machinery
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CE	Coeficiente Partição Entrópica
CP	Coeficiente de Partição
DB	Davies-Bouldin
DBSCAN	Density-based Spatial Clustering of Applications With Noise
DCBD	Descoberta de Conhecimento em Bases de Dados
EFLD	Extended Fisher Linear Descriminant
FLD	Fisher Linear Descriminant
FS	Fukuyama-Sugeno
ICC	Inter Class Contrast
IEEE	Institute of Eletrical and Eletronics Engineers
JDBC	Java Database Connectivity
KDD	Knowledge Discovery in Database
PB	Point Bi-Serial
RBF	Rasion Basis Function Network
RCP	Robust C-Prototypes
SACA	Standard Ant Clustering Algorithm
SD	Separação e Dispersão
SGBD	Sistemas Gerenciadores de Banco de Dados
UFRJ	Universidade Federal do Rio de Janeiro
URCP	Unsupervised Robust C-Prototypes
VRC	Variance Ratio Criterion
XB	Xie-Beni

SUMÁRIO

1 INTRODUÇÃO

Diariamente, grandes quantidades de dados são coletadas e armazenadas em bancos de dados, aumentando a necessidade por métodos eficientes e efetivos de análise para fazer uso da informação contida implicitamente nos dados (ANKERST et al, 1999, tradução nossa). Desta forma, para satisfazer a necessidade da descoberta de conhecimento útil nessas bases de dados, em 1996, por Usama Fayyad e seus colaboradores, originou-se o conceito de *data mining*, que é definido como um processo que busca encontrar padrões em um conjunto de dados, sendo que esses padrões devem ser úteis, válidos e compreensíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

O *data mining* consiste na aplicação de diferentes conhecimentos, como de Inteligência Computacional, Banco de dados, Estatística, Aprendizado de Máquina e Reconhecimento de Padrões. Essa técnica constitui-se na principal etapa do processo de Descoberta de Conhecimento em Bases de Dados (DCBD), sendo responsável pela identificação das relações entre os padrões.

Na etapa de *data mining* empregam-se tarefas e métodos com características distintas, que são aplicadas em bases de dados, conforme o objetivo da descoberta de conhecimento. Dentre as tarefas tem-se a classificação, associação, sumarização, regressão, previsão de séries temporais e clusterização (GOLDSCHMIDT; PASSOS, 2005).

A clusterização, segundo Han e Kamber (2006, tradução nossa) reúne um conjunto de dados em grupos de objetos similares, formando *clusters*. Na clusterização busca-se maximizar a semelhança entre objetos do mesmo *cluster* e minimizar a semelhança entre *clusters* (LAROSE, 2005, tradução nossa).

Clusterização é uma tarefa onde não há grupos pré-definidos e exemplos que mostrem se os *clusters* encontrados pelos algoritmos são totalmente válidos (HALKIDI et al, 2002, tradução nossa). Assim, para validar os resultados dos algoritmos de *data mining*, é necessário aplicar alguns critérios. Além disso, em muitos algoritmos de clusterização o usuário tem de informar o número de *clusters* que serão encontrados no conjunto de dados. Esta tarefa é complexa e sujeita a erros, sendo necessário ter um conhecimento prévio e específico da base utilizada. Mediante isso, a fim de se identificar a qualidade dos resultados gerados pela

clusterização deve-se empregar medidas de validação voltadas a essa tarefa de *data mining* (GAN; MA; WU, 2007, tradução nossa).

Na aplicação de *data mining* empregam-se ferramentas computacionais, denominados *shells*, as quais em sua maioria são comerciais, tendo-se algumas iniciativas gratuitas, como por exemplo, a *Shell Orion Data Mining Engine*. Esta ferramenta encontra-se em desenvolvimento no Grupo de Pesquisa em Inteligência Computacional Aplicada da UNESC.

Dentre as tarefas de *data mining* disponibilizadas na *Shell Orion*, tem-se a clusterização pelos algoritmos *K-Means*, *Kohonen*, *Gustafson–Kessel*, *Gath-Geva*, *Robust C-Prototypes* (RCP)*, Unsupervised Robust C-Prototypes* (URCP), *Fuzzy C-Means, Density-based Spatial Clustering of Applications With Noise* (DBSCAN), *Ant Based Clustering* (ABC), *Adaptative Ant-Clustering Algorithm* (A²CA) e *Standard Ant Clustering Algorithm* (SACA). Porém, muitos desses algoritmos não foram validados com mais de uma base de dados, observando-se os resultados do particionamento por meio de medidas de qualidade.

Conforme Sommerville (2003) validar é o processo de verificação e análise capaz de garantir que o software cumpra com suas especificações. No caso dos algoritmos de clusterização realiza-se a validação dos *clusters* identificados.

Esta pesquisa enfatiza uma revisão de literatura na área de medidas de qualidade, destacando-se aquelas aplicadas na clusterização, bem como a análise dos resultados gerados pelas medidas Coeficiente de Partição, Coeficiente de Partição Entrópica, Xie-Beni, Dunn e C-Index nos algoritmos FCM, RCP, URCP, DBSCAN, SACA, A²CA, ABC da tarefa de clusterização da Shell Orion, avaliando os resultados das suas partições, empregando-se quatro bases de dados.

1.1 OBJETIVO GERAL

Analisar medidas de qualidade aplicáveis a algoritmos da tarefa de clusterização em *data mining*.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos desta pesquisa consistem em:

a) compreender o conceito de *data mining,* clusterização*,* medidas de qualidade em *data mining* e técnicas de validação;
b) compreender o método de lógica *fuzzy*, densidade e colônias de formiga;
c) aplicar medidas de qualidade para validação dos resultados de alguns algoritmos de clusterização da *Shell Orion Data Mining Engine*;
d) analisar os resultados dos testes e das medidas empregadas.

1.3 JUSTIFICATIVA

Data mining, segundo Tan, Steinbach e Kumar (2009), é um processo relevante que busca descobrir de forma automática informações úteis em grandes bases de dados, onde, por meio de suas tarefas e métodos procura encontrar padrões válidos. Padrões estes, que de acordo com Goldschmidt e Passos (2005), poderiam ser ignorados pelas outras formas de análise de dados. Isso se consegue pela aplicação de algoritmos específicos que mesmo sob algumas limitações aceitáveis de eficiência computacional, tanto de tempo quanto de processamento, identificam uma enumeração particular de padrões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

O processo de *data mining* possui tarefas distintas, onde uma delas é a de clusterização, que consegue identificar padrões ou tendências, apenas observando os dados, sem definir rótulos, como na classificação (HAN; KAMBER, 2006, tradução nossa).

As metodologias de clusterização têm sido largamente utilizadas em numerosas aplicações, incluindo reconhecimento de padrões, análise de dados, processamento de imagens e pesquisa de mercado (CARLANTONIO, 2001). Ainda, Goldschmidt e Passos (2005) e Serra (2002) trazem, que os resultados provenientes dessa tarefa podem ser utilizados como etapas de pré-processamento para a aplicação de outras tarefas, tais como a classificação e sumarização.

Como forma de auxílio na determinação da qualidade dos *clusters* encontrados pelos algoritmos de clusterização, o resultado do agrupamento deve ser validado, verificando-se a solução encontrada, sendo que essa tarefa é feita geralmente por meio de medidas de qualidade, que consistem em índices

estatísticos (FERNANDÈZ; CORBACHO, 2010, tradução nossa; JAIN; DUBES, 1988, tradução nossa).

Os algoritmos devem separar corretamente os *clusters*, porém muitas vezes na clusterização tem-se um resultado desagradável, na melhor das hipóteses e errado na pior delas, ilustrando assim a necessidade de abordagens do problema da validação de *clusters* (BEZDEK et al, 2005, tradução nossa).

Dentre os fatores que levaram a escolha da *Shell Orion Data Mining Engine*, está o fato de ser uma ferramenta gratuita e por conter diferentes tarefas e métodos, dentre essas tarefas, a de clusterização. Além disso, na *Shell Orion,* durante o desenvolvimento dos seus módulos, alguns algoritmos de clusterização não foram submetidos a mais de uma base de dados, onde não se sabe se os mesmos geram *clusters* com exatidão, no qual é preciso avaliar a qualidade dos *clusters* por meio de medidas de qualidade.

1.4 ESTRUTURA DO TRABALHO

Este trabalho é constituído por 5 capítulos, sendo o Capítulo 1 destinado a introdução, que traz uma breve descrição do trabalho e a justificativa da pesquisa.

No Capítulo 2 se tem os conceitos do processo de descoberta de conhecimento em bases de dados, as etapas que o compõem, em especial a de *data mining*, bem como suas principais tarefas e métodos. A tarefa de clusterização também é abordada, bem como as principais ferramentas de data mining existentes e a Shell Orion que foi utilizada na pesquisa.

Os conceitos de medidas de qualidade na clusterização, onde, traz-se uma pesquisa das principais medidas de qualidade encontradas na literatura, são apresentadas no Capítulo 3.

No Capítulo 4 é possível identificar alguns trabalhos correlatos na área de medidas de qualidade em clusterização, tanto nacionais como internacionais.

O trabalho desenvolvido, com a metodologia adotada e os resultados obtidos são expostos no Capítulo 5. Finalizando, tem-se a conclusão e as possibilidades de trabalhos futuros.

2 PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

A evolução da Tecnologia da Informação com relação à coleta, armazenamento, transmissão de dados, aumento da velocidade de comunicação e custos acessíveis a estas tecnologias, possibilitou que organizações pudessem armazenar e processar seus dados de forma digital, onde, estes dados geralmente eram analisados e interpretados por especialistas de forma manual por meio de Sistemas Gerenciadores de Banco de Dados (SGBD) e planilhas eletrônicas, porém com o aumento dessas bases de dados era visível a demora e o custo para aplicação desse tipo de análise (MELANDA, 2005).

O processo *Knowledge Discovery in Database* (KDD) ou Descoberta de Conhecimento em Bases de Dados (DCBD) surgiu em resposta a essa necessidade, de novos métodos capazes de descobrir informações relevantes e de forma automática em grandes conjuntos de dados (BORGES, 2010; ROMÃO 2002).

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996, tradução nossa) o DCBD é um processo não trivial para a identificação de padrões em grandes quantidades de dados, padrões estes devendo ser válidos, novos e úteis de forma à serem humanamente compreensíveis.

Silva (2002) enfatiza os termos usados por Fayyad na definição do processo de DCBD:

a) **processo**: no DCBD processo refere-se a uma sequência de vários passos envolvendo preparação dos dados, busca por padrões, avaliação do conhecimento e refinação envolvendo iteração e modificação;
b) **padrão**: representa uma expressão E em uma linguagem L que descrevem fatos individuais em um subconjunto FE de F. É considerado um padrão E se este for mais simples do que a enumeração de todos os fatos no subconjunto FE. Por exemplo, o padrão: *"se renda < r então a pessoa não recebe financiamento"* seria aplicável para uma escolha de *r*;
c) **válidos**: os padrões encontrados devem ser válidos em novos dados com algum grau de certeza. Uma medida de certeza é uma função *C* que mapeia expressões em *L* para um espaço de medidas *MC*. Por

exemplo, se um limite de padrão de crédito é ampliado, então a medida de certeza diminuiria, uma vez que mais financiamentos seriam concedidos a um grupo até então restrito a esta operação;

d) **novos**: refere-se a um conhecimento implícito nos dados, podendo ser medida por uma função *N(E,F),* onde, pode ser uma função booleana ou uma medida que expresse o grau de “novidade” ou “surpresa”. O primeiro exemplo seria de um fato que não é novo: sejam *E* = “usa tênis” e *F* = “alunos de colégio” então *N(E,F)* = 0 ou *N(E,F) = false.* Já no próximo exemplo, seria o de um fato novo e desconhecido, onde: *E* = “bom pagador” e *F* = “trabalhador da construção civil” então *N (E,F)* = 0,85 ou *N (E,F)= true;*

e) **potencialmente úteis**: os padrões encontrados no processo de DCBD devem de alguma forma, levar a alguma atitude prática. Como por exemplo, regras obtidas no processo, podem ser usadas para aumentar o retorno financeiro de uma instituição;

f) **compreensíveis**: um dos objetivos do DCBD é a representação de padrões de forma compreensível aos humanos, de modo que aumente o entendimento dos próprios dados.

Conforme Martins (2012), o principal objetivo do processo de DCBD é fazer com que dados brutos passem a identificar conhecimentos úteis em bases de dados.

Uma característica importante na DCBD é a variedade de atividades em sua aplicação, pois se inicia pelo domínio de aplicação, posteriormente a preparação e análise dos dados, avaliação, entendimento até a aplicação dos resultados gerados (APPEL, 2010).

O processo de DCBD é composto de diversas etapas, sendo estas interseccionadas, de modo que o resultado de uma etapa pode ser usado para melhorar os resultados de etapas subsequentes (SASSI, 2006). Sendo as principais etapas (figura 1):

a) **seleção:** também denominada coleta de dados, é a fase onde se identifica dentre os atributos das bases de dados, quais devem ser considerados na análise (GOLDSCHMIDT; PASSOS, 2005). A etapa de seleção pode ser julgada como trabalhosa, pois alguns problemas

podem surgir durante o processo, como por exemplo: questões legais e éticas, motivos estratégicos, razões políticas, formato arbitrário dos dados, conectividade, banco de dados ou aplicações obsoletas (BATISTA, 2003). Uma das vantagens da etapa de seleção é o ganho no tempo de processamento, pois diminui o espaço de busca, pelo fato de analisar apenas um subconjunto de atributos (SASSI, 2006);

b) **pré-processamento:** embora seja possível minerar os dados brutos armazenados nas bases de dados, é de suma importância que se efetue a tarefa de pré-processamento dos dados (WINCK, 2012). Quando há dados precários, consequentemente o resultado de qualquer tarefa tende a ser de acordo com a qualidade dos mesmos (ROMÃO, 2002). Na etapa de pré-processamento busca-se melhorar a qualidade dos dados e consequentemente aumentar a acurácia e eficiência do processo de *data mining* (APPEL, 2010). Para garantir esta qualidade aos dados, faz-se a limpeza de valores nulos e redundantes presentes nos conjuntos de dados, verificando a consistência das informações, removendo também dados duplicados e corrompidos, com isso melhora-se o processamento do algoritmo (SASSI, 2006);

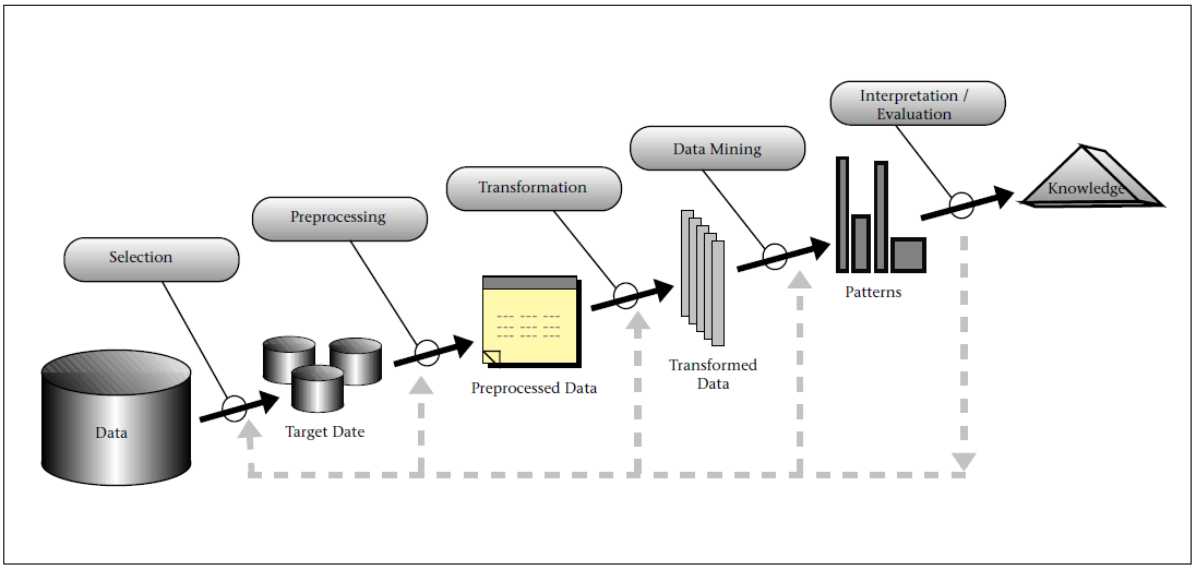
c) **transformação:** uma vez pré-processados, os dados devem ser transformados de acordo com os algoritmos de *data mining* à serem usados, tendo em vista a limitação que cada algoritmo tem com relação aos tipos de dados aceitáveis. Por exemplo, um algoritmo que aceita somente atributos numéricos, se os dados da base não estiverem de acordo, estes devem ser transformados para numéricos. Outro objetivo da transformação dos dados é a sua padronização em intervalos de valores, como por exemplo, criando-se categorias de idade (0-18,19-23, 24-29, 30-35) ao invés de apenas idades. Dentre as vantagens da transformação, tem-se a melhora do tempo de processamento e no entendimento do conhecimento descoberto, podendo facilitar a tomada de decisão do algoritmo (SASSI, 2006);

d) ***data mining*:** é a fase onde ocorre a busca por padrões nos conjuntos de dados, por meio de algoritmos específicos de *data mining*, sendo

ela efetivamente a responsável pela descoberta do conhecimento (GOLDSCHMIDT; PASSOS, 2005);

e) **interpretação dos resultados:** a última etapa do processo de DCBD também chamada de pós-processamento é onde os especialistas da DCBD e do domínio de aplicação analisam, interpretam e avaliam os resultados gerados pelo *data mining* e definem novas alternativas de investigação aos dados (GOLDSCHMIDT; PASSOS, 2005). Segundo Sassi (2006) o objetivo principal desta etapa é a melhora na compreensão dos conhecimentos encontrados e a validação dos mesmos por meio de medidas de qualidade.

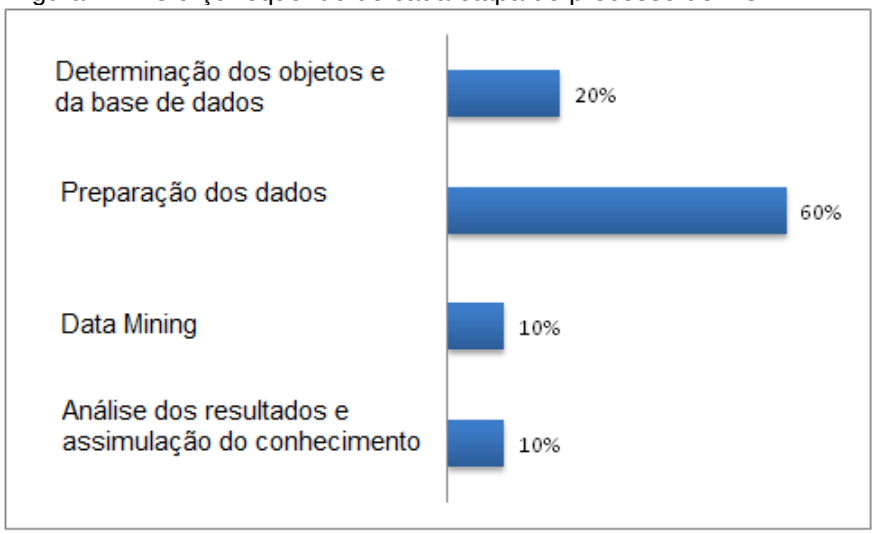
Figura 1 - Etapas do processo de DCBD



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smith (1996).

Dentre todas as etapas da DCBD, a de preparação dos dados (seleção, pré-processamento e transformação) é a que exige mais esforços do processo, onde, pela figura 2 é possível visualizar a grande diferença do tempo gasto pela etapa de preparação com relação as demais (KANTARDZIC, 2011, tradução nossa; PYLE, 1999).

Figura 2 – Esforço requerido de cada etapa do processo de DCBD.



Fonte: Adaptado de Cabena et al (1997) e Kantardzic (2011).

O processo de DCBD está diretamente ligado ao *data mining*, muitas vezes sendo considerados como sinônimos, porém Fayyad, Piatetsky-Shapiro e Smyth (1996, tradução nossa) os diferencia referenciando o *data mining* como sendo o modo pelo qual os padrões são extraídos e enumerados a partir dos dados, enquanto o DCBD é o processo de avaliação e interpretação desses padrões, separando os válidos dos inúteis, ou ainda, dos já conhecidos. Outro fator que colaborou para que fossem tidos como sinônimos é o *data mining* ser considerado a principal etapa do processo de DCBD.

2.1 DATA MINING

Segundo Fayyad, Piatetsky-Shapiro e Smith (1996, tradução nossa) o crescimento dos bancos de dados empresariais, governamentais e científicos acarretou a incapacidade humana de interpretar e analisar essas informações, de modo que fossem necessários novos métodos e ferramentas capazes de analisar e extrair informações úteis a partir desses dados armazenados.

Sendo assim, de acordo com Sassi (2006) o *data mining* explora repositórios de dados à procura de padrões e relacionamentos implícitos, encontrando dados que possam prever tendências ou comportamentos futuros.

Na aplicação do *data mining* o usuário precisa ter de forma clara quais os objetivos que deseja alcançar e que problemas pretende resolver, pelo fato de ter

que escolher um algoritmo que atenda as necessidades do problema a ser resolvido e alcance os objetivos propostos (SILVA, 2007).

Data mining é considerado um campo novo e interdisciplinar, tendo origem em áreas como: sistemas gerenciadores de banco de dados, *data warehouses*, estatística, aprendizagem de máquina e visualização de dados, contando com o auxílio também de áreas como: redes neurais, reconhecimento de padrões, análise de dados espaciais, bases de dados de imagem e processamento de sinais (HAN; KAMBER, 2006, tradução nossa).

O *data mining* tem sido aplicado em diversas áreas, com a finalidade, tanto de descrever características do passado, quanto predizer tendências futuras, dentre essas, destacam-se áreas como: finanças, energia, saúde, recursos hídricos, genética e biologia (DINIZ, 2009).

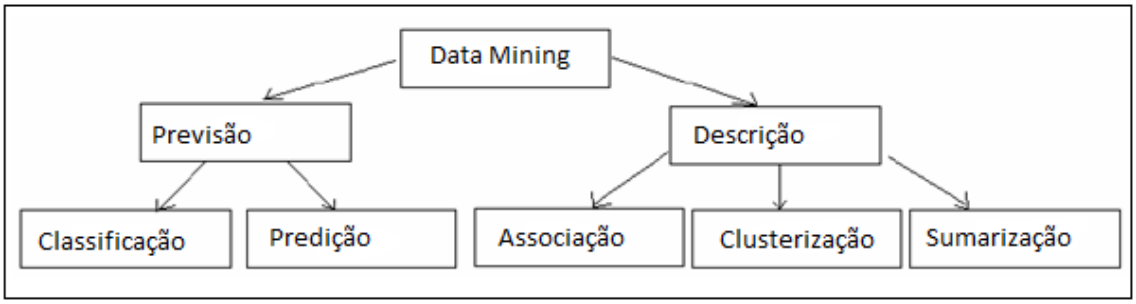
Witten, Frank e Hall (2011, tradução nossa) destacam ainda outros campos de pesquisa distintos, onde o *data mining* é aplicado, como: mineração de dados na web, segmentação de dados de satélite, previsão de carga do sistema elétrico, *marketing* e vendas. No comércio e indústria vem sendo aplicado à vendas de supermercados, produção industrial e *marketing*; em operações bancárias, financeiras e seguradoras; pelo governo em estatísticas de saúde e levantamentos demográficos; na tecnologia em *logs* de servidores, sistemas e simuladores e na descoberta científica em áreas como genética, astronomia, geologia e medicina (SILVA, 2002).

A etapa de *data mining* é composta por tarefas e métodos, ambos possuindo suas particularidades, onde a escolha adequada de cada um irá depender do conhecimento acerca do domínio de aplicação do conjunto de dados utilizado (KANTARDZIC, 2011).

2.1.1 Tarefas e métodos de *data mining*

A técnica de *data mining* é capaz de realizar tarefas preditivas e descritivas (figura 3), onde as preditivas têm como principal objetivo a predição do valor de um atributo a partir de valores de outros atributos e as descritivas consistem na análise dos dados em busca de padrões (correlações, tendências e anomalias) (GONZALES, 2012).

Figura 3 - Tarefas preditivas e descritivas de *data mining*



Fonte: Adaptado de Rezende (2005).

Dentre as principais tarefas de *data mining* destacam-se:

a) **tarefas preditivas/supervisionadas:**

- **classificação:** é considerada a tarefa mais importante e popular, basicamente, consiste em criar classes pré-definidas de acordo com semelhanças de algumas características, determinando um conjunto de classes ou padrões que podem ser usados para classificar outros padrões (PASTA, 2011). De forma geral, a tarefa de classificação busca associar cada registro da base de dados a um rótulo categórico, também chamado de classe (RABELO, 2007),
- **regressão/estimativa:** a regressão, assim como na classificação busca funções que associem registros a rótulos (PIRES, 2011). Podendo ser aplicada, por exemplo, para prever a temperatura em um determinado dia ou local (COSTA, 2012);

b) **tarefas descritivas/não-supervisionadas:**

- **sumarização:** consiste em identificar e mostrar de forma clara e compreensível características dos dados, podendo ser aplicado, por exemplo, na identificação de características dos assinantes de uma revista que moram na região sudeste do Brasil, descobrindo-se que possuem faixa salarial de X reais, nível superior completo e residência própria (GOLDSHMIDT; PASSOS, 2005),
- **associação:** tem como objetivo principal a busca por transações de itens que ocorram frequentemente de forma simultânea em uma base de dados (PIRES, 2011). Para tal tarefa, utilizam-se algoritmos de regras de associação, que após minerar itens frequentes, buscam por

relacionamentos recorrentes em um mesmo conjunto de dados (WINCK, 2012),

- **clusterização:** a tarefa de clusterização, a qual será utilizada neste trabalho, de forma geral é a identificação de classes sem algum tipo de pré-definição, sendo denominadas de *clusters*, baseando-se na similaridade dos dados, onde seus resultados podem servir de base para outras tarefas.

Dias (2001), esclarece as principais tarefas de *data mining*, descrevendo-as e exemplificando-as, como pode ser visto na tabela 1.

Tabela 1 - Tarefas de *data* mining

Tarefas	Descrição	Exemplos
Classificação	Constrói um modelo que de alguma forma possa ser aplicado à dados não classificados a fim de categorizá-los	• Escolher limite de crédito • Esclarecer pedidos de seguros fraudulentos
Estimativa/Regressão	Usada para definir um valor a alguma variável contínua desconhecida	• Estimar o número de filhos ou a renda total de uma família • Estimar o valor em tempo de vida de um cliente • Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos • Prever a demanda de um consumidor para um novo produto
Associação	Usada para determinar quais itens tendem a co-ocorrerem em uma mesma transação	• Determinar associações, como quais produtos costumam ser colocados juntos em um carrinho de supermercado
Clusterização	Processo de uma partição de uma população heterogênea em vários subgrupos, ou grupos mais homogêneos	• Agrupar clientes por região do país • Agrupar clientes com comportamento de compras similar • Agrupar seções de usuários Web para prever o comportamento futuro do usuário
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados	• Tabular o significado e desvio padrão para todos os itens de dados

Fonte: Adaptado de Dias (2001).

Assim como há diversas tarefas de *data mining*, existem vários métodos

disponíveis para dar suporte à implementação dessas tarefas, onde a escolha do método mais apropriado depende essencialmente das necessidades e dos resultados que se deseja atingir.

A seguir, são descritos de forma sucinta alguns dos métodos de *data mining*:

a) **redes neurais:** assemelhando-se ao cérebro humano, uma rede neural é um processador constituído de unidade de processamento simples, capaz de armazenar e disponibilizar o conhecimento experimental, por meio de um processo de aprendizagem, denominados de neurônios, responsáveis pela conexão e armazenagem do conhecimento adquirido (HAYKIN, 2001);

b) **lógica *fuzzy*:** pode ser vista como uma extensão da teoria clássica de conjuntos (0 ou 1), pois incorpora graus de pertinência em relação a um conjunto (REZENDE, 2005). A lógica *fuzzy* permite uma modelagem do modo aproximado de raciocínio, no qual se parece muito com a habilidade humana em tomar decisões em ambientes imprecisos (GOLDSCHMIDT; PASSOS, 2005);

c) **indução de árvores de decisão:** uma árvore de decisão é aquela onde cada nó não terminal representa um teste ou decisão no item de dado considerado (GOEBEL; GRUENWALD, 1999, tradução nossa). Sendo um método apropriado à tarefas de classificação e regressão (DIAS, 2001);

d) **descoberta de regras de associação:** estabelece uma correlação estatística entre a ocorrência de certos atributos em um conjunto de dados (GOEBEL; GRUENWALD, 1999, tradução nossa);

e) **algoritmos genéticos:** são modelos computacionais inspirados em princípios da reprodução genética, para buscar e otimizar soluções de problemas complexos (LINDEN, 2006).

É interessante observar que cada método ou algoritmo de *data mining* adapta-se melhor em alguns problemas do que em outros, impossibilitando que exista um método universal e que seja o melhor. Geralmente, para cada problema particular, há um algoritmo que se adapta melhor. Com relação a escolha das tarefas e métodos pode-se concluir que, quanto mais experiência e intuição tem o analista,

maior é a chance de sucesso da tarefa ou método escolhido (DINIZ; LOUZADA NETO, 2000).

2.2 TAREFA DE CLUSTERIZAÇÃO

Muitas denominações podem ser encontradas na literatura para Clusterização de dados, como *Data Clustering*, Taxonomia Numérica, Análise de *Cluster,* Agrupamento e Análise de Aglomerados, sendo, Clusterização a denominação adotada neste trabalho.

A tarefa de clusterização é definida em alguns trabalhos como:

a) clusterização, segundo Han e Kamber (2006, tradução nossa), é o processo de agrupar um conjunto de dados físicos ou abstratos dentro de uma classe de objetos similares;

b) Mary e Kumar (2012, tradução nossa) referem como a divisão de dados dentro de grupos de objetos similares, representando os dados por grupos, ou *clusters*, sendo a procura por estes, uma aprendizagem não supervisionada e o sistema resultante representa um conceito de dados;

c) conforme Velmurugan e Santhanam (2010, tradução nossa) o processo de organizar objetos em grupos, cujos membros são similares de alguma maneira, sendo que, um *cluster* é consequentemente uma coleção de objetos no qual são similares entre eles e dissimilares a objetos pertencentes a outros *clusters*;

d) Jain, Murty e Flynn (1999, tradução nossa) definem clusterização como a organização de uma coleção de padrões (geralmente representado como um vetor de medidas, ou um ponto em um espaço multidimensional) em *clusters*, baseado na similaridade presente nos dados;

e) a clusterização é uma tarefa descritiva que procura identificar um conjunto finito de *clusters* a partir dos dados (REZENDE, 2005);

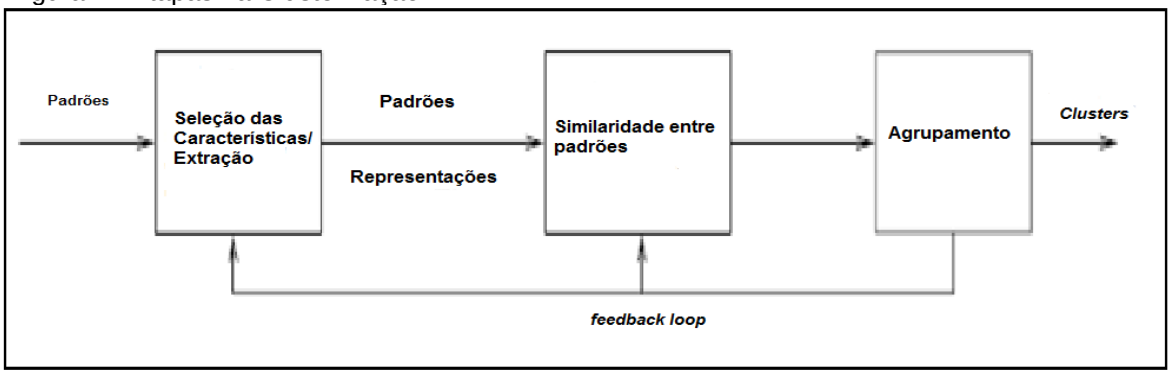
f) segundo Naldi (2011), a clusterização se destaca em relação às demais tarefas, quando aplicada a problemas que estejam em estados iniciais, ou que se tenha pouco conhecimento acerca dos dados.

De forma geral, a maioria dos autores traz definições muito parecidas com relação à clusterização, onde geralmente, sustentam, como sendo a busca por *clusters* em uma base de dados de forma automática, por meio da similaridade entre os dados, sendo uma tarefa não-supervisionada.

A importância da clusterização se dá pelo fato da possibilidade de ser uma tarefa primária no *data mining*, podendo-se utilizar os resultados para a aplicação de outras tarefas, como a classificação, a fim de classificar cada *cluster* gerado em determinadas classes.

Tipicamente, a clusterização é composta de algumas etapas (figura 4) (JAIN; MURTY; FLYNN, 1999, tradução nossa):

Figura 4 - Etapas na Clusterização



Fonte: Adaptado de Jain, Murty e Flynn (1999).

A representação de objetos refere-se as informações disponíveis para o algoritmo de clusterização, a seleção de características é o processo que seleciona e transforma atributos originais que melhor representam o conjunto de dados e a similaridade entre os objetos por meio de funções de similaridade ou distância (JAIN; MURTY; FLYNN, 1999, tradução nossa).

2.2.1 Requisitos recomendáveis em métodos de clusterização

Atualmente existem muitos métodos disponíveis para a clusterização, porém não existe um método, ou algoritmo totalmente perfeito e que consiga atender todos os requisitos de um método ideal (LAURO, 2008).

Segundo Han e Kamber (2006, tradução nossa), um método de clusterização para ser considerado ideal, deve atender alguns requisitos, sendo estes:

a) **escalabilidade:** muitos algoritmos de clusterização trabalham muito

bem com bases de dados pequenas com menos de 200 dados, porém existem bases de dados com milhões de objetos. Nesses casos se utilizam apenas amostras das bases de dados, os resultados podem ser tendenciosos, por isso ressalta-se a importância da escalabilidade em algoritmos de clusterização, onde, os mesmos devem trabalhar tanto com pequenas, quanto com grandes bases de dados;

b) **habilidade para tratar diferentes tipos de atributos:** alguns algoritmos são projetados para clusterizar apenas dados numéricos (baseados na distância), porém aplicações podem requerer clusterizar outros tipos de dados, tais como binários, categóricos (nominais) e ordinais, ou ainda a mistura desses tipos de dados;

c) **descoberta de *clusters* com formatos arbitrários:** vários algoritmos de clusterização determinam *clusters* baseados nas medidas de distância *Manhattan*[1]ou Euclidiana[2]. Algoritmos baseados em tais distâncias tendem a encontrar *clusters* esféricos com tamanhos e densidades similares, porém um *cluster* pode ser de qualquer formato. Assim, é importante desenvolver algoritmos que possam detectar *clusters* de formatos arbitrários;

d) **redução dos parâmetros de entrada:** muitos algoritmos de clusterização requerem que os usuários entrem com parâmetros iniciais. Os resultados da clusterização podem ser sensíveis aos parâmetros de entrada, estes sendo muitas vezes difícil de determinar, especialmente quando a base de dados contém objetos com alta dimensionalidade. Isto, além de se tornar custoso ao usuário, também pode dificultar o controle da qualidade da clusterização;

e) **habilidade em lidar com dados ruidosos:** a maioria das bases de dados reais possuem ruídos, dados faltantes, desconhecidos ou errôneos e alguns algoritmos de clusterização são sensíveis a estes dados, onde, estes *outliers* podem afetar a qualidade dos *clusters*;

f) **clusterização independente da ordem dos registros:** alguns

[1] A função de distância *Manhattan*, ou *City-Block*, representa o somatório do módulo das diferenças dos valores dos atributos entre dois pontos (KAUFMAN; ROUSSEEAW, 1990, tradução nossa).

[2] A distancia euclidiana é definida como aquela entre pares de indivíduos em um plano cartesiano (CLIFFORD; STEPHENSON, 1975, tradução nossa).

algoritmos são sensíveis a ordem em que os dados são inseridos, onde, se mudasse a ordem dos dados, afetaria a clusterização. Portanto, é importante desenvolver algoritmos de clusterização que sejam imparciais nos resultados, quanto à ordem em que os dados estão apresentados;

g) **capacidade de tradar dados com alta dimensionalidade:** uma base de dados pode conter muitas dimensões ou atributos. Muitos algoritmos de clusterização são bons em tratar dados com poucos atributos (um ou dois), porém falham em dados com alta dimensionalidade, recomenda-se que os algoritmos trabalhem também com esses dados;

h) **clusterização baseada em restrições:** há momentos no mundo real em que se necessita instituir limites na clusterização, como por exemplo a separação de clientes por região, para tal, seria uma tarefa desafiante encontrar *clusters* de dados que satisfizessem estas limitações impostas pelo usuário;

i) **interpretação e usabilidade:** usuários esperam que os resultados de clusterizações sejam interpretáveis, compreensíveis e usáveis. Devendo os resultados da clusterização serem associados as suas aplicações e interpretações semânticas, deste modo, ressalta-se a importância de estudar como um resultado pode influenciar os métodos e características da clusterização.

2.2.2 Métodos de clusterização

A escolha do método de clusterização depende do tipo de dados que se irá trabalhar e os resultados ou propósitos que se deseja alcançar, considerando isso, os métodos comumente utilizados na clusterização podem ser categorizados em (LAURO, 2008):

a) **métodos de particionamento:** têm por finalidade dividir uma base de dados em *k* grupos, onde, o valor de *k* é definido pelo usuário. Para isto, o algoritmo começa escolhendo *k* objetos como sendo os centros dos *k clusters*, depois divide os objetos em *k clusters* baseando-se na

medida de similaridade escolhida (CARLANTONIO, 2001). Posteriormente a divisão inicial, existem duas formas de escolher qual elemento irá ser o centro de um *cluster* e consequentemente servir de referência à medida de similaridade (CARLANTONIO, 2001):

- ***k-means:*** média dos objetos pertencentes ao *cluster* relacionado (centro de gravidade do *cluster*),
- ***k-medoids:*** baseia-se no objeto que se encontra mais perto do centro de gravidade do *cluster;*

b) **métodos hierárquicos:** criam decomposições hierárquicas, ou seja, dividem os grupos de forma gradual, tendo como objetivo principal construir uma sequência hierárquica de agrupamentos. Podem ser classificados em aglomerativos e divisivos (DUARTE, 2008):

- **aglomerativos (*botton-up*):** objetos com maior similaridade são clusterizados formando um único grupo, podendo repetir o processo e como decréscimo da similaridade, todos os outros subgrupos se unem formando novamente um grupo único com todos os objetos (ALVES, 2007),
- **divisivos (*top-down)*:** iniciam considerando um grande conjunto como um único *cluster* e posteriormente subdivide-se em grupos menores, removendo a cada iteração o elemento mais distante do centróide do *cluster* (SANTOS, 2009);

c) **métodos baseados em densidade:** são métodos onde a densidade da vizinhança dos objetos ou as funções de densidade específicas irão definir os *clusters*, no qual, para cada objeto pertencente a um *cluster,* a vizinhança deste, dentro de um determinado raio deve conter um número mínimo de objetos (BOSCARIOLI, 2008);

d) **métodos baseados em grades:** divide o espaço dos dados em um número finito de células formando assim uma estrutura de grades, de modo que todas as operações da clusterização se realizam nesta estrutura de grades (LAURO, 2008);

e) **métodos baseados em modelos:** criam hipóteses a um modelo para cada um dos *clusters*, encontrando o melhor ajuste dos dados ao modelo fornecido. Os algoritmos baseados em modelos podem

> localizar *clusters* por meio da construção de uma função, que reflete a distribuição espacial dos pontos de dados, conduzindo a uma maneira de determinar automaticamente o número de *clusters* com base em padrões estatísticos. Para isso, leva em consideração ruídos ou *outliers*, produzindo métodos de particionamento robustos (HAN; KAMBER, 2006, tradução nossa).

Na tabela 2 exemplificam-se alguns algoritmos de clusterização para cada método.

Tabela 2 - Exemplos de algoritmos de clusterização por método

Método	Exemplos de algoritmos
Particionamento	*K-means, k-medoids*, CLARA, CLARANS, CLUSTER, PAM
Hierárquico	AGNES, DIANA, CURE, CHAMELEON, BIRCH, ROCK, CLINK
Densidade	DBSCAN, OPTICS, DENCLUE, GDDBSCAN, DBCLASD
Grade	STING, CLIQUE, WAVECLUSTER, BANG CLUSTERING
Modelo	COWEB, CLASSIT, AUTOCLASS, SNOB, MCLUST, SOM, ART

Fonte: Adaptado de Boscarioli (2008).

A clusterização, assim como as outras tarefas citadas anteriormente, é o alicerce para que se consiga aplicar o *data mining* a uma base de dados, porém, o usuário tendo definido a mesma, precisa de programas que façam isso de forma automatizada. Para isso, existem ferramentas computacionais, em sua maioria pagas, para a aplicação do *data mining*. Cada uma dessas ferramentas possui particularidades com relação as tarefas, algoritmos, métodos e conectividade com bancos de dados. A tabela 3 traz uma listagem de ferramentas de *data* mining, algumas características destas ferramentas e quais bancos de dados essas são suportados. Para facilitar a interpretação da mesma, as colunas representam: Versão do software: Final (F), Beta (B); Licença: Comercial (C), *Freeware* e *Shareware* (F), Pública (P); Disponibilidade de *Download*: Sim (S), Não (N); Tipo da Ferramenta: Acadêmica (A), Comercial (C) ou ambas (A/C); Arquitetura: *Stand Alone* (S), *Client*/*Server* (C/S), Processamento Paralelo (PP).

Tabela 3 - Ferramentas de Data Mining

Ferramenta	Características gerais						Conectividade a base de dados					
	Final/Beta	Licença	Demo/Download	Acad./Comercial	Arquitect. (S,C/S,PP)	ASCII	Dbase	ODBC	MySQL	MS Excel	LOTUS	Outros
Alyuda Neuro Inteligence	F	C	S	C	S	X						
BrainMaker	F	C	N	A/C	S	X	X			X	X	
BSVM	F	F	S	A	S	X						
Clementine	F	C	N	C	S/CS	X		X		X	X	Informix, Oracle, Sybase
DTREG	F	C	S	A/C	S							
EQUBITS Foresight™	F	C	S	A/C	S	X						
EWA Systems	F	C	N	A/C	S/CS				X			
GhostMiner	F	C	N	A/C	S	X		X	X	X		
Gist	F	F	S	A	S	X						
Gornik	F	C	N	C	S/CS	X		X	X	X		
Insightful Miner	F	C	S	A/C	S/CS	X	X	X	X	X		SAS, SPSS, Sybase
Kernel Machines	F	F	S	A	S							
Knowledge Miner	F	C	S	A/C	S							
KXEN	F	C	N	C	S/CS				X			
LIBSVM	F	F	S	A	S	X						
MATLAB NN Toolbox	F	C	S	A	S							
MCubiX from Diagnos	F	C	N	C	S	X		X	X	X		Access, freetext, imagen
MemBrain	F	F	S	A	S							
NeuralWorks Predict	F	C	S	C	S	X						
NeuroSoluctions	F	C	S	A/C	S/CS	x				x		Access
NeuroXL	F	C	N	C	S					X		
IPNNL Software	B	F	S	A	S	X						
Oracle DM	F	C	S	C	S,CS, P	X	X	X	X	X		Oracle, Sybase
Orange	F	F	S	A	S	X						
pcSVM	B	P	S	A	S							
R	F	P	S	A	S	X						
SAS Enterprise Miner	F	C	S	A/C	CS			X				
StarProbe	F	C	S	A/C	S/CS			X				JDBC
STATISTICA NN	F	C	S	A	S/CS							
SvmFu 3	B	P	S	A	S	X						
SVM-light	F	F	S	A	S	X						
TANAGRA	F	F	S	A	S	X						
Hhink-Analitics	F	C	N	C	CS		X		X			
Tiberius	F	C	S	A/C	S/CS	X		X	X	X		
Weka	F	P	S	A	S	X						
XLMiner	F	C	S	A/C	S							
Shell Orion	B	F	S	A					X			XE, Postgres, Firebird

Fonte: Adaptado de Cruz e Cortes (2009).

2.2.3 Shell Orion Data Mining Engine

A *Shell Orion Data Mining Engine* é um projeto acadêmico que vem sendo desenvolvido desde 2005 por acadêmicos do curso de Ciência da Computação da Universidade do Extremo Sul Catarinense (UNESC), no qual, todos os algoritmos implementados até o momento foram desenvolvidos como Trabalhos de Conclusão de Curso (TCC).

Pensando-se em fornecer uma ferramenta gratuita de *data mining*, surgiu a *Shell* Orion, pois as ferramentas disponíveis para esse fim, geralmente são comerciais e de alto custo.

A *Shell Orion* vem sendo implementada na linguagem de programação Java, pois segundo Pelegrin (2005), essa linguagem permite a reutilização de código, é independente de plataforma e possui ambientes de desenvolvimento gratuitos.

Outra vantagem da utilização da plataforma Java para desenvolvimento da *Shell Orion* é a sua Interface de Programação de Aplicações (*Application Programming Interface*-API) denominada *Java Database Connectivity* (JDBC), onde com a utilização dessa API permite que a *Orion* se conecte a qualquer Sistema Gerenciador de Banco de Dados (SGBD) que possua um *driver* disponível, o que traz à ferramenta flexibilidade.

Até o presente momento, como se pode observar na tabela 4, a *Shell Orion* é composta por três tarefas: associação, classificação e clusterização. Na tarefa de associação encontra-se desenvolvido o algoritmo de *Apriori* e FP-Growth, na classificação os algoritmos CART, ID3, C4.5 e *Radial Basis Function Network* (RBF). A clusterização é a tarefa que possui mais algoritmos implementados, pelo método de lógica *fuzzy* encontram-se o *Fuzzy C-Means* (FCM)*, Robust C-Prototypes* (RCP)*, Unsupervised Robust C-Prototypes* (URCP)*, Gustafson-Kessel* (GK) e *Gath-Geva* (GV); pelo método de particionamento o *K-Means;* pelo método de densidade tem-se o Algoritmo *Density-Based Spatial Clustering of Applications With Noise* (DBSCAN); pelo método de redes neurais encontra-se o *Kohonen* e finalmente, pelo método de colônia de formigas tem-se o algoritmo S*tandard Ant Clustering Algorithm* (SACA), Ant-Based Clustering e A²CA.

Tabela 4 - Evolução da Shell Orion Data Mining Engine

Ano	Tarefa	Método	Algoritmo	Atributos	Referência
2005	Associação	Regras de Associação	Apriori	Numéricos	(CASAGRANDE, 2005)
2005	Classificação	Árvores de Decisão	ID3	Nominais	(PELEGRIN, 2005)
2007	Classificação	Árvores de Decisão	CART	Nominais e numéricos	(RAIMUNDO, 2007)
2007	Clusterização	Particionamento	*K-means*	Numéricos	(MARTINS, 2007)
2007	Clusterização	Redes Neurais	*Kohonen*	Numéricos	(BORTOLOTTO, 2007)
2008	Clusterização	Lógica *Fuzzy*	Gustafson-Kessel	Numéricos	(CASSETARI JUNIOR, 2008)
2009	Clusterização	Lógica *Fuzzy*	Gath-Geva	Numéricos	(PEREGO, 2009)
2009	Classificação	Árvores de Decisão	C4.5	Nominais e Numéricos	(MONDARDO, 2009)
2010	Classificação	Redes Neurais	RBF	Numéricos	(SCOTTI, 2010)
2010	Clusterização	Lógica *Fuzzy*	RCP	Numéricos	(CROTTI JUNIOR, 2010)
2010	Clusterização	Lógica *Fuzzy*	URCP	Numéricos	(CROTTI JUNIOR, 2010)
2010	Clusterização	Lógica Fuzzy	FCM	Numéricos	(CROTTI JUNIOR, 2010)
2011	Clusterização	Densidade	DBSCAN	Numéricos	(GAVA, 2011)
2012	Clusterização	Enxame de Formigas	SACA	Numéricos	(GHELLERE, 2012)
2012	Clusterização	Enxame de Formigas	Ant-Based Clustering	Numéricos	(GHELLERE, 2012)
2012	Clusterização	Enxame de Formigas	A^2CA	Numéricos	(GHELLERE, 2012)
2013	Classificação	Teoria de Bayes	Naive Bayes	Nominais/ Numéricos	(NOVASKI, 2012)

Fonte: Adaptado de Novaski (2012).

Todos os algoritmos da Shell Orion foram testados e analisados quanto ao processamento, porém em sua grande maioria, não se levou em consideração a qualidade das partições geradas por esses algoritmos e nem a aplicação de diferentes bases de dados. Assim, a fim de se analisar esses resultados, o uso e a interpretação de medidas de qualidade em *data mining* são imprescindíveis.

3 MEDIDAS DE QUALIDADE APLICADAS À CLUSTERIZAÇÃO EM DATA MINING

O particionamento de um conjunto de dados resultante da aplicação de um algoritmo de clusterização pode ser tradicional, com valores {0,1} ou *fuzzy*, com graus de pertinência variando entre [0,1] (SILVA; GOMIDE, 2004).

A divisão de um conjunto de dados gerada por algoritmos de clusterização, geralmente não leva em consideração se essa estrutura realmente existe, tornando assim, importante a busca em saber se esses padrões encontrados são válidos, ou se apenas um particionamento sem significado foi gerado pelo algoritmo (BOSCARIOLI, 2008).

Rezaee et al (1998, tradução nossa) salienta que como a clusterização é um processo não-supervisionado, onde não se tem conhecimento *a priori*, é inevitável o uso de algum tipo de avaliação das partições geradas. Essas avaliações, ao mesmo tempo em que são uma etapa importante no processo, muitas vezes, podem ser consideradas fatigantes para o usuário, pois, de certa forma, identificar qual algoritmo se adapta melhor aos seus dados, pode não ser uma tarefa trivial (DUARTE, 2008).

Após a execução dos algoritmos de clusterização, é o momento em que podem surgir algumas dúvidas aos usuários relacionadas ao particionamento, como por exemplo (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001, tradução nossa):

a) quantos *clusters* possuem o conjunto de dados?
b) estes resultados representam os dados reais?
c) esta é a melhor maneira de particionar o conjunto de dados?

Considerando isso, tem-se a necessidade de métodos capazes de responder a estas perguntas, qualificando o particionamento gerado pelos algoritmos de clusterização, de modo que se possa, além de avaliar o particionamento dos dados, identificar a quantidade exata de *clusters* presentes em um conjunto de dados, dados estes, que geralmente são não rotulados (FRANCO, 2002).

Conforme Jain e Dubes (1988, tradução nossa) a validação de *cluster* se refere ao procedimento de avaliar os resultados da clusterização de um modo quantitativo e objetivo.

As medidas de validação aplicam-se às partições feitas em um conjunto de dados, geradas a partir de um método de clusterização, com o intuito de estimar a qualidade do particionamento produzido pelo método aplicado (FRANCO, 2002).

Estes métodos ou medidas que indicam o quão “bom” foi o agrupamento, podem ser encontradas na literatura com diversas denominações, tais como: Índices de validação de *clusters* (BEZDEK; PAL, 1998; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002; PAKHIRA; BANDYOPADHYAY; MAULIK, 2004; TAN; STEINBACH; KUMAR, 2009; WU; YANG, 2005), Critérios de validação (XIE; BENI, 1991), medida de separação de *clusters* (DAVIES; BOULDIN, 1979), análise de *cluster* (CALINSKI; HARABASZ, 1974; JOLLIFFE; PHILIPP, 2010; ROUSSEEUW, 1987) e Medidas de Qualidade (GUILLET; HAMILTON, 2010), sendo esta última a denominação utilizada no trabalho.

Tan, Steinbach e Kumar (2005, tradução nossa), trazem algumas abordagens na aplicação de medidas de qualidade para a clusterização que devem ser resolvidas com a aplicação das mesmas:

a) determinar o número “correto” de *clusters*;
b) avaliar quanto os resultados de um agrupamento se ajustam aos dados, sem auxílio de informações externas;
c) comparar os resultados da análise de agrupamento com os resultados externos conhecidos, como rótulos de classes fornecidos;
d) comparar dois conjuntos de grupos para determinar qual é melhor;
e) determinar e distinguir a existência de dados aleatórios nos dados.

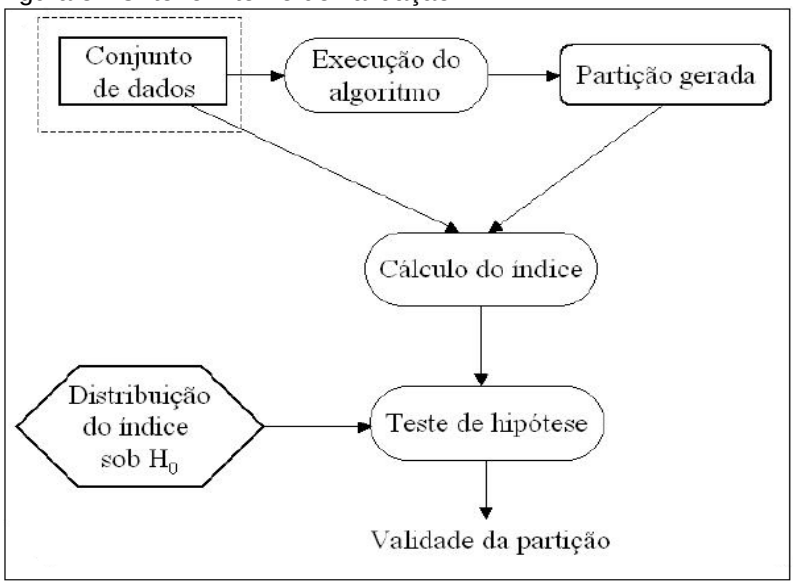
Geralmente, essas medidas na clusterização, são definidas pela combinação da coesão e separabilidade dos *clusters*, no qual, coesão refere-se as medidas de proximidade dos conjuntos, como por exemplo, a variância; e separabilidade indica o quão distintos são os dois *clusters*, como a distância entre objetos de dois *clusters*, por exemplo (RENDÓN et al, 2011, tradução nossa).

A validação de uma estrutura de clusterização pode ser expressa de três formas (JAIN; DUBES,1988, tradução nossa):

a) **critério interno:** em uma verificação interna de validação busca-se determinar se a estrutura descoberta é intrinsecamente apropriada para os dados, sendo avaliações subjetivas e dependentes do domínio.

Na figura 5 pode-se observar o funcionamento da estrutura interna (KANTARDZIC, 2011, tradução nossa);

Figura 5 - Critério Interno de validação

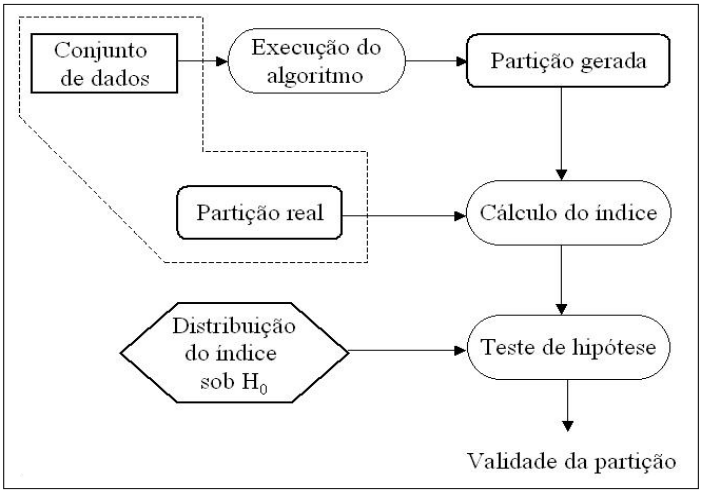


Fonte: Adaptado de Faceli, Carvalho e Souto (2005).

b) critério externo: avaliam-se os resultados de um algoritmo de clusterização baseando-se em uma estrutura pré-especificada na qual é imposta em um conjunto de dados e reflete a estrutura intuitiva do conjunto de dados, havendo duas formas diferentes (figura 6):

- comparando a estrutura da clusterização resultante *C* para uma partição independente dos dados *P*, que foi construído para a intuição das pessoas sobre a estrutura de agrupamento do conjunto de dados,
- comparando-se a matriz de proximidade *Q* à partição *P*, (GAN; MA; WU, 2007, tradução nossa);

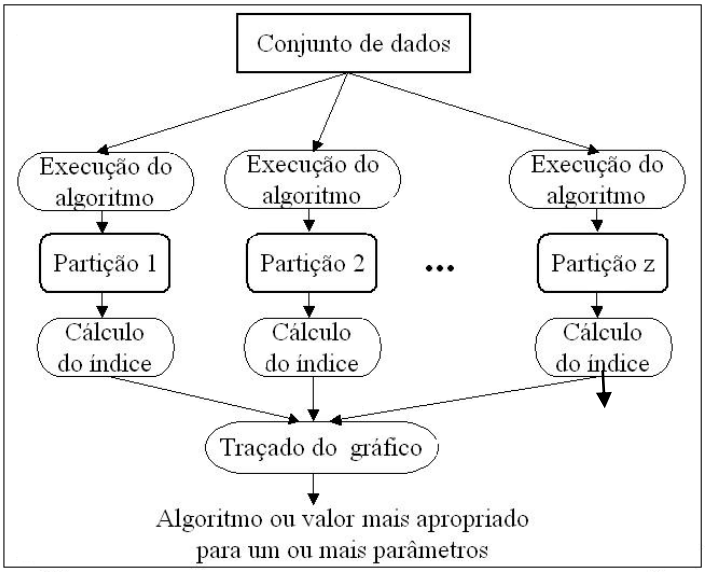
Figura 6 - Critério Externo de validação



Fonte: Adaptado de Faceli, Carvalho e Souto (2005).

c) **critério relativo (híbrido):** faz várias tentativas de agrupamento de um conjunto contra outro, geralmente por meio de uma "pontuação" (BAARSCH; CELEBI, 2012, tradução nossa). Basicamente o critério relativo (figura 7) é a escolha do melhor esquema de agregação de um conjunto de esquemas definidos de acordo com um critério pré-determinado. Destaca-se que o critério relativo se sobressai perante os critérios externos e internos, pelo fato destes serem testes estatísticos, possuindo a desvantagem de sua demanda computacional se tornar elevada (HALKIDI et al, 2002, tradução nossa).

Figura 7 - Critério Relativo de validação



Fonte: Adaptado de Faceli, Carvalho e Souto (2005).

Apesar de Jain e Dubes (1988, tradução nossa) mostrarem os três critérios (interno, externo e relativo) das medidas, Baarsch e Celebi (2012, tradução nossa) destacam, que na realidade no decorrer dos anos apenas dois critérios têm recebido atenção de pesquisadores da área, o externo e o relativo.

Após o entendimento das medidas de qualidade na clusterização, realizou-se uma pesquisa na literatura das principais medidas existentes, do período de 1973 a 2013, pela tabela 5, pode-se observar além das medidas, os autores, o ano de publicação, o título do artigo onde a medida foi descrita e o critério (interno, externo, relativo).

Tabela 5 - Medidas de qualidade.

(continua)

Nome da Medida	Ano	Autores	Artigo Proposto	Critério
Dunn	1973	DUNN	A fuzzy relative of the ISODATA process and its use in detecting compacted well-separated clusters	Relativo
Variance Ratio Criterion (VRC)	1974	CAKLINSKI, HARABASZ	A dendrite method for cluster analysis	Relativo
Coeficiente de Partição (CP)	1974	BEZDEK; DUNN	Numerical taxonomy with fuzzy sets Some recent investigations of a new fuzzy partition algorithm and its application to pattern classification problem	Relativo
Partição Entrópica (PE)	1974	BEZDEK	Cluster validity with fuzzy sets	Relativo
Gamma	1975	BAKER, HUBERT	Measuring the power of hierarchical cluster analysis	Interno
C-Index	1976	HUBERT, LEVIN	A general statistical framework for assessing categorical clustering in free call	Interno
Davies-Bouldin (DB)	1979	DAVIES, BOULDIN	A clustering separation measure	Relativo
Performance da Nebulosidade	1978	ROUBENS	Fuzzy clustering algorithms and their cluster validity	Relativo
Silhouette	1987	ROUSSEUW	Silhouette: a graphical aid to the interpretation and validation of cluster analysis	Relativo
Fukuyama-Sugeno (FS)	1989	FUKUYAMA, SUGENO	A new method of choosing the number of clusters for fuzzy c-means method	Relativo
Compacidade e Separação (XIE BENI)	1991	XIE, BENI	A validity Measure for Fuzzy clustering	Relativo
Variações do Davies-Bouldin e Dunn, baseados em teoria dos grafos:	1997	PAL, BISWAS	Cluster validation using graph theoretic concepts	Relativo
Medidas baseadas no Dunn generalizados:	1998	BEZDEK, PAL	Some new indexes of cluster validity	Relativo

Tabela 5- Medidas de qualidade.

(conclusão)

Nome da Medida	Ano	Autores	Artigo Proposto	Critério
Separação e Dispersão (SD)	2000	HALKIDI, VAZIRGIANNIS	Quality Scheme Assessment in the Clustering Process	Relativo
S_Dbw	2001	HALKIDI, VAZIRGIANNIS	Clustering validity assessment: find the optimal partitioning of a data set	Relativo
Separação da Partição	2001	YANG, WU	A cluster validity index for fuzzy clustering	Relativo
Contraste entre classes	2002	FRANCO	A validity Measure for Hard and Fuzzy Clustering derived from Fisher's linear discriminant	Relativo
CS	2004	CHOU, SU, LAI	A new cluster validity measure and its application to image compression	Relativo
PBM	2004	PAKHIRAA, BANDYOPADHYAYB, MAULIKC	Validity index for crisp and fuzzy clusters	Relativo
Davies-Bouldin*	2005	KIM, RAMAKRISHMA	New indices for cluster validity assessment	Relativo
Score Function	2007	SAITTA, RAPHAEL, SMITH	A bounded index for cluster validity	Interno
Sym	2008	BANDYOPADHYAY, SAHA	A point symmetry-based clustering technique for automatic evolution of clusters	Relativo
Distância Baseada em Pontos Simétricos	2009	SAHA, BANDYOPADHYAY	Performance evaluation of some symmetry-based cluster validity indexes	Relativo
COP	2010	GURRUTXAGA et al	SEP/COP: an efficient method to find the best partition in hierarchical clustering based on a new cluster validity index	Relativo
Negentropy increment	2010	FERNÁNDÈZ, CORBACHO	Normality-based validation for crisp clustering	Relativo
SV	2011	ŽALIK, ŽALIK	Validity index for clusters of different sizes and densities	Relativo
OS	2011	ŽALIK, ŽALIK	Validity index for clusters of different sizes and densities	Relativo

Fonte: Do Autor.

Após a pesquisa realizada acerca das principais medidas de qualidade utilizadas na tarefa de clusterização, mostrada pela tabela 5, tornou-se necessário

uma descrição sucinta de cada medida, a fim de se obter um melhor entendimento da mesma e de suas principais equações.

3.1 VARIANCE RATIO CRITERION (VRC)

O VRC proposto por Calinski, Harabasz em 1974 é uma medida que considera a homogeneidade interna dos clusters e a separação entre clusters. Basicamente, é dividida em duas funções: soma da distância quadrática intra-cluster do inglês *Within-Group Sum of Squares* (WGSS) e a soma da distância quadrática entre clusters do inglês *Between-Groups Sum of Squares* (BGSS), sendo estas representadas pelas equações (1) e (2) (CALINSKI; HARABASZ, 1974, tradução nossa):

$$WGSS = \sum_{i=1}^{k} d_{intra}(C_i) \tag{1}$$

$$BGSS = \sum_{i=1}^{k} d_{inter}(C_i) \tag{2}$$

Onde, *k* é o número de *clusters* do particionamento e $d_{intra}(C_i)$ e $d_{inter}(C_i)$ são os cálculos das distâncias intra-clusters e inter-clusters. Resultados satisfatórios referentes aos particionamento dos clusters são obtidos, quando os valores do *WGSS* são baixos e os valores do *BGSS* são altos. A equação que efetivamente vai indicar o VRC é mostrada pela equação 3:

$$VRC = \frac{BGSS}{k-1} / \frac{WGSS}{n-k} \tag{3}$$

3.2 COEFICIENTE DE PARTIÇÃO

Assim como a VRC, o Coeficiente de Partição (CP) também foi proposta em 1974, porém, por J. C. Bezdek, no qual, pode ser definida como um somatório dos graus de pertinência de todos os elementos, com relação aos *clusters* ao quadrado, por fim, dividindo-os pelo total de elementos presentes na base de dados (BEZDEK, 1974, tradução nossa).

$$CP = \frac{\sum_{j=1}^{N}\sum_{i=1}^{C}\mu_{ij}^{2}}{N} \quad (4)$$

Onde:

a) ***V$_{CP}$****:* índice do coeficiente de partição;

b) ***C****:* quantidade de *clusters;*

c) ***N****:* quantidade de elementos;

d) μ_{ij} : grau de pertinência do j-ésimo elemento no i-ésimo cluster.

O valor de *CP* deve ser maximizado, isto é, o maior possível no intervalo de [1/*C*, 1], quanto mais próximo de 1/*C* o valor indicar, menor a existência de grupos bem definidos no particionamento (BEZDEK et al, 2005, tradução nossa).

3.3 COEFICIENTE DE PARTIÇÃO ENTRÓPICA

Medida proposta por Nikhil R. Pal e James C. Bezdek, em 1995, tendo como equação principal:

$$V_{PE} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C}\left[\mu_{ij}\log(\mu_{ij})\right] \quad (5)$$

Onde:

a) ***V$_{PE}$***: índice de partição entrópica;

b) ***C***: quantidade de *clusters*;

c) ***N***: quantidade de elementos;

d) μ_{ij} : grau de pertinência do j-ésimo elemento no i-ésimo *cluster*.

Ao contrário do coeficiente de partição, o valor do partição entrópica deve ser minimizado, ou seja, quanto mais próximo de 0 mais bem definidos serão os *clusters*, sendo que, quando mais próximo do limite superior no intervalo de [0, log(*C*)] menos definidos serão os *clusters*.

3.4 PERFORMANCE DA NEBULOSIDADE

Proposta em 1978 por Mark Roubens, sua denominação do inglês (*Fuzziness Performance*), é uma medida para validação do número “ideal” de

clusters à um conjunto de dados, sendo uma derivação do Coeficiente de Partição. Basicamente, por um método de categorização que busca estimar o grau de nebulosidade dos *clusters*, sendo que, quando um valor baixo é encontrado, presume-se que o número de categorias é o ideal, tendo como equação principal (ROUBENS, 1978, tradução nossa):

$$FP = 1 - \frac{\sum_{x} \min\{\mu_k(x), \mu_k,(x)\}}{\sum_{x} \max\{\mu_k(x), \mu_k,(x)\}} \tag{6}$$

3.5 DUNN

Medida proposta por J. C. Dunn em 1974 é uma medida com muitas variações, onde a coesão é estimada pela distância do vizinho mais próximo e a separação pelo diâmetro máximo do *cluster*, sendo que, quando valores elevados são encontrados em *D(K)* mais separados e compactos são os clusters. O ponto de máximo no gráfico de *D(K)* contra *K* pode ser uma indicação do número de *clusters* que mais se ajusta aos dados. Os problemas do índice Dunn original são a sua complexidade e sensibilidade a ruídos. As medidas originais são definidas pela equação 7, em que $\delta(c_i, c_l)$ é uma função de dissimilaridade entre os *clusters* C_i e C_j e $\Delta(c_i)$ é a distância intra-cluster do *cluster* C_i, uma medida de dispersão do cluster.

$$D(K) = \min_{i=1,\ldots,k} \left\{ \min_{j=i+1,\ldots,K} \left\{ \frac{\delta(C_I, C_j}{\max_{k=1,\ldots K}} \right\} \right\} \tag{7}$$

Onde:

$$\delta(c_k, c_l) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{8}$$

$$\Delta(C_i) = \max_{x\ ,y\ \in c_i} \{d\ (x\ , x\ )\} \tag{9}$$

3.6 GAMMA

Medida proposta por Frank B. Baker e Lawrence J. Hubert em 1975, denominada Gamma, sendo uma adaptação do Gamma Kruskal e Goodman[1] e pode ser descrita como (BAKER; HUBERT, 1975, tradução nossa):

$$G(C) = \frac{\sum_{C_k \in C} \sum_{x_i, x_j \in c_k} dl(x_i, x_j)}{n_w \left(\binom{N}{2} - n_w \right)} \tag{10}$$

Onde, $dl(x_i, x_j)$ indica o número de todos os objetos pares em *x*, ou seja, x_k e *j*, que contempla duas condições: (a) x_k e x_i estão em diferentes *clusters*, e (b) *De* (x_k, x_l) < *De* (x_i, x_j). Neste caso o denominador é apenas o fator de normalização (BAKER; HUBERT, 1975, tradução nossa);

3.7 DAVIES-BOULDIN (DB)

A medida Davies-Bouldin foi proposta em 1979 por David L. Davies e Donald W. Bouldin e provavelmente seja a medida mais usada em estudos de validação de *clusters* (ARBELAITZ et al, 2013), de forma geral ela estima a coesão baseado na distância de pontos em um cluster para o centróide e a separação baseada na distância entre os centróides, sendo que, valores baixos de $DB(C)$ indicam *clusters* compactos e separados. Sendo sua equação definida como:

$$DB(C) = \frac{1}{K} \sum_{x_i \in C_k} d_e \left(x_i, \overline{c_k} \right) \tag{11}$$

Onde:

$$S(c_k) = 1/|c_k| \sum_{x_i \in c_k} d_e \left(x_i, \overline{c_k} \right) \tag{12}$$

3.8 SILHOUETTE

A medida de qualidade Silhouette foi proposta por Peter J. Rousseeuw em 1987, sendo uma medida do tipo sumarização normalizada (*normalized summation-type índex*). A coesão é medida baseada na distância entre todos os pontos no mesmo *cluster* e a separação é baseada na distância do vizinho mais próximo, sendo que quando os valores forem maximizados de Sil há presença de grupos bem definidos, sendo definida como (ROUSEEUW, 1987, tradução nossa):

$$Sil(C) = 1/N \sum_{C_K \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{\max\{a(x_i, c_k), b(x_i, c_k)\}} \tag{13}$$

Onde:

$$a(x_i, x_k) = 1/|c_k| \sum_{x_j \in c_k} d_e(x_i, x_j) \tag{14}$$

$$b(x_i, c_k) = \min_{c_l \in C \setminus c_k} \left\{ 1/|c_l| \sum_{x_j \in c_l} d_e(x_i, x_j) \right\} \quad (15)$$

3.9 C-INDEX

Proposta por L. J. Hubert e J. R. Levin em 1976 é uma medida que verifica a coesão interna de um *cluster*, sendo que, quanto mais semelhantes forem os pontos contidos em um agrupamento e mais compacto for o *cluster*, mais próximo de zero será o resultado do *C-Index*, em um intervalo entre [0,1] (MILLIGAN; COOPER, 1985, tradução nossa).

$$C - Index = \frac{d_w(C_i) - \min(n_w)}{\max(n_w) - \min(n_w)} \quad (16)$$

Onde:

a) $d_w(C_i)$: somatório das distâncias entre todos os pares de pontos no cluster $C_{i;}$

b) n_w : quantidade de pares de pontos no cluster $C_{i;}$

c) $\min(n_w)$: somatório das n_w menores distâncias entre os pares de objetos do conjunto de dados;

d) $\max(n_w)$: somatório das n_w maiores distâncias entre pares de objetos do conjunto de dados.

3.10 SEPARAÇÃO E DISPERSÃO (SD)

Medida proposta por Maria Halkidi, Y. Batistakis e M. Vazirgiannis em 2001 é basicamente baseada nos conceitos de dispersão média (*Disp*) e separação (*Sep*) dos *clusters,* sendo um valor alto o parâmetro para a definição do número de categorias. Sua equação pode ser definida como:

$$SD(K) = a \cdot Disp(K) \cdot Sep(K) \quad (17)$$

Onde:

$$Disp(K) = \frac{1}{K} \sum_{i=1}^{K} \|\sigma(c_i)\| / \|\sigma(X)\| \quad (18)$$

$$Sep(K) = \frac{D_{\max}}{D_{\min}} \sum_{q=1}^{K} \left(\sum_{l=1}^{K} \|c_q - c_l\| \right)^{-1} \quad (19)$$

3.11 S_DBW

A S_Dbw assim como a SD também foi proposta por Maria Halkidi e Michalis Vazirgiannis porém em 2002, é uma medida tipo relativa que tem uma formulação mais complexa baseada na distância Euclidiana normalizada, sendo que, chega-se ao número “ideal” de *clusters* quando o menor valor de S_Dbw for atingido. A S_Dbw pode ser definida como:

$$S\_Dbw(K) = Disp(K) + Dens\_IC(K) \tag{20}$$

Onde:

$$Disp(n_c) \tag{21}$$

$$Dens\_Ic(K) = \frac{1}{K \cdot (K-1)} \sum_{i=1}^{K} \left(\sum_{i \neq j}^{K} \frac{densidade(u_{ij})}{\max\{densidade(c_i), densidade(c_i)\}} \right) \tag{22}$$

3.12 CS-INDEX

Medida apresentada em 2004 por C-H. Chou, M-C. Su, E. Lai, no qual, foi proposta à um ambiente de compressão de imagem, porém pode ser usada à qualquer ambiente, é uma medida do tipo relativa, que estima a coesão pelo diâmetro dos *clusters* pela distância pelo vizinho mais próximo, sendo que, um valor baixo indica um bom particionamento,podendo ser definida pela equação (CHOU; SU;LAI, 2004, tradução nossa):

$$CS(C) = \frac{\sum_{C_k \in C} \{1/|c_k| \sum_{x_i \in C_k} \max_{x_j \in C_k} \{d_e(x_i, x_j)\}\}}{\sum_{c_k \in C} \min_{c_l \in C \setminus c_k} \{d_e(\overline{c_k}, \overline{c_l})\}} \tag{23}$$

3.13 FUNÇÃO DE PONTUAÇÃO

Proposta por S. Saitta, B. Raphael e I. Smith em 2007 do inglês (*score function*) é uma medida do tipo soma, onde a separação é medida baseada na distância dos centroides dos *clusters* e a coesão é baseada na distância dos pontos em um *cluster* ao centroide, sendo que, quanto maior for o valor de *SF*, mais “perto do ideal” será o número de clusters, sendo sua equação (SAITTA; RAPHAEL; SMITH, 2007, tradução nossa):

$$SF(C) = 1 - \frac{1}{e^{e^{bcd(c)+wcd(c)}}} \tag{24}$$

Onde:

$$bcd(c) = \frac{\sum_{c_k \in C} |c_k| d_e(\overline{c_k}, \overline{X})}{N \times K} \tag{24}$$

$$wcd(C) = \sum_{c_k \in C} 1/|c_k| \sum_{x_i \in C_k} d_e(x_i, \overline{c_k}) \tag{25}$$

3.14 SV-INDEX

Apresentada em 2011 por K. R. Zalik, B. Zalik, seu critério é do tipo relativo, onde, a medida estima a separação pela distância do vizinho mais próximo e a coesão baseada na distância dos pontos da borda do *cluster* em um *cluster* para seu centroide, sendo que quanto mais maximizado for o valor de *SV*, mais particionados estão os grupos e mais "ideal" é o numero de *clusters*, onde, sua equação é definida por (ŽALIK; ŽALIK, 2011, tradução nossa):

$$SV(C) = \frac{\sum_{C_k \in C} \min_{c_l \in C \setminus c_k} \{d_e(\overline{c_k}, \overline{c_l})\}}{\sum_{c_k \in C} 10/|c_k| \sum \max_{x_i \in C_k} (0.1|c_k|)\{d_e(x_i, \overline{c_k})\}} \tag{26}$$

3.15 OS-INDEX

Este é outra medida do tipo relativa, proposta em 2011 por Krista Rizman Žalik e Borut Žalik, onde um estimador de separação mais complexo é utilizado, onde, ao contrário do SV-Index, quanto menor for o valor de *OS,* mais bem definidos serão os *cluster* e mais "próximo da realidade" é o número de clusters gerados, sendo definido por (ŽALIK; ŽALIK, 2011, tradução nossa):

$$OS(C) = \frac{\sum_{c_k \in C} \sum_{x_i \in C_k} ov(x_i, c_k)}{\sum_{c_k \in C} 10/|c_k| \sum \max_{x_i \in C_k} (0.1|c_k|)\{d_e(x_i, \overline{c_k})\}} \tag{25}$$

Onde:

$$ov(x_i, c_k) = \begin{cases} \frac{a(x_i, c_k)}{b(x_i, c_k)} \; if \; \frac{b(x_i, c_k) - a(x_i, c_k)}{b(x_i, c_k) + a(x_i, c_k)} < 0.4 \\ 0 \quad otherwise \end{cases} \tag{26}$$

$$a(x_i, c_k) = 1/|c_k| \sum_{x_i \in C_k} d_e(x_i, x_j) \quad (27)$$

$$b(x_i, c_k) = 1/|c_k| \sum_{x_i \in C_k} \min(|c_k)\{d_e|(x_i, x_j)\} \quad (28)$$

3.16 VARIAÇÕES DO DB E DUNN

Propostas em 1997 por N. R. PAL e J. BISWAS são variações aplicadas ao Davies-Boudin e Dunn baseados em teoria dos grafos, as variações afetam a forma como os estimadores de coesão são calculados, $\Delta(C_k)$ para o Dunn e $S(C_k)$ para o Davies-Bouldin. Para cada uma das três versões (MST, RNG E GG) essas duas funções são calculadas da mesma forma. Primeiramente, um determinado tipo de gráfico é calculado para c_k, tornando os objetos no cluster como vértices e a distância entre objetos como o peso de cada aresta. Em seguida, o maior peso é tomado como o valor para $\Delta(c_k)$ e $S(c_k)$. A diferença entre as e variações vem da seleção do grafo selecionado. Para o MST uma árvore geradora mínima é construída, para o RNG um gráfico de Vizinhança Relativa e GG o grafo de Gabriel;

3.17 GENERALIZAÇÕES DO DUNN

Medidas propostas por James C. Bezdek e Nikhil R. Pal em 1998, no qual todas são variações do Dunn, sendo uma combinação de três variantes de δ: estimador de separação e coesão e duas variações de Δ. As variações 3, 4 e 5 de δ e 1 e 3 pra Δ são mostradas abaixo:

$$\delta^5(c_k, c_l) = \frac{1}{|c_k| + |c_l|} \left(\sum_{x_i \in C_k} d_e(x_i, \overline{c_k}) + \sum_{x_i \in C_l} d_e(x_j, \overline{c_l}) \right) \quad (29)$$

$$\delta^4(c_k, c_l) = d_e(\overline{c_k}, \overline{c_l}) \quad (30)$$

$$\delta^3(c_k, c_l) = \frac{1}{|c_k||c_l|} \sum_{x_i \in C_k} \sum_{x_j \in C_l} d_e x_i, x_j \quad (31)$$

$$\Delta^3(c_k) = 2/|c_k| \sum_{x_i \in c_k} d_e(x_i, \overline{c_k}) \quad (32)$$

$$\Delta^1(c_k) = \Delta(c_k) \quad (33)$$

3.18 SIMÉTRICA

Medida proposta por Sanghamitra Bandyopadhyay, e Sriparna Saha, em 2008 no qual é uma adaptação da I (medida proposta por Ujjwal Maulik e Sanghamitra Bandyopadhyay em 2002). A Simétrica (SYM), assim como a I, é baseada na Distância entre pontos simétricos do inglês (*Point Symmetry Distance*), sendo que, quanto maior é o valor de Sym, mais próximo o número atual de clusters é o "ideal", podendo ser definida como:

$$Sym(C) = \frac{\max_{c_k, c_l \in C} \{d_e(\overline{c_k}, \overline{c_l})\}}{K \sum_{c_k \in C} \sum_{x_i \in c_k} d_{ps}{}^*(x_i, c_k)} \quad (34)$$

3.19 MEDIDAS BASEADAS NA DISTÂNCIA EM PONTOS SIMÉTRICOS (MBDPS)

As MBDPS é uma adaptação de três medidas já conhecidas, a Davies-Bouldin, Dunn e Dunn generalizado. Os autores que propuseram as medida foram S. Saha e S. Bandyopadhyay, no ano de 2009. A MBDPS do Davies-Bouldin se difere no calculado de S:

$$S(c_k) = 1/|c_k| \sum_{x_i \in c_k} d^*_{ps}(x_i, c_k) \quad (35)$$

A variação da MBDPS do Dunn foi a seguinte:

$$\Delta(c_k) = \max_{x_i \in c_k} \left\{ d^*_{ps}(x_i, c_k) \right\} \quad (36)$$

E finalmente, a variação do Dunn Generalizado:

$$\Delta(c_k) = 2/|c_k| \sum_{x_i \in c_k} d^*_{ps}(x_i, c_k) \quad (37)$$

3.20 COP

Embora esta medida fosse proposta pela primeira vez para ser usada em conjunção com um algoritmo de pós-processamento de *cluster* hierárquico, ela também pode ser usada como uma medida normal de validação. A COP foi proposta por Ibai Gurrutxaga et al em 2010, sendo do tipo relativa e a coesão é estimada pela

distância entre os pontos de um cluster ao seu centroide e a separação baseia-se na distância vizinho mais próximo. Sua equação é definida por:

$$COP(C) = \frac{1}{N}\sum_{c_k \in C}|c_k| \frac{1/|c_k|\sum_{x_j \in c_k} d_e(x_i, \overline{c_k})}{\min_{x_i \notin c_k} \max_{x_j \in c_k} d_e(x_i, x_j)} \tag{38}$$

3.21 NEGENTROPY INCREMENT

Proposta por Luis F. Lago-Fernández e Fernando Corbacho em 2010 é uma medida baseada na estimativa normalizada do cluster, portanto, não se baseia na estimativa de coesão e separação, sendo representada por:

$$NI(C) = {}^{1}\!\!\diagup\!\!{}_{2}\sum_{c_k \in C} p(c_k)\left|\sum_{c_k}\right| - 1/2\log\left|\sum_{x}\right| - \sum_{c_k \in C} p(c_k)\log p(c_k) \tag{39}$$

3.22 XIE-BENI (COMPACIDADE E SEPARAÇÃO)

O Xie-Beni, denominado assim, pelos seus autores Xuanli Lisa Xie e Gerardo Beni, sendo proposta em 1991, podendo ser definida por:

$$XB(K) = \frac{\sum_{I=1}^{k}\sum_{J=1}^{n} u_{ij}{}^{2}\left\|c_i - x_j\right\|^2}{n \cdot \min_{i,l}\left\|c_i - c_l\right\|^2} \tag{40}$$

Onde, grupos compactos e bem separados são obtidos quando um valor baixo é encontrado (XIE; BENI, 1991, tradução nossa);

3.23 PBM

Denominado dessa forma pelas iniciais dos sobrenomes de seus autores, Malay K. Pakhiraa, Sanghamitra Bandyopadhyayb e Ujjwal Maulik, assim como as outras medidas, buscam avaliar a coesão e separação de *clusters,* sendo que, o valor máximo de (PBM) indica o particionamento correto. Sendo definido pela equação (PAKHIRA; BANDYOPADHYAY; MAULIK, 2004, tradução nossa):

$$PBM(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^2 \tag{41}$$

3.24 FUKUYAMA-SUGENO

O nome vem dos autores M. Fukuyama e M. Sugeno, sendo proposta em 1989, onde, um valor pequeno indica bom particionamento dos grupos, sendo definida como (FUKUYAMA; SUGENO, 1989, tradução nossa):

$$FS = \sum_{k=1}^{n} \sum_{i=1}^{K} (u_{ik})^m \cdot \left(\left\| x_k - \upsilon_i \right\|_A^2 - \left\| \upsilon_i - \overline{\upsilon} \right\|_A^2 \right) \tag{42}$$

3.25 SEPARAÇÃO DA PARTIÇÃO

Proposto em 2001 por Miin-Shen Yang e Kuo-Lung Wu, bons resultados são encontrados quando seu valor é maximizado, definido como :

$$S(K) = \sum_{k=1}^{K} S_k \tag{43}$$

Onde:

$$S_k = \sum_{i=1}^{n} u_{ik}^2 / u_M - \exp\left(- \min_{j \neq k} \left\{ \left\| \upsilon_k - \upsilon_j \right\|^2 \right\} / \beta_T \right) \tag{44}$$

$$u_M = \max_{1 \leq k \leq K} \sum_{i=1}^{n} u_{ik}^2 \tag{45}$$

$$\beta_T = \frac{\sum_{k=1}^{K} \left\| \upsilon_K - \overline{\upsilon} \right\|^2}{K} \tag{46}$$

3.26 CONTRASTE ENTRE CLASSES

Medida proposta em 2002 por Cláudia Rita de Franco, originalmente chamada de *Inter Class Contrast* (ICC), é uma medida para avaliar particionamentos nebulosos, ou tradicionais, de forma geral, que foi criada para identificar centros de *clusters* alocados muito próximos. Valores altos de ICC demostram que melhor tende a ser o particionamento, suas equações principais são (FRANCO, 2002, tradução nossa):

$$ICC = \frac{S_{Be}}{n} \bullet D_{\min} \bullet \sqrt{c} \tag{47}$$

Onde:

$$D_{\min} = \min_{1 \le i \le c} \left[\min_{i+1 \le j \le c-1} \left\| m_{ei} - m_{ej} \right\| \right] \quad (48)$$

Uma das dificuldades encontradas em trabalhos relacionados à validação de *clusters* é a necessidade de um *framework* para interpretação de uma medida. Considerando-se que o valor final encontrado por uma medida seja, por exemplo, o número 10, qual o significado desse valor, ele é bom, ruim ou aceitável (TAN; STEINBACH; KUMAR, 2005, tradução nossa). Devido a isso, muitas medidas são desconsideradas e acabam não sendo comumente usadas por pesquisadores da área, pois, muitas delas não trazem uma parametrização do valor, ou ainda se trazem, tratam de forma vaga, como por exemplo, baixo ou alto, sem uma definição de intervalo, tornando assim, difícil a aplicação da mesma na validação dos resultados de algoritmos de *data mining*.

4 TRABALHOS CORRELATOS

Nos últimos 40 anos, alguns trabalhos vêm sendo propostos na área de avaliação de *clusters*, ou medidas de qualidade para validação dos mesmos, onde, em sua maioria são pesquisas relacionadas à aplicação de medidas já existentes, ou trabalhos em que os autores criam novas medidas de qualidade. Contudo, por ser um trabalho custoso e que demanda conhecimentos específicos da área de clusterização, algoritmos, tarefas e métodos de data mining poucos trabalhos deste tipo são apresentados.

4.1 NOVOS MÉTODOS DE CLASSIFICAÇÃO NEBULOSA E DE VALIDAÇÃO DE CATEGORIAS E SUAS APLICAÇÕES A PROBLEMAS DE RECONHECIMENTO DE PADRÕES

Essa Dissertação foi submetida por Cláudia Rita de Franco, para obtenção do grau de Mestre pela Universidade Federal do Rio de Janeiro (UFRJ), no ano de 2002.

A dissertação propõe, além de uma pesquisa acerca de algumas medidas existentes, a criação e aplicação de duas novas medidas de validação: a Extensão do Discriminante Linear de Fisher, do inglês *Extended Fisher Linear Descriminant* (EFLD), e o Contraste entre Classes, do inglês *Inter Class Contrast* (ICC). Essas medidas foram aplicadas a problemas de reconhecimento de padrões, propondo-se também uma ferramenta denominada Sistema ICC-KN.

Segundo a autora, a EFLD é uma versão mais rápida da medida original, o *Fisher Linear Descriminant* (FLD), e foi estendida para ser capaz de lidar tanto com particionamentos nebulosos, quanto tradicionais. O ICC é uma medida criada para ser rápida e eficiente, também trata ambas as partições, nebulosas e tradicionais, tendo como objetivo a medição da separação e compacidade do particionamento (FRANCO, 2002).

Como resultados obtidos pelo trabalho, pode-se ressaltar a melhora na validação do número de *clusters* pelas medidas, a EFLD apontou um aumento na velocidade do cálculo, a ICC se mostrou eficaz em ambos os tipos de partições, conseguindo avaliar os dois aspectos propostos pela mesma (compacidade e

separação), provando ser uma medida com alto grau de acertos e com um melhor tempo de processamento com relação a medidas semelhantes (FRANCO, 2002). Um ponto levantado pela autora é a falta de aplicabilidade de alguns algoritmos de clusterização e de medidas de qualidades para validação, tendo, geralmente sua implementação e definição só nos artigos originais dos próprios autores.

4.2 INVESTIGAÇÃO DE MEDIDAS DE VALIDAÇÃO INTERNAS PARA CLUSTERIZAÇÃO K-MEANS

Trabalho desenvolvido por Jonathan Baarsch e M. Emre Celebi, foi publicado nos *Proceedings of the International, MultiConference of Engineers and Computer Scientists* no ano de 2012, em Hong Kong.

Neste artigo os autores abordaram algumas medidas de qualidade sendo aplicadas ao algoritmo k-means em bases de dados artificiais, aquelas compostas por valores que não são reais, a fim de se obter uma forma de identificar qual medida se comporta melhor perante o particionamento. Foram avaliadas sete medidas: Dunn, DB, Sillhouette, VRC, Sum of Squares (Soma quadrática), Point Bi-Serial (PB) e PBM.

Como resultados da pesquisa, destaca-se a superioridade da medida Soma Quadrática em relação as outras, pois, de todas as execuções feitas no artigo, esta obteve o maior índice de resultados satisfatórios, o de 77%, contra 64% do Silhouette, 54% do CH, 47% do DB, 46% do Dunn, 36% do PBM e 19% do PB. Assim, nesta pesquisa percebeu-se a superioridade do Soma Quadrática (BAARSCH; CEBELI, 2012, tradução nossa).

4.3 INDICES DE VALIDAÇÃO DE CLUSTERS INTERNOS VERSUS EXTERNOS

Esta pesquisa foi desenvolvida por Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi e Elvia M. Quiroz, em 2011, tendo sido publicada no *International Journal of Computers and Communications,* no Reino Unido.

A pesquisa é um estudo comparativo de algumas medidas internas e externas de qualidade para validação de *clusters*. Nas internas, os autores fizeram o uso de seis medidas, incluindo: Bic Index, VRC, DB, Silhouette, Dunn e Niva índex,

já para medidas externas utilizaram quatro: F-Measure, NMI, Purity, Entropy. As medidas foram aplicadas aos algoritmos k-means e Bissection k-means, empregando-se treze bases de dados distintas.

Os resultados obtidos pelos autores, pela aplicação do algoritmo Bissection k-means quanto aos valores resultantes das medidas internas, em 86% dos casos o número correto de *clusters* foram encontrados, contra, 51% das medidas externas. Na aplicação do k-means, em 76,9% dos casos as medidas internas atingiram resultados satisfatórios, contra 61% das medidas externas. Assim, concluíram que as medidas internas se comportam melhor na validação de algoritmos de clusterização na busca pelo número ideal de *clusters.*

5 ANÁLISE DAS MEDIDAS DE QUALIDADE APLICADAS AOS ALGORITMOS DE CLUSTERIZAÇÃO DA SHELL ORION DATA MINING ENGINE

A pesquisa desenvolvida consiste na análise de medidas de qualidade para validação dos resultados de algoritmos de clusterização da Shell Orion Data Mining Engine, tendo em vista a importância destas medidas para a avaliação dos resultados gerados por esses algoritmos.

A literatura sobre este assunto ainda não conseguiu identificar medidas que sejam aplicadas a qualquer algoritmo, tarefa ou método, o que há até então, são várias medidas desenvolvidas por pesquisadores da área, que possibilitam a avaliação do particionamento quanto à coesão, separação e o número ideal de *clusters* que o conjunto de dados possui.

5.1 METODOLOGIA

No desenvolvimento deste trabalho as etapas metodológicas consistiram em: realização do levantamento bibliográfico; revisão da literatura sobre as medidas de qualidade para clusterização; seleção e definição das bases de dados a serem utilizadas; pré-processamento dessas bases de dados; seleção dos algoritmos da tarefa de clusterização da Shell Orion a serem aplicados; aplicação dos algoritmos nas bases de dados selecionadas e análise dos resultados obtidos pelos algoritmos por meio das medidas de qualidade.

Na etapa do levantamento bibliográfico fundamentaram-se todos os assuntos que envolvem medidas de qualidade, desde DCBD, *data mining*, clusterização até as medidas de qualidade voltadas à clusterização.

5.1.1 Revisão de Literatura sobre Medidas de Qualidade para Clusterização

Como pode-se observar no Capítulo 3, medidas de qualidade empregadas na clusterização é uma área que apresenta pesquisas que datam de 40 anos. Porém, mesmo sendo um tempo considerável, pode-se dizer que é uma área nova em *data mining*, pois ainda é forte a pesquisa por novas medidas que se comportem melhor a um determinado cenário. Ainda no Capítulo 3, apresentou-se uma tabela originada de uma pesquisa realizada neste trabalho de conclusão de

curso sobre as medidas de qualidade, onde foram encontradas na literatura 29 medidas em um período de tempo de 40 anos (1973 a 2013).

Contudo, no decorrer do trabalho foi identificada a necessidade de se fazer uma revisão de literatura com relação a aplicação dessas medidas, como por exemplo, quantificar quais medidas são mais usadas, quais algoritmos são mais aplicados nessas pesquisas e quais bases de dados são comumente usadas pelos autores. Sendo assim, a pergunta a qual esta revisão de literatura destina-se a responder é: Quais medidas, algoritmos e bases de dados são comumente usados com relação as medidas de qualidade na tarefa de clusterização?

A fim de responder essa pergunta, a primeira etapa da revisão de literatura consistiu na definição das palavras-chave a serem empregadas para a busca destes trabalhos em periódicos e artigos de conferências internacionais. As palavras-chave que se concluiu serem as mais adequadas foram: índice de validação, medidas de validação, medidas de qualidade na clusterização, métricas de validação e análise de *cluster*.

Finalizada esta parte, realizou-se uma busca com estas palavras-chave em repositórios de trabalhos científicos renomados e utilizados em muitas áreas de estudos da computação, sendo estes: ScienceDirect, Association for Computing Machinery (ACM), Institute of Eletrical and Eletronics Engineers (IEEE), Scirus, SpringerLink, Periódicos CAPES e Scielo.

A pesquisa realizada nos periódicos geraram centenas de artigos, porém os que não condiziam com o propósito da pesquisa foram descartados durante o processo, como por exemplo, artigos que apenas traziam a fundamentação de uma nova medida e não a comparava com outras, bem como aqueles que não utilizavam medidas de qualidade, foram eliminados.

Outro fator considerado na pesquisa, foi a identificação dos periódicos que fossem cientificamente relevantes, para isto empregou-se a medida por meio da classificação do Qualis[3] com relação as áreas de Computação e Engenharias IV.

A verificação do Qualis dos periódicos foi realizada utilizando-se o sistema online WebQualis[4], que é um *site* disponível pela Coordenação de Aperfeiçoamento

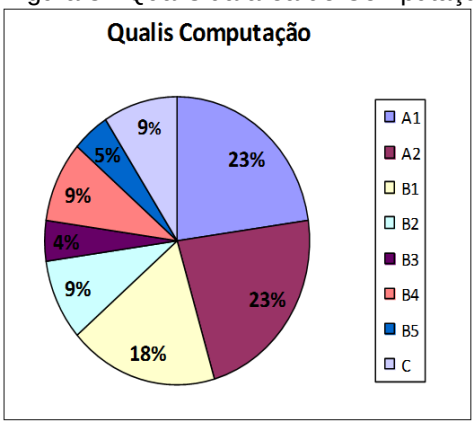
[3] Qualis é um conjunto de procedimentos utilizados pela CAPES para estratificação da qualidade da produção intelectual dos programas de pós-graduação, os periódicos são classificados com estratos indicativos da qualidade dos mesmos, sendo classificados em A1-A2, B1- B5 e C (CAPES, 2009).

de Pessoal de Nível Superior (CAPES) ou também denominado como Portal CAPES.

Os periódicos e conferências cujos artigos foram selecionados na pesquisa são mostrados na tabela 6, bem como sua classificação quanto ao Qualis nas áreas de Computação e Engenharias IV, que são consideradas as principais áreas que envolvem trabalhos nesta área de pesquisa.

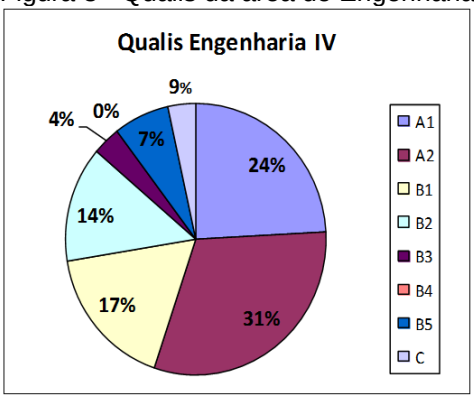
Na área de Computação, como pode ser visto pela figura 9, os periódicos e conferências com Qualis entre A1 e B1 totalizam 64 % do total, seguido de 22% entre os Qualis B2, B3 e B4 e 14% referentes aos Qualis B5 e C.

Figura 8 - Qualis da área de Computação



Na área de Engenharia IV (figura 10), os Qualis A1, A2 e B1 somam 72% e os demais Qualis totalizam 34% das conferências e periódicos encontrados.

Figura 8 - Qualis da área de Engenharia IV



Com relação aos periódicos e conferências utilizadas neste trabalho, pode-se afirmar que ambos são relevantes e de qualidade, pois tanto na computação, quanto na Engenharia IV, os Qualis predominantes foram os mais altos (A1, A2 e B1).

[4] Sistema que fornece a classificação do Qualis de periódicos, por meio de um site de busca online, no qual é atualizado anualmente pelo CAPES: http://qualis.capes.gov.br/webqualis/principal.seam

Tabela 6 - Periódicos utilizados na pesquisa

Nome do Periódico/Evento	QUALIS	
	Comp.	Eng. IV
ACM SIGMOD	A1	-
Advances in Electrical and Computer Engineering	B4	B5
Advances in Engineering Software	B1	B1
Bioinformatics	A1	A2
Communications in Statistics - Theory e Methods	B4	B2
Computers & Mathematics with Applications	A2	B1
Electronics Letters	A1	A1
European Journal of Operational Research	A1	A2
Expert Systems with Applications	A1	A2
Fuzzy Sets and Systems	A1	A1
IEEE Transactions on Pattern Analysis and Machine Intelligence	A1	A1
IEEE Transactions on Systems, Man, and Cybernetics	A1	A1
International Journal of Bioinformatics Research	B3	B3
International Journal of Computer Science and Information Technologies	B5	B5
International Journal of Computer Science Issues	C	-
International Journal Of Computers And Communications	-	-
International Journal of Research in Computer Science	-	-
Journal of Computational and Applied Mathematics	A2	B1
Journal of Intelligent Information Systems	B1	B2
Journal of the American Statistical Association	-	A2
Lecture Notes in Computer Science	C	C
Neural Processing Letters	B1	A2
Pattern Analysis and Applications	B1	A2
Pattern Recognition	A1	A1
Pattern Recognition Letters	A1	B1
Psychometrika	-	B1
IEEE International Conference on Data Mining	Evento	A2
IEEE International Conference on Fuzzy Systems	Evento	A2
IEEE Transactions on Fuzzy Systems	Evento	A1
IEEE Transactions on Knowledge and Data Engineering	Evento	A1
International Conference Methodology, Systems, and Applications.	Evento	B2
International Conference on Computational Intelligence for Modelling, Control and Automation	Evento	B2
International Conference on Scientific and Statistical Database Management	Evento	A2

Fonte: Do Autor.

Após a seleção dos artigos e a verificação da qualidade dos periódicos e das conferências, realizou-se a leitura dos artigos selecionados e a coleta de informações relevantes à pesquisa. Na leitura dos artigos, identificou-se que no

período de 1999 a 2012 tem-se um predomínio de trabalhos na área, então se definiu este como sendo o ponto de corte do período da pesquisa.

Os trabalhos resultantes desta pesquisa realizada podem ser visualizados na tabela 7, que informa: os autores e o ano dos artigos; as medidas que foram aplicadas ou comparadas e se houve definição de uma nova medida no trabalho; os algoritmos e as principais bases de dados que foram usadas, estas separadas por bases de dados artificiais e reais. As bases de dados artificiais são aquelas criadas por meio de simulações, enquanto as bases de dados reais representam valores de problemas reais, não sendo simuladas.

Pode-se observar pela tabela 7 que grande parte dos trabalhos apresentados são internacionais, no qual, dos 38 trabalhos encontrados apenas dois são nacionais, sendo estes: Moraes, Evsukoff e Ebecken (2005) e Silva e Gomide (2004). Com isso, ressalta-se o quão pouco a área de medidas de qualidade na clusterização é explorada em pesquisas nacionais. A fim de contribuir com mais um estudo na área, o presente trabalho também foi inserido na tabela 7 (última linha).

Como o intuito desta revisão de literatura era responder quais medidas, algoritmos e bases de dados são mais aplicados, observou-se que os algoritmos que mais apareceram nos artigos foram: *Fuzzy C-Means*, *K-Means* e DBSCAN.

As medidas de qualidade mais empregadas nos estudos apresentados foram: CP, CE, XB, Dunn, C-Index, DB, Silhouette, CH, PBM e Fukuyama-Sugeno. Com relação as bases de dados que mais foram utilizadas nos trabalhos, destacaram-se: *Iris*, *Wine*, *Breast Cancer Wisconsin* e *Glass*.

Tabela 7 - Revisão de Literatura sobre Medidas de Qualidade para clusterização

(continua)

Artigo	Medidas Empregadas	Propõe nova Medida	Algoritmos aplicados	Bases de dados	
				Artificiais	Reais
Ansari et al (2011)	Dunn, Davies-Bouldin (DB), C-Index, Rand Index, Jaccard Index, Silhouette, Fowlkes Mallows, Sum of the Squared Error (SSE)	-	Leader, DBSCAN, K-Means, K-Medoid, Single Link Aglomerative Hierarchical (SLAH)	-	P.A. Faculdade de Engenharia, Karnataka, INDIA
Baarch e Celebi (2012)	Dunn, DB, Calinski-Harabasz, Silhouette, Point Biserial (MILLIGAN, 1981), PBM, Sum-of-Squares (ZHAO; XU; FRANTI, 2009)	-	K-Means	4 bases de dados Artificiais	-
Bandyopadhyay e Maulik (2001)	DB, Dunn, Generalizações do Dunn	I- Index	VGA-Clustering	AD_5_2, AD_10_2, AD_4_3	Iris e Breast Cancer
Bandyopadhyay e Saha (2008)	I-Index	Sym-Index, PS-Based	VGAPS-Clustering	9 bases de dados artificiais	Iris, Breast Cancer, Newthyroid, Glass, LiverDisorder
Boudraa (1999)	Coeficiente de Partição (CP), Coeficiente de Entropia (CE), Xie-Beni (XB), Fukuyama-Sugeno (FS), MH (DUBES, 1987), CJ (COGGINS, JAIN, 1985)	Bcrit	Fuzzy C-Means	-	IMOX, Iris, Starfield, Speaker, Lenses, Balance
Chou, Su e Lai (2004)	Dunn, CP, XB, DB, Gath-Geva (GATH; GEVA, 1989), CE	CS	Fuzzy C-Means e Gustafson-Kessel	5 bases de dados artificiais	-
Ding e Harrison (2007)	Silhouette, DB, Calinski-Harabasz, Dunn, C-Index, K (KRZANOWSKI; LAI, 1985) e H (HARTIGAN, 1978)	Relational Visual Cluster Validity (RVCV)	K-Means, PAM, Hierarchical, SOM, Neuro-Gas	-	Wisconsin Breast Cancer, Iris
Fernandèz e Corbacho (2010)	DB, Dunn, PBM, Silhouette	Negentropy	Algoritmo Genético	Gaussians 2D, 3D	Iris, Wisconsin Breast Cancer e Wine
Gurrutxaga et al (2010)	Calinski-Harabasz, C-Index, Gamma, DB, Dunn e Rand Index	COP	Single-Linkage, Average-Linkage, Complete-Linkage, SEP	3 Bases de dados artificiais	Ecoli, Ionosphere, Iris, Page-blocks, Segmentation, Vehicle, Yeast

Tabela 7 - Revisão de Literatura sobre Medidas de Qualidade para clusterização

(continua)

Artigo	Medidas Empregadas	Propõe nova Medida	Algoritmos aplicados	Bases de dados	
				Artificiais	Reais
Hashimoto, Nakamura e Miyamoto (2008)	XB, FS, DB e suas versões kernelizadas	-	Hard C-Means Clustering(HCM) e sFCM	4 adaptações de 1 base artificial	-
He et al (2008)	CP, CE, XB, FS e o Kwon-Index (KWON, 1998)	-	Fuzzy C-Means	X30, Bensaid	China Statistic Almanac (1999-2003)
Hu et al (2011)	CP, CE, MPC (DAVE, 1996), OS, PACES (WU; YANG, 2005), CO, COR (ZALIK, 2010), W (ZHANG etal, 2008)	MPO	Fuzzy C-Means	A3, B6, C2, D4	Iris, Wine
Kim, Lee e Lee (2004)	PC, PE, XB, FS, Kwon-Index, CWB (REZAEE,1998) e Bcrit	OS	Fuzzy C-Means	Bensaid, X30	Starfield, Iris e Butterfly
Kim e Ramakrishma (2005)	SV, SDunn, XB, DB e I-Index	SV*,SDunn*,XB*, XB**, DB* e DB**	K-Means	DB1, DB2, DB3, DB4, DB5, X30, Bensaid	Starfield, Iris e NE (endereços postais de New York, Philadelphia e Boston)
Kim, Yoo e Ramakrishna (2004)	SV, SDunn, XB, DB e I-Index	Adaptação para dados com alta dimensionalidade (SVHD, IHD, SDHD, XBHD, DBHD)	PROCLUS	T1, T2, T3, T4, T5, T6, T7, T8	-
Kovács et al (2006)	Dunn, DB, SD e S_Dbw	-	K-Means, DBSCAN	1 Base de dados artificial	-
Kwon (1998)	XB	Kwon-Index	Fuzzy C-Means	3 Bases de dados artificiais	-
Maulik e Bandyopadhyay (2002)	DB, Calinski-Harabasz, Dunn, e I-Index	-	K-Means, Single Linkage e Simulated Annealing (AS)	2D (AD_10_2, AD_4_3N, e AD_2_10. AD_10_2) 3D (AD_4_3N)	Crude Oil (JOHNSON; WICHERN, 1982) e Wisconsin Breast Cancer
Saad e Alimi (2012)	PCAES (WU; YANG, 2005), SVI (TSEKOURAS; SARIMVEIS, 2004) e XB	Modified Partition Coefficient And Exponential Separation (MPCAES)	Fuzzy C-Means	1 Base de dados artificial	Iris
Mok et al (2012)	PC, PE, XB e FS	W	Fuzzy C-Means	6 Bases de dados artificiais	Iris, Wine e Breast Cancer

Tabela 7 - Revisão de Literatura sobre Medidas de Qualidade para clusterização

(continua)

Artigo	Medidas Empregadas	Propõe nova Medida	Algoritmos aplicados	Bases de dados	
				Artificiais	Reais
Moraes, Evsukoff e Ebecken (2005)	PBM	-	Unsupervised Vector Quantiser (UVQ) FCM, SOM (HEXAGONAL) SOM (RETANGULAR)	-	Base de dados Sísmicos da UFRJ
Pakhira, Bandyopadhyay e Maulik (2004)	DB, XB e Dunn	PBM	Fuzzy C-Means	Circular 5 2, Circular 6 2, Elliptical 10 2 and Spherical 4 3	Iris, Wisconsin Breast Cancer, Crude Oil e Kalazaar (MITRA; PAL, 1995)
Rendón et al (2011)	F-Measure (LARSEN; AONE, 1999), NMIMeasure (STREHL, GHOSH, 2002), Entropy (SHENNON, 1948), BIC (RAFTERY, 1986), Calinski-Harabasz, DB, Silhouette, Dunn, NIVA (RENDON et al, 2002)	-	k-Means, Bissections K-means (BKM)	12 Bases de dados artificiais	-
Rezaee (2010)	PC, PE, XB, Kwon-Index e SC (KIM et al, 2004)	VSC	Fuzzy C-Means e Gustafson-Kessel	6 Bases de dados artificiais	Iris e Mammographic mass (Institute of Radiology of the University Erlangen-Nuremberg 2003-2006)
Sadr e Momtaz (2011)	VPC, PXIE, VPE, VFS (BEZDEK, 1998)	CABRP	K-Means e Fuzzy C-Means	8 Bases de dados artificiais	-
Saha e Bandyopadhyay (2009)	DB, Dunn, Dunn gener., OS, I-Index, XB, FS, Kwon-Index e SV	SymDB, SymDunn, SymDGen, SymOS, SymI, SymXB, SymK, SymSV	K-Means, Fuzzy C-Means GAPS (Bandyopadhyay; Saha, 2007)	Ring_3_2, Ellip_2_2, Mixed_5_2, Line_2_2, Sph_6_2, Bensaid	Iris, Breast Cancer, Glass, Lung Cancer e Wine
Saitta, Raphael e Smith (2007)	I-Index, Silhouette, DB, Dunn	Score Function (SF)	K-Means	4 Bases de dados artificiais	Iris e Wine
Shim, Chung e Choi (2005)	13 medidas (MILLIGAN; COOPER, 1985) e S_Dbw	-	K-Means	-	(MILLIGAN, 1995)
Silva e Gomide (2004)	PC, PE, XB, FS, AD (GEVA,1999)	-	Participatory Learning (AP) e Gustafson-Kessel	3 Bases de dados artificiais	Iris
Tasdemir e Merényi (2011)	Dunn, DB, CDBw, Silhouette, Calinski-Harabasz, PBM	Conn_Index	K-Means	2 Bases de dados artificiais	Wisconsin Breast Cancer, Iris e Wine

Tabela 7 - Revisão de Literatura sobre Medidas de Qualidade para clusterização

(conclusão)

Artigo	Medidas Empregadas	Propõe nova Medida	Algoritmos aplicados	Bases de dados	
				Artificiais	Reais
Wang e Zhang (2007)	CP, CE, SC, P (CHEN; LINKENS, 2004), FS, XB, Kwon-Index, T (TAN;SUN;SUN, 2005), Gath-Geva, PCAES, SVI, CWB, WSJ (SUN; WANG; JIANG, 2004), OS, PBM, SCG (BOUGUESSA, WANG, 2004), Dunn Generalizado, BS (CHO; YOO, 2005) e IG (RHEE; OH, 1996)	-	Fuzzy C-Means	DataSet_3_3, DataSet_4_3, DataSet_4_2, DataSet_4noise, DataSet_5_2, DataSet_6_2, DataSet_10_2 e DataSet_15_2	Iris, Wisconsin Breast Cancer (WBCD), Wisconsin Diagnostic Breast Cancer(WDBC), Wine, Bupa Liver Disorder, Butterfly
Wu e Yang (2005)	FHV (Gath-Geva, 1989), CP, CE, MPC (DAVE, 1996), FS, XB, SC,	PCAES	Fuzzy C-Means	5 bases de dados artificiais	Iris, Glass e Vowel
Yang e Lai (2011)	CP, CE, MPE, FS, XB, FHV,	AM-PCM	Fuzzy C-Means	8 Bases de dados artificiais	-
Yang, Song e Cao (2006)	XB, FS, CP, CE, FHV, PD (GATH;GEVA, 1989)	Vapnik–Chervonenkis-bound (VC-Bound)	Deterministic Anneling (DA)	Noisy Data, Overlapped Data, Unbalanced Data, Synthetic Data,	Iris, Butterfly, Wine
Yang e Wu (2001)	CP, CE, FS, XB, DB	Partition Separation (PS)	Fuzzy C-Means	1 Base de dados artificial	Butterfly
Zahid, Limouri e Essaid (1999)	CP, CE, XB, FS	Separation Compactness (SC)	Hard C-Means Clustering (HCM) e Fuzzy C-Means	4 Bases de dados artificiais	Iris
Zhang et al (2008)	CP, CE, MPC, FS, XB, Kwon-Index, FHV,	w e PBMF	Fuzzy C-Means	Dataset_3_3, DataSet_4_3, DataSet_4_2, DataSet_5_2, DataSet_6_2, DataSet_10_2 e DataSet_15_2	Wine, Liver Disorder, Iris, Wisconsin Diagnostic Breast Cancer, Bupa Liver Disorder, Butterfly, Wisconsin Breast Cancer
Zhao, Xu e Franti (2009)	Calinski-Harabasz, Ball-Hall, MSE, Hartigan, Xu, Dunn, DB, XB, Bayes Information Criterion (BIC), Silhouette	WB	Random Local Search Algorithm	4 Bases de dados artificiais	Iris e Breast Cancer Wisconsin
Palhano (2013)	CP, CE, XB, Dunn e C-Index	-	Fuzzy C-Means RCP, URCP, DBSCAN, SACA, A²CA e ABC	-	Iris, Breast Cancer Wisconsin, Wine e Bacias Hidrográficas da região de Criciúma

Fonte: Do Autor.

5.1.2 Bases de Dados Utilizadas

Na etapa de escolha das bases de dados que seriam utilizadas, definiram-se quatro bases de dados para analisar o comportamento dos algoritmos e das medidas de qualidade com relação a cada uma. Foi definido também que as bases de dados teriam de ser reais, pois bases de dados artificiais geralmente possuem dados perfeitos e consequentemente geram partições exatas, impossibilitando de analisar como o algoritmo e as medidas se comportam com dados imperfeitos. Foram escolhidas as bases de dados *Iris*, *Wine*, *Breast Cancer Wisconsin* e das bacias hidrográficas.

As três primeiras são conjuntos de dados obtidos a partir do UCI *Machine Learning Repository*[5], o qual consiste em um repositório de dados reais administrado pelo Departamento de Informação e Ciência da Computação da Universidade da Califórnia, localizada na cidade de Irvine. Foram escolhidos dados do UCI, pelo fato de conter bases de dados gratuitas e bem aceitas por pesquisadores da área de aprendizado de máquina, como visto anteriormente pela tabela 7, onde muitos trabalhos fizeram uso destas bases.

A quarta base de dados utilizada no trabalho refere-se a dados locais da região de Criciúma. Essa base de dados possui registros das bacias hidrográficas afetadas por poluentes provenientes da extração do carvão mineral na região (GTA, 2009), a base foi fornecida pelo professor Dr. Carlyle Torres de Bezerra Menezes do curso de Engenharia Ambiental da UNESC.

5.1.2.1 Base de dados da Iris

A base de dados da *Iris* refere-se às flores da família *Iridaceas*, com predominância nas espécies *Iris-Setosa*, *Iris-Versicolor* e *Iris-Virginica* (figura 11). O conjunto de dados é composto por 150 registros, distribuídos em 50 registros para cada espécie e constituída por quatro atributos: largura da sépala, largura da pétala, comprimento da sépala e comprimento da pétala.

[5] Repositório de dados disponível em: http://archive.ics.uci.edu/ml/

Figura 9- Flores Iris-Setosa, Iris-Versicolor e Iris-Virginica

5.1.2.2 Base de Dados *Breast Cancer Wisconsin*

Breast Cancer é um conjunto de dados que disponibiliza informações sobre pacientes com câncer de mama obtidas a partir dos Hospitais da Universidade de Wisconsin, ambas localizadas em Madison nos Estados Unidos. A base de dados é composta por 699 registros, porém foram retirados 16 registros que não estavam coerentes, restando 683 registros, sendo composta por 10 atributos numéricos que são métricas que avaliam as características das células analisadas, onde as características de cada atributo podem ser visualizadas pela tabela 8.

As amostras foram coletadas periodicamente pelo Dr. William H. Wolberg a partir dos seus casos clínicos, resultando em um agrupamento cronológico dos dados entre janeiro de 1989 a novembro de 1991.

Tabela 8 - Descrição base de dados *Breast Cancer Wisconsin*

Atributo	Descrição	Valores
id	Identificador do Registro	Núm. Inteiros
clump_thickness	Camadas de agrupamento de células	1-10
uniformity_ofcell_size	Variação de tamanho das células	1-10
uniformity_ofcell_shape	Variação das formas das células	1-10
marginal_adhesion	Nível de perda de aderência das células	1-10
single_epithelial_cell_size	Tamanho das células epiteliais	1-10
bare_nuclei	Identificação de núcleos não envolvidos pelo citoplasma	1-10
bland_chromatin	Nível de textura uniforma no núcleo das células	1-10
normal_nucleoli	Identificador do nucléolo	1-10
Mitoses	Nível de multiplicação celular (mitose)	1-10
class	Identificador: 2 para “benigno” e 4 para “maligno”	2-4

Fonte: Adaptado de UCI *Machine Learning Repository* (2013).

5.1.2.3 Base de Dados *Wine*

A base de dados *Wine* corresponde do resultado de análises químicas para determinar a origem de vinhos cultivados em uma região da Itália, porém provenientes de diferentes cultivadores de videiras. Sendo composta por 178 registros e 13 atributos (*alcohol, malic_acid, ash, alcalinity_of_ash, magnesium, total_phenols, flavanoids, nonflavanoid_phenols, proanthocyanins, color_intensity, hue, od280/od315_of_diluted_wines* e *proline)* que envolvem informações técnicas e específicas de características químicas e visuais de cada tipo de vinho.

5.1.2.4 Base de dados Bacias Hidrográficas

Esta base de dados corresponde a índices que trazem dados com relação a qualidade da água de três bacias hidrográficas da região de Criciúma: Araranguá, Tubarão e Urussanga.

A região de Criciúma é conhecida por sua abundância em relação à extração de carvão mineral, o que traz para a cidade e municípios próximos, geração de emprego e renda em minas de carvão, porém a extração do mesmo acarreta vários problemas ambientais, ressaltando o de recursos hídricos.

Considerando isso, foi criado um grupo denominado Grupo Técnico de Assessoramento (GTA) para monitorar as empresas responsáveis pela contaminação das bacias e solos e recuperar as áreas degradadas pela contaminação do rejeito da mineração de carvão.

O GTA divulga relatórios de monitoramento dos indicadores ambientais das bacias hidrográficas que tem por objetivo verificar a qualidade dos recursos hídricos da região carbonífera catarinense, a fim de avaliar a eficácia dos trabalhos de recuperação ambiental executados por empresas do ramo carbonífero (GTA, 2009). Por meio do quarto Relatório no ano de 2010 foi disponibilizado os dados com indicadores ambientais, somando 1723 registros e 20 atributos identificados pela tabela 9.

Na aplicação dos algoritmos foi selecionada apenas a bacia Araranguá, pois é considerada a que contém mais registros na base de dados (954) e a que possui um índice maior de poluição com relação as demais.

Tabela 9 - Base de dados Bacias Hidrográficas

Atributo	Descrição	Valor
id	Atributo identificador do registro.	Número inteiro
id_cab	Identificador do ponto de monitoramento.	Número inteiro de 1 até 140
id_bacia	Atributo identificador da bacia hidrográfica onde o registro foi coletado.	1 para Araranguá, 2 para Tubarão e 3 para Urussanga
campanha	Identificador da campanha de monitoramento.	Número inteiro de 1 até 20
data	Atributo que faz referência a data de coleta do registro.	Data
ph	Grandeza físico-química conhecida como "potencial hidrogeniônico". É um valor entre 0 e 14 que indica se uma solução qualquer é ácida (*pH* < 7), neutra (*pH* = 7), ou alcalina (*pH* > 7). A faixa para o *pH* recomendável está entre 6 a 9.	Número decimal
vazão	Volume de água ou um fluido qualquer que passa, em uma determinada unidade de tempo, por meio de uma superfície.	Número decimal
acidez	Quantidade de ácido necessária para titular uma amostra a um determinado *pH*	Número decimal
cond	Condutividade elétrica da água, ou seja, capacidade da água conduzir corrente elétrica.	Número decimal
so, al, fe, mn	Respectivamente representam as concentrações de sódio, alumínio, ferro e manganês que estão presentes nas águas analisadas.	Número decimal
c_acidez	Carga de acidez. Calculada pela multiplicação da vazão pela concentração de acidez. Possui relação direta com a carga de contaminantes presentes na água.	Número decimal
c_fe, c_al, c_so, c_mn	Respectivamente representam as cargas de ferro, alumínio, sódio e manganês presentes nas águas.	Número decimal
precip	Precipitação regional (chuva) obtidos em estações meteorológicas das empresas carboníferas da região.	Número decimal
id_estacao	Atributo identificador da estação de monitoramento.	Número inteiro de 1 a 14

Fonte: Adaptado de GTA (2009).

Após pré-processar as quatro bases de dados para que ficassem de acordo com os algoritmos de clusterização e retirar os dados incoerentes ou faltantes, as bases de dados foram submetidas as aplicações nos algoritmos selecionados.

5.1.3 Seleção das medidas de qualidade e algoritmos utilizados

Dentre os algoritmos de clusterização da Shell Orion, foram selecionados sete algoritmos para a aplicação das quatro bases de dados citadas anteriormente, sendo estes: FCM, RCP, URCP, DBSCAN, SACA, A²CA e ABC.

Com relação as medidas de qualidade analisadas, foram consideradas cinco medidas: Dunn, C-Index, Xie-Beni, Coeficiente de Partição e Coeficiente de Partição Entrópica, pois pela tabela 7 é visível a ampla aplicabilidade destas medidas em trabalhos da área, outro fator relevante foi o fato das mesmas já estarem implementadas nos algoritmos da Shell Orion, sendo nos algoritmos FCM, RCP e URCP implementadas as medidas CP, CE e Xie-Beni e nos algoritmos DBSCAN, SACA, A²CA, ABC as medidas Dunn e C-Index.

5.1.3.1 Significado dos Valores encontrados pelas medidas de qualidade

No Capítulo 3 na parte da pesquisa realizada sobre as medidas de qualidade foi mostrado o significado de cada medida, incluindo as medidas que foram usadas no trabalho, contudo será apresentado novamente o que representa os valores encontrados por cada medida:

a) **Coeficiente de partição:** no intervalo de [1/*C*, 1], quanto mais próximo de um o valor da medida, mais compactos e bem separados serão os *clusters*;

b) **Coeficiente de partição entrópica:** esta medida é o contrário do coeficiente de partição, onde, quanto mais próximo de zero melhor será a qualidade dos *clusters*;

c) **Dunn:** esta medida encontra *clusters* compactos e separados quando valores maximizados são encontrados pela mesma;

d) **C-index:** diferentemente das medidas anteriores a C-Index avalia a coesão de cada *cluster* e não a separação de todos os *cluster*, onde no intervalo de [0,1], quanto mais próximo de zero, mais bem compacto é o *cluster* avaliado.

5.1.4 Aplicação dos Algoritmos

5.1.4.1 Aplicação do Algoritmo FCM

O algoritmo FCM é uma adaptação do tradicional *K-Means*, sendo sua versão final proposta por James C. Bezdek (BEZDEK et al, 2005, tradução nossa).

O FCM é considerado um dos primeiros algoritmos a utilizar o elemento *fuzzificador*, sendo este o responsável pelo grau de fuzzificação entre os elementos, no qual, quanto maior seu valor, mais relacionados serão os dados (CROTTI JUNIOR et al, 2012).

Os parâmetros de entrada do algoritmo são: quantidade de *clusters*, parâmetro de fuzzyficação (fixo definido como 2), quantidade de iterações e taxa de erro (fixo definido como 0,00001), os parâmetros escolhidos para cada experimento realizado pelo FCM pode ser visualizado pela tabela 10.

Tabela 10 - Escolha dos Parâmetros de Entrada FCM

Algoritmo FCM	**Bases de Dados**			
	Iris	***Wine***	***Breast Cancer Wisconsin***	***Bacias Hidrográficas***
1 º Experimento				
Quantidade de Clusters	2	2	2	2
Quantidade de Iterações	12	16	17	34
2º Experimento				
Quantidade de Clusters	3	3	3	3
Quantidade de Iterações	32	30	44	28
3º Experimento				
Quantidade de Clusters	4	4	4	4
Quantidade de Iterações	67	77	96	95

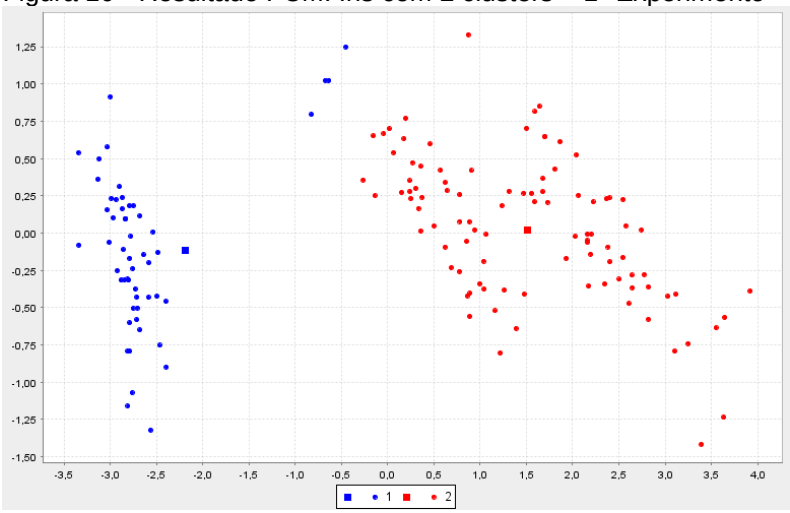
Fonte: Do Autor.

O algoritmo FCM não identifica de forma automática o número ideal de *clusters*, tendo que o usuário definir essa informação como parâmetro de entrada. Portanto, para as quatro bases de dados, mesmo em algumas se sabendo previamente o número ideal de *clusters*, foram realizados três experimentos com dois, três e quatro *clusters,* respectivamente. Com isso, pode-se verificar as variações dos valores das medidas de qualidade para que por meio das mesmas, fosse possível identificar o número ideal de *clusters* presentes no conjunto de dados.

5.1.4.1.1 Base de dados Iris

O resultado da clusterização para o primeiro experimento (dois *clusters*) foi de 54 elementos para o *cluster* um (36% dos dados) e de 96 elementos para o *cluster* dois (64% dos dados), podendo ser observado pela figura 11 o particionamento gerado.

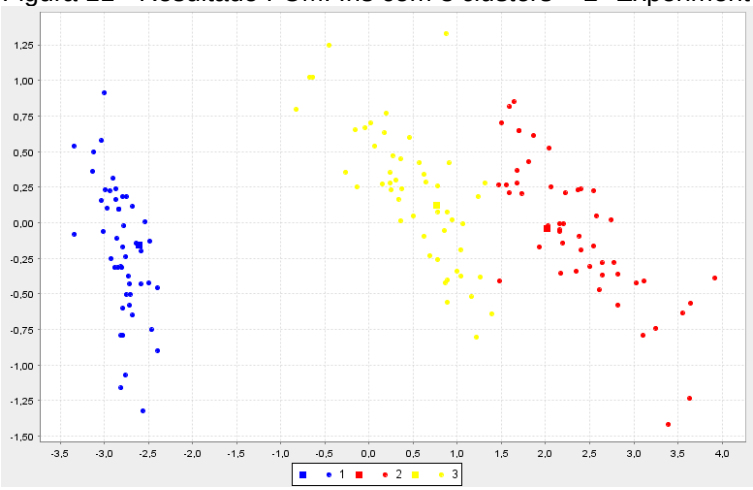
Figura 10 - Resultado FCM: Iris com 2 clusters – 1º Experimento



Fonte: Do Autor.

No segundo experimento do FCM (três *clusters*), o resultado do particionamento foi de três *clusters*, sendo 50 elementos para cada *cluster*, equivalendo a 33,33% dos dados em cada *cluster* (figura 12).

Figura 11 - Resultado FCM: Iris com 3 *clusters* – 2º Experimento

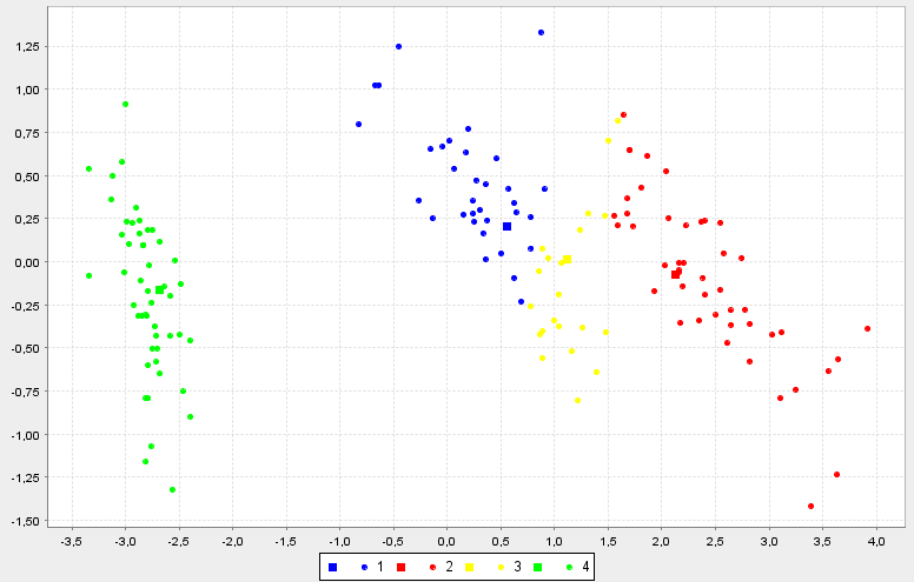


Fonte: Do Autor.

No terceiro experimento do FCM (quatro *clusters*), o resultado da clusterização para o *cluster* um foi de 33 elementos (22,00% dos dados), do *cluster* dois 46 elementos (30,67%), para o *cluster* três 21 elementos (14% dos

dados) e finalmente o *cluster* quatro com 50 elementos correspondentes a 33,33% dos dados (figura 13).

Figura 12 - Resultado FCM: Iris com 4 *clusters* – 3º Experimento



Fonte: Do Autor.

Com relação as medidas de qualidade quanto ao FCM (tabela 11) pode-se observar que as três medidas (CP, CE e XB) indicaram dois, como sendo o número ideal de *clusters* para a base de dados da Íris.

Tabela 11 - Medidas de qualidade dos experimentos FCM na *Iris*

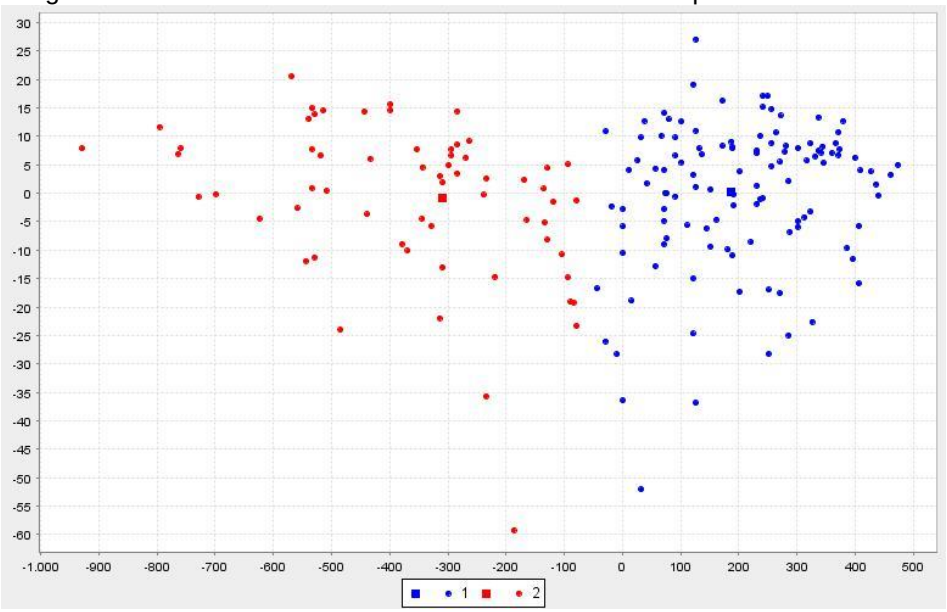
Experimentos	Algoritmo FCM: Núm. de *Clusters*	Coeficiente de Partição (+)	Coeficiente de Entropia (-)	Xie-Beni (-)
1º Experimento	***2***	***0,682402***	***0,493657***	***0,212019***
2º Experimento	**3**	0,544811	0,788098	0,340005
3º Experimento	**4**	0,430202	1,064938	0,315331

Legenda: Função maximizadora (+); Função minimizadora (-).

5.1.4.1.2 Base de dados Wine

A resposta da clusterização nesta base de dados para o primeiro experimento (dois *clusters*) foi de 116 itens para o *cluster* um (65,17% dos dados) e de 62 elementos para o cluster dois que corresponde a 64% dos dados (figura 14).

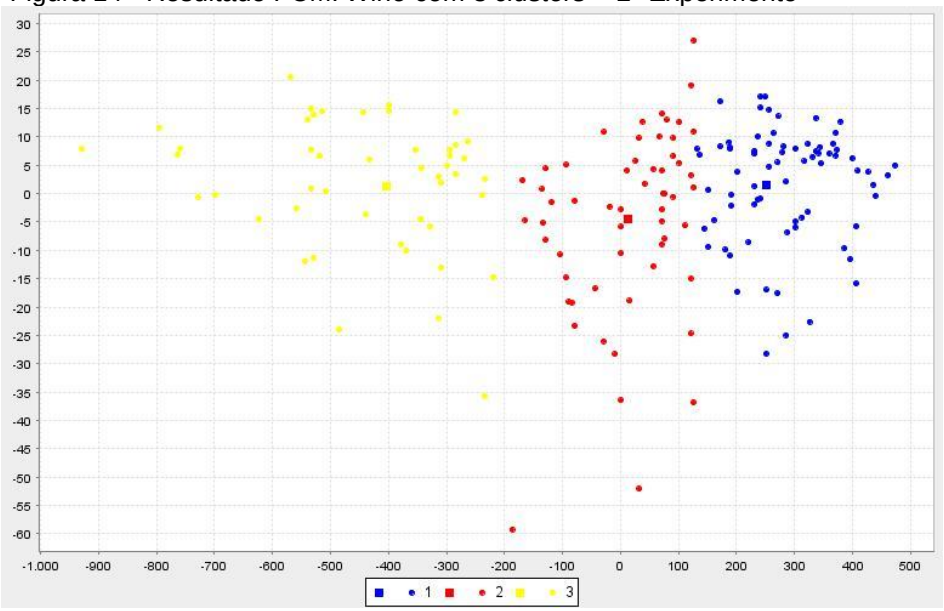
Figura 13 - Resultado FCM: *Wine* com 2 *clusters* – 1º Experimento



Fonte: Do Autor.

No segundo experimento do FCM (três *clusters*), o resultado da particionamento foi de 74 elementos para o *cluster* um (41,57 % dos dados), 57 elementos pertinentes ao *cluster* dois (32,02%) e 47 elementos correspondentes ao *cluster* três (26,40% dos dados), podendo observar-se pela figura 15.

Figura 14 - Resultado FCM: *Wine* com 3 *clusters* – 2º Experimento

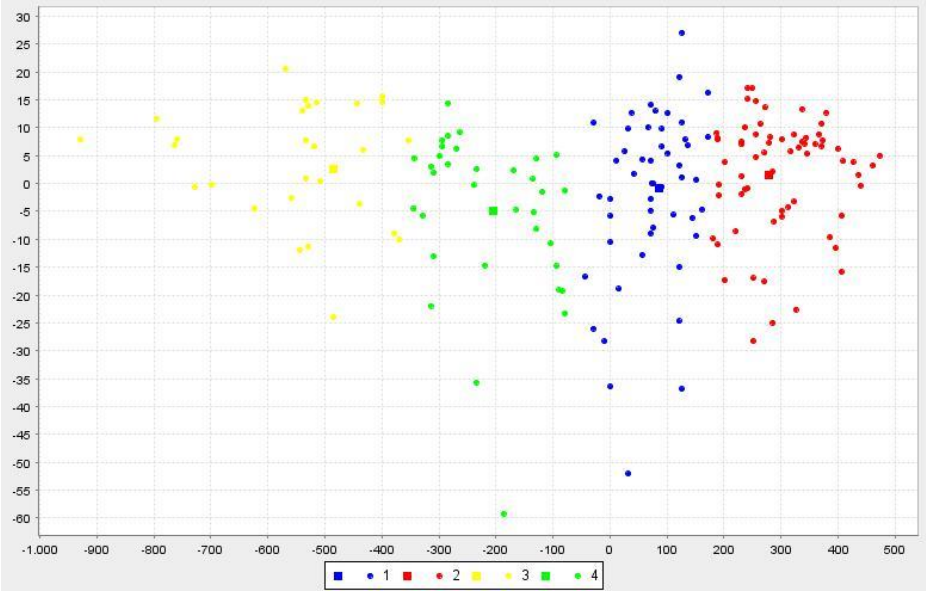


Fonte: Do Autor.

No terceiro experimento (quatro *clusters*) realizado, obteve-se o *cluster* um composto por 50 elementos (28,09% dos dados), o *cluster* dois 66 elementos (37,08%), para o *cluster* três 28 elementos (15,73% dos dados) e finalmente o

cluster quatro com 34 elementos correspondentes a 19,10% dos dados (figura 16).

Figura 15 - Resultado FCM: *Wine* com 4 *clusters* – 3º Experimento



Fonte: Do Autor.

Em relação as medidas de qualidade quanto à base de dados *Wine* no algoritmo FCM (tabela 12) pode-se observar que as três medidas (CP, CE e XB) indicaram dois, como sendo o número ideal de *cluster*, pois a função maximizadora CP obteve o valor mais alto no primeiro experimento e as funções minimizadoras (CE e XB) encontraram valores baixos, no primeiro experimento.

Tabela 12 - Medidas de qualidade dos experimentos FCM na *Wine*

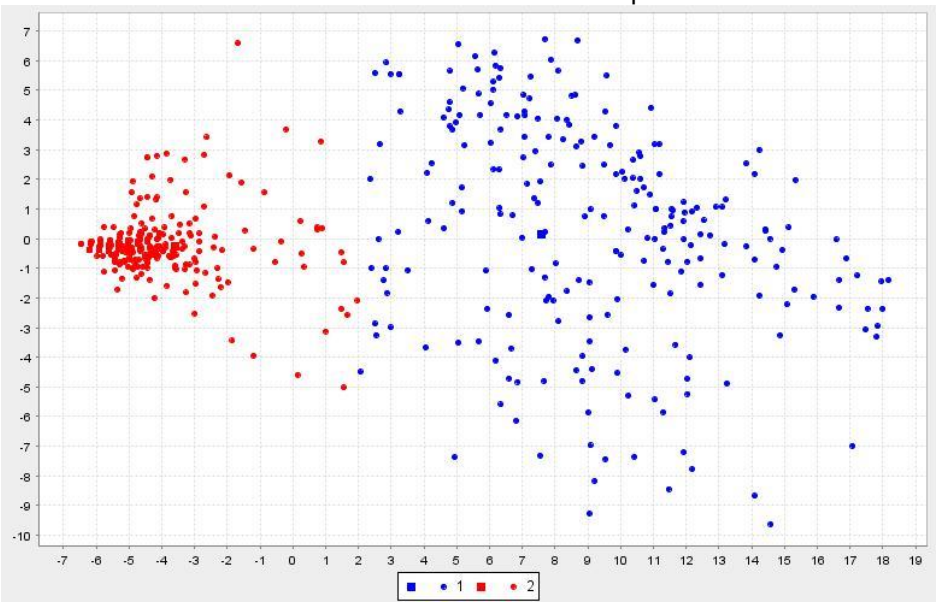
Experimentos	**Algoritmo FCM: Número de *Clusters***	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie-Beni (-)**
1º Experimento	**2**	***0,715227***	***0,445603***	***0,197425***
2º Experimento	**3**	0.5723236	0.745642	0.247785
3º Experimento	**4**	0.507056	0.934491	0.213641

Fonte: Do Autor.

5.1.4.1.3 Base de dados Breast Cancer Wisconsin (BCW)

No primeiro experimento (dois *clusters*) da base de dados BCW no FCM, como resultado do particionamento, obteve-se no *cluster* um 234 elementos (34,26% dos dados) e no *cluster* dois 449 elementos e 65,74% em relação aos dados (figura 17).

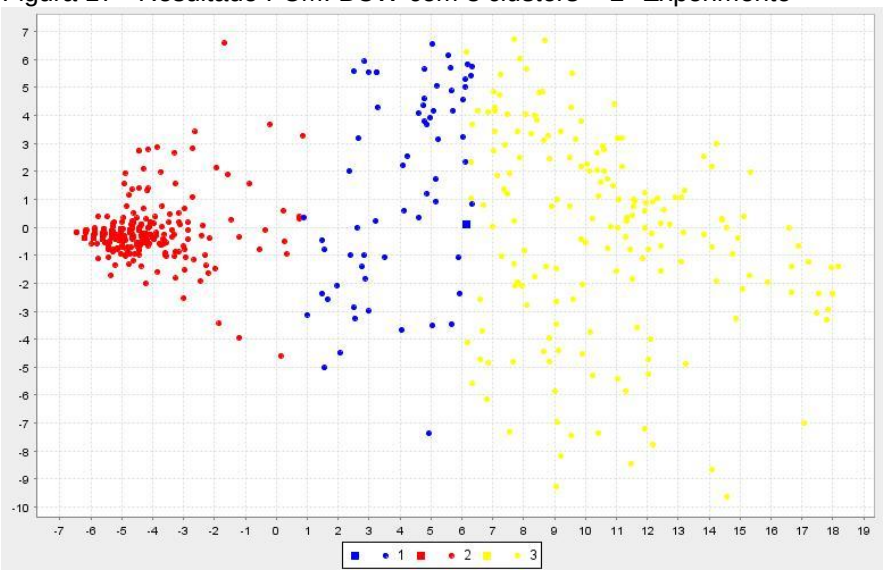
Figura 16 - Resultado FCM: BCW com 2 *clusters* – 1º Experimento



Fonte: Do Autor.

Com relação ao segundo experimento (três *clusters*) do FCM nesta base de dados, a clusterização resultante foi: 64 elementos (09,37%) pertencentes ao *cluster* um, 441 elementos (64,57% dos dados) referentes ao *cluster* dois e 178 elementos (26,06%) pertinentes ao *cluster* três (figura 18).

Figura 17 - Resultado FCM: BCW com 3 *clusters* – 2º Experimento

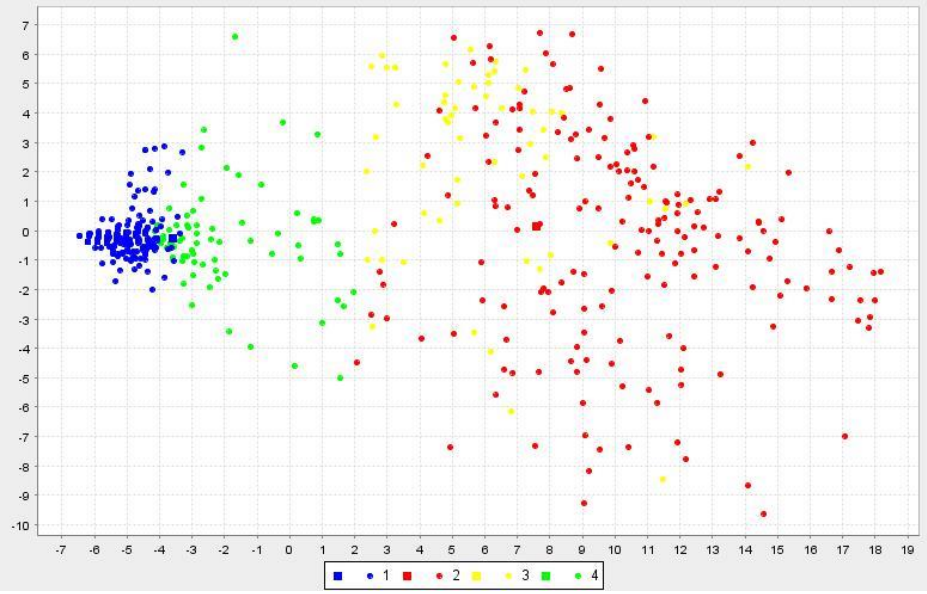


Fonte: Do Autor.

A fim de executar a base de dados BCW com quatro clusters, o resultado do terceiro experimento pode ser visualizado pela figura 19, onde encontrou-se: 371 elementos (54,32% dos dados) pelo *cluster* um, 174 elementos (25,48%) pertencentes ao *cluster* dois, 60 elementos (08,78% dos dados)

pertinentes ao *cluster* três e 78 elementos representando 11,42% dos dados pelo cluster quatro.

Figura 18 - Resultado FCM: BCW com 4 *clusters* – 3º Experimento



Fonte: Do Autor.

Os valores encontrados pelas medidas de qualidade com relação à BCW no FCM podem ser observados pela tabela 12, onde se observa que o Coeficiente de Partição e Coeficiente de Entropia indicam o 1º experimento como sendo o ideal e o Xie-Beni indica o segundo experimento com três *clusters* como sendo o ideal.

Tabela 13 - Medidas de qualidade dos experimentos FCM na *Breast Cancer Wisconsin*

Experimentos	**Algoritmo FCM: Número de *Clusters***	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie-Beni (-)**
1º Experimento	**2**	***0.653785***	***0.525657***	0.290974
2º Experimento	**3**	0.481494	0.889662	***0.212514***
3º Experimento	**4**	0.3268947	1.218800	5.662159

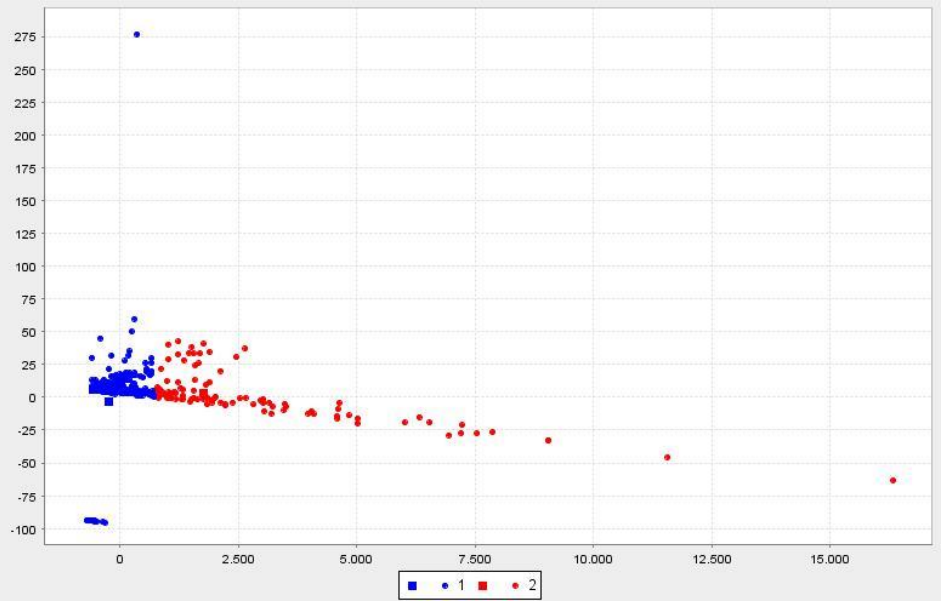
Fonte: Do Autor.

5.1.4.1.4 Base de dados de Bacias Hidrográficas

Para a execução do algoritmo FCM com a base de dados de Bacias Hidrográficas, ao contrário das demais, não foram selecionados todos os atributos. Isso aconteceu, em função de serem atributos com muita variação de valores, portanto escolheram-se os atributos *ph, acidez* e *mg*, que possibilitam constatar que há uma relação entre os valores, pois quando o fator pH é ruim, os valores de manganês e acidez também são alterados.

No primeiro experimento (dois *clusters*) do FCM na base de dados a clusterização gerada pelo algoritmo ficou da seguinte forma: *cluster* um com 848 elementos (88,89% dos dados) e o segundo *cluster* com 106 elementos (11,11%), podendo ser visualizada de forma gráfica pela figura 20.

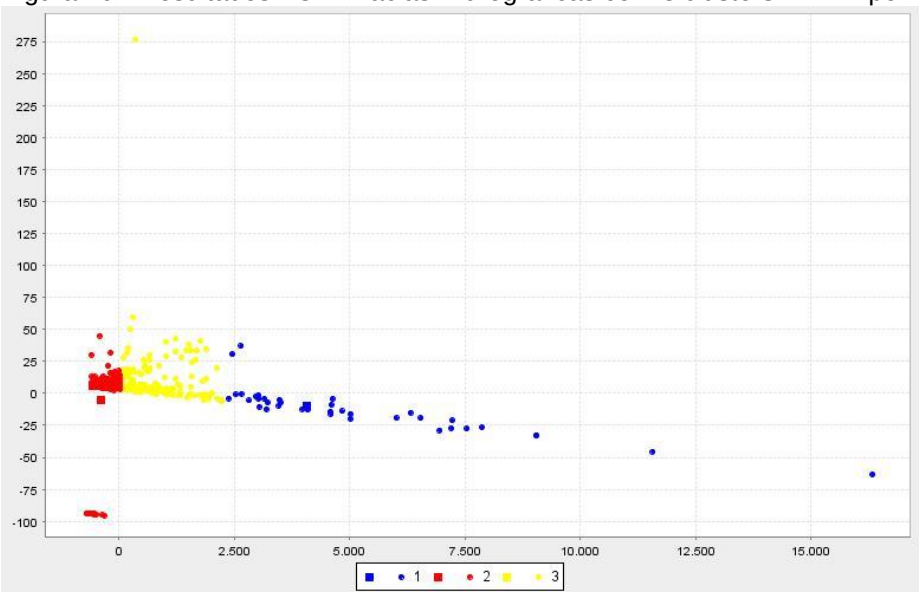
Figura 19 - Resultado FCM: Bacias Hidrográficas com 2 *clusters* – 1º Experimento



Fonte: Do Autor.

No experimento seguinte (três *clusters*) o resultado proveniente do particionamento pode ser visto pela figura 21, no qual encontrou-se: 38 elementos pertencentes ao *cluster* um (3,98% dos dados), 652 elementos pertinentes ao *cluster* dois com predominância de 68,34% dos dados e o terceiro *cluster* com 264 elementos com 27,67% dos dados.

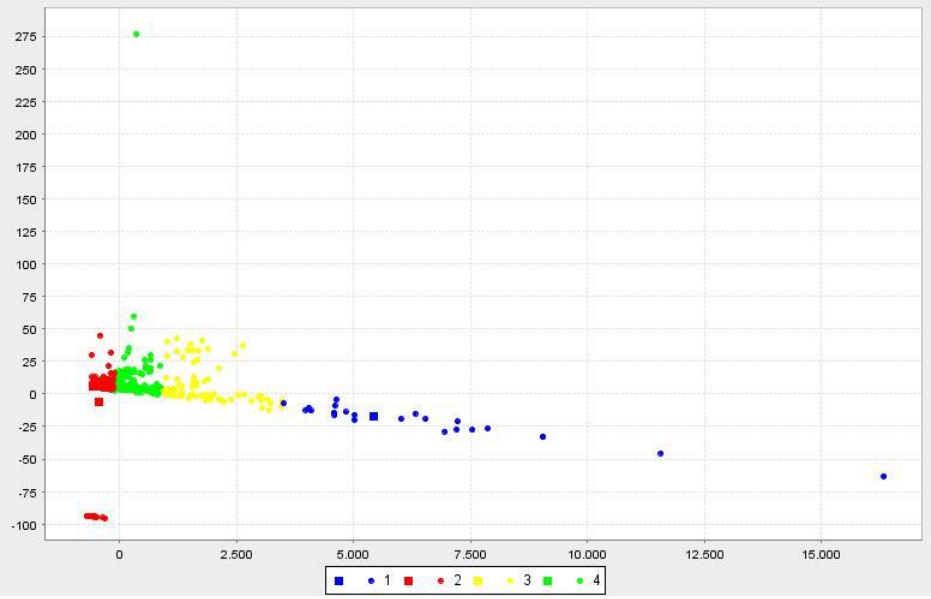
Figura 20 - Resultados FCM: Bacias Hidrográficas com 3 *clusters* – 2º Experimento



Fonte: Do Autor.

Para finalizar as execuções com o algoritmo FCM o resultado do terceiro experimento na base de dados das bacias hidrográficas, foi o seguinte: *cluster* um com 23 elementos (2,41%), *cluster* dois com 593 elementos com predominância de 62,16% dos dados, o terceiro *cluster* encontrou 68 elementos (7,13%) e o *cluster* quatro 270 elementos com (28,30% dos dados) (figura 22).

Figura 21 - Resultado FCM: Bacias Hidrográficas com 4 *clusters* – 3º Experimento



Fonte: Do Autor.

Os valores das medidas de qualidade com relação à base de dados das Bacias hidrográficas podem ser visualizados pela tabela 13, onde é possível visualizar que apenas o Xie-Beni indica a presença de quatro *clusters*, as demais indicam como sendo dois, o número ideal de *clusters*.

Tabela 14 - Medidas de qualidade dos experimentos FCM na base Bacias Hidrográficas

Experimentos	Algoritmo FCM: Número de *Clusters*	Coeficiente de Partição (+)	Coeficiente de Entropia (-)	Xie Beni (-)
1º Experimento	**2**	***0.765836***	***0.383198***	0.162791
2º Experimento	**3**	0.682614	0.561969	0.205993
3º Experimento	**4**	0.6442495	0.682163	***0.094975***

Fonte: Do Autor.

5.1.4.2 Aplicação do Algoritmo RCP

O algoritmo RCP foi criado pelos autores Hichem Frigui e Raghu Krishnapuram no ano de 1996, ele utiliza-se de uma matriz de covariância para

reduzir a interferência de dados ruidosos na clusterização (OLIVEIRA; PEDRYCZ, 2007, tradução nossa).

Os parâmetros de entrada do algoritmo RCP são: quantidade de *clusters*, parâmetro de fuzzyficação (definido o valor 2), quantidade de iterações e abrangência da função de peso (definido como sendo 12), a escolha dos parâmetros com relação aos experimentos pode ser visualizada pela tabela 15 .

Assim como o FCM, o algoritmo RCP não identifica de forma automática o número ideal de *clusters*, então a aplicação das bases de dados foi feita da mesma forma que o FCM.

Tabela 15 - Escolha dos Parâmetros de Entrada RCP

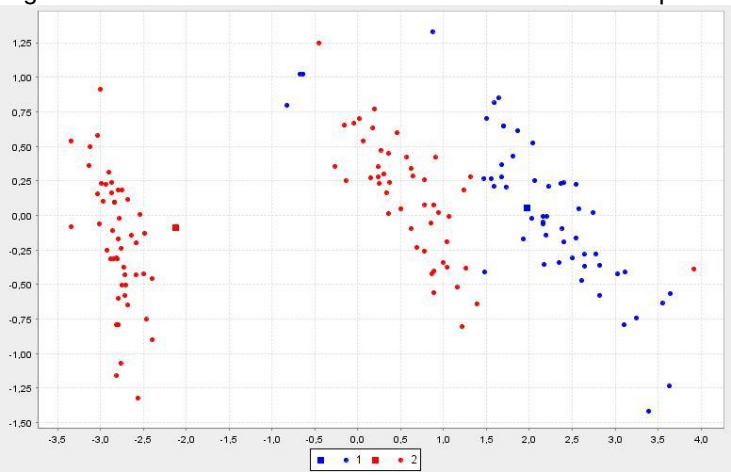
Algoritmo RCP	**Bases de Dados**			
	Iris	***Wine***	***Breast Cancer Wisconsin***	***Bacias Hidrográficas***
1 º Experimento				
Quantidade de Clusters	2	2	2	2
Quantidade de Iterações	51	40	10	10
2º Experimento				
Quantidade de Clusters	3	3	3	3
Quantidade de Iterações	10	10	10	91
3º Experimento				
Quantidade de Clusters	4	4	4	4
Quantidade de Iterações	10	10	10	73

Fonte: Do Autor.

5.1.4.2.1 Base de dados Iris

No primeiro experimento da *Iris* no algoritmo RCP pode ser observada pelo gráfico representado pela figura 23, onde mostra o resultado da clusterização: 53 elementos pertencentes ao *cluster* um (35,33% dos dados) e 97 elementos pertinentes ao segundo *cluster* com predominância de 64,67% dos dados da base.

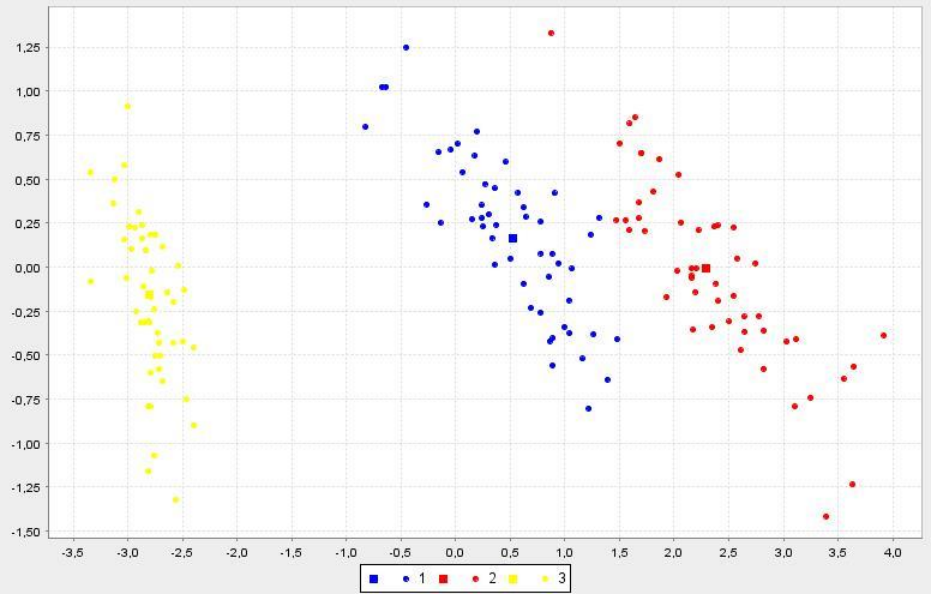
Figura 22 - Resultado RCP: Iris com 2 *clusters* – 1º Experimento



Fonte: Do Autor.

No segundo experimento, como resultado do particionamento obteve-se três *clusters*, cada um com 50 elementos correspondentes a 33,33% dos dados (figura 24).

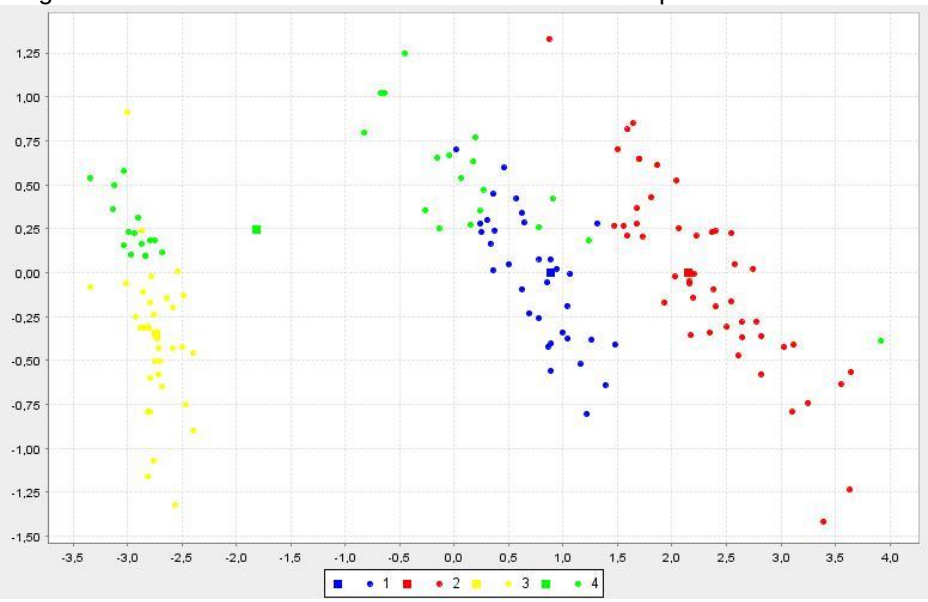
Figura 23 - Resultado RCP: Iris com 3 *clusters* – 2º Experimento



Fonte: Do Autor.

O terceiro experimento foi realizado com quatro *clusters*, resultando a seguinte clusterização: *cluster* um: 33 elementos (22% dos dados), *cluster* dois com 49 elementos (32,67%) e os terceiro e quarto *cluster* ambos encontraram 34 elementos e 22,67% dos dados cada *cluster* (figura 25).

Figura 24 - Resultado RCP: Iris com 4 *clusters* – 3º Experimento



Fonte: Do Autor.

Os valores correspondentes a cada medida com relação ao RCP pode ser visualizada pela tabela 16, onde se observa que neste caso apenas o Xie-Beni indica a presença de dois *clusters*, as demais medidas (CP e CE) indicam três *clusters* como o número ideal.

Tabela 16 - Medidas de qualidade dos experimentos RCP na *Iris*

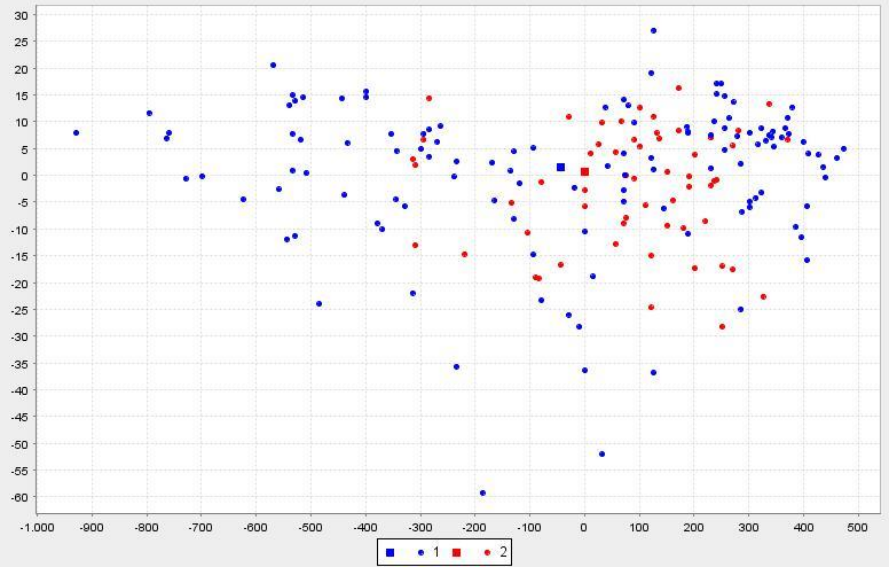
Experimentos	**Algoritmo RCP: Número de *Clusters***	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie Beni (-)**
1º Experimento	**2**	0.651881	0.521349	***0.105121***
2º Experimento	**3**	***0.949697***	***0.121707***	3.435614
3º Experimento	**4**	0.428762	1.054721	0.274151

Fonte: Do Autor.

5.1.4.2.2 Base de dados Wine

No primeiro experimento com o RCP na base de dados *Wine,* a clusterização resultante foi: *cluster* um com 120 elementos (67,42%) e o segundo *cluster* com 58 elementos (32,58%), o gráfico pode ser observado pela figura 26.

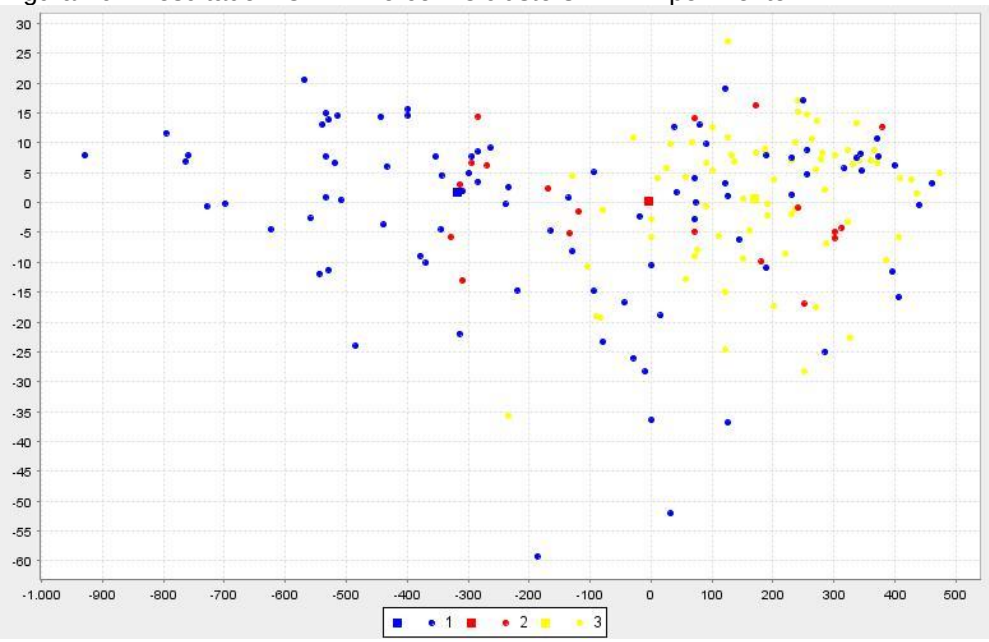
Figura 25 - Resultado RCP: *Wine* com 2 *clusters* – 1º Experimento



Fonte: Do Autor.

No segundo experimento a clusterização com três *clusters* ficou da seguinte forma: *cluster* um com 86 elementos (48,31%), *cluster* dois 19 elementos (10,67%) e *cluster* três 73 elementos (41,01%) (figura 27).

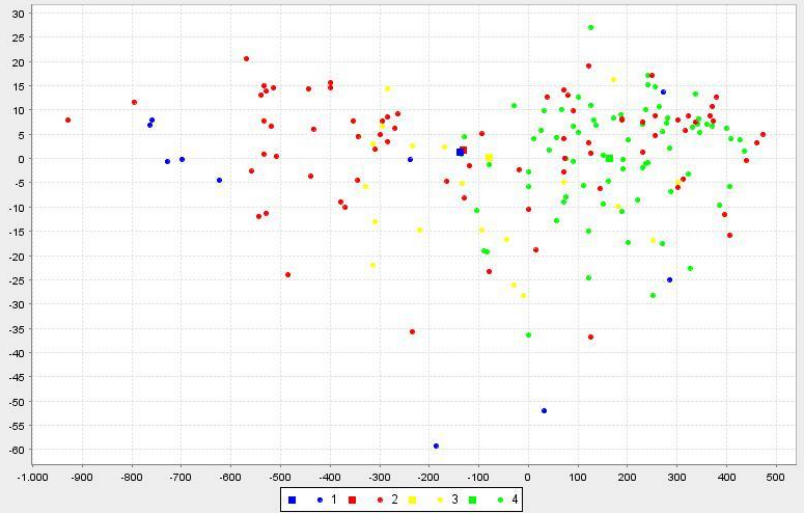
Figura 26 - Resultado RCP: *Wine* com 3 *clusters* – 2º Experimento



Fonte: Do Autor.

Foi efetuado um experimento com quatro *clusters* e obteve-se a seguinte clusterização: *cluster* um 11 elementos (06,18%), *cluster* dois 25 elementos (14,04%), *cluster* três 86 elementos (48,31%) e *cluster* quatro 56 elementos (31,46%) (figura 28).

Figura 27 - Resultado RCP: *Wine* com 4 *clusters* – 3º Experimento



Fonte: Do Autor.

Podem ser visualizados pela tabela 17 os valores obtidos pelas medidas de qualidade na base de dados Wine no RCP. Observa-se pelos valores de CP e CE, que não houve uma boa clusterização, pois ambas não obtiveram valores bons, o melhor valor encontrado foi no primeiro experimento. Já com relação ao Xie-Beni, no segundo experimento foi encontrado o melhor valor para o mesmo, indicando três como o número ideal de *clusters*.

Tabela 17 - Medidas de qualidade dos experimentos RCP na *Wine*

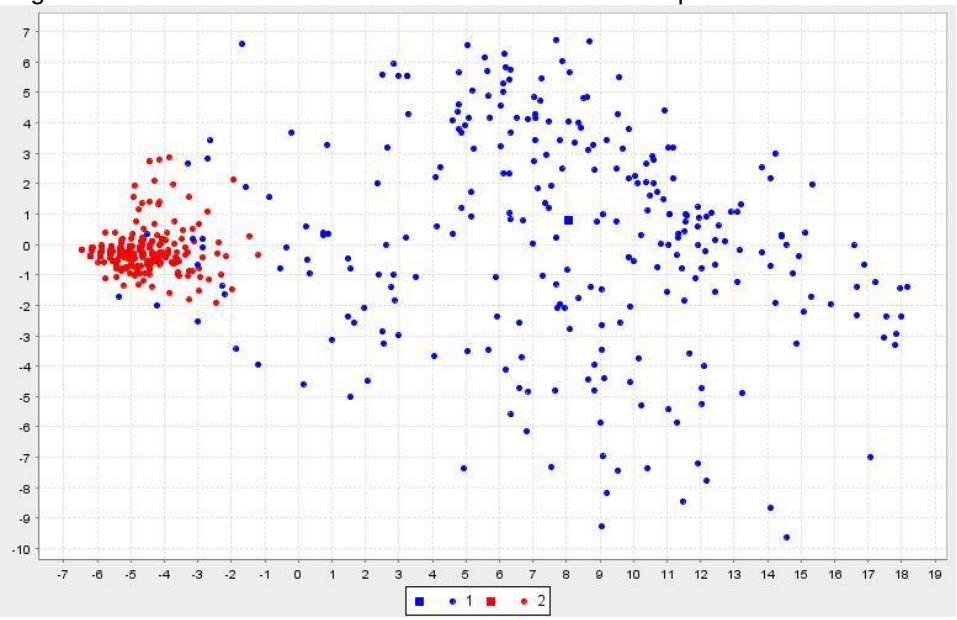
Experimentos	**Algoritmo RCP: Número de *Clusters***	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie Beni (-)**
1º Experimento	**2**	***0.503968***	***0.689162***	0.143280
2º Experimento	**3**	0.385060	1.018044	***0.031201***
3º Experimento	**4**	0.266823	1.348804	0.0349544

Fonte: Do Autor.

5.1.4.2.3 Base de dados Breast Cancer Wisconsin

Na base de dados BCW o resultado encontrado pelo primeiro experimento foi o seguinte: *cluster* um com 270 elementos (39,53%) e *cluster* dois com 413 elementos (60,47%) (figura 29).

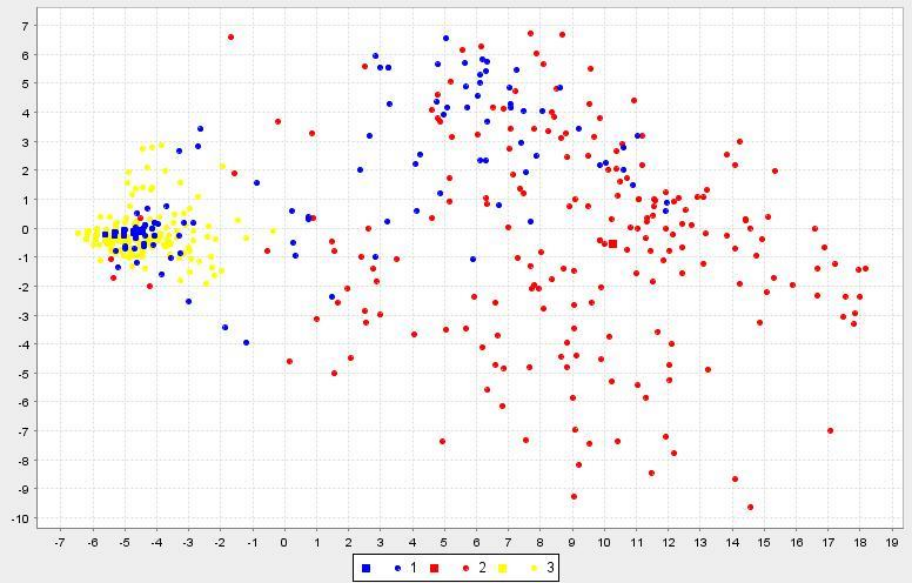
Figura 28 - Resultado RCP: BCW com 2 *clusters* – 1º Experimento



Fonte: Do Autor.

No segundo experimento foram definidos três *clusters*, o resultado da clusterização foi o seguinte: *cluster* um 211 elementos (30,89%), *cluster* dois 199 elementos (29,14%) e o terceiro *cluster* 273 elementos (39,97%) (figura 30)

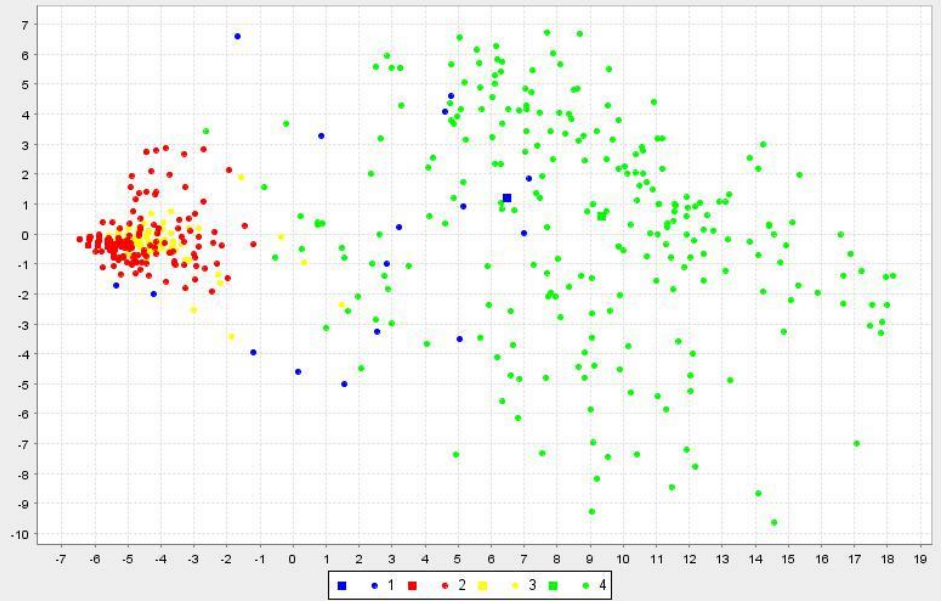
Figura 29 - Resultado RCP: BCW com 3 *clusters* – 2º Experimento



Fonte: Do Autor.

Para o último experimento o resultado obtido do particionamento (figura 31) ficou da seguinte forma: 16 elementos (02,34%) para o *cluster* um, 266 elementos (38,95%) pertencentes ao *cluster* dois, 162 elementos (23,72%) pertinentes ao *cluster* três, com relação ao quarto *cluster* foram encontrados 239 elementos (34,99%).

Figura 30 - Resultado RCP: BCW com 4 *clusters* – 3º Experimento



Fonte: Do Autor.

Os resultados das medidas são visualizados pela tabela 18, onde é possível visualizar que todas as medidas indicam como sendo dois *clusters* o número ideal para a base de dados BCW.

Tabela 18 - Medidas de Qualidade dos experimentos RCP na BCW

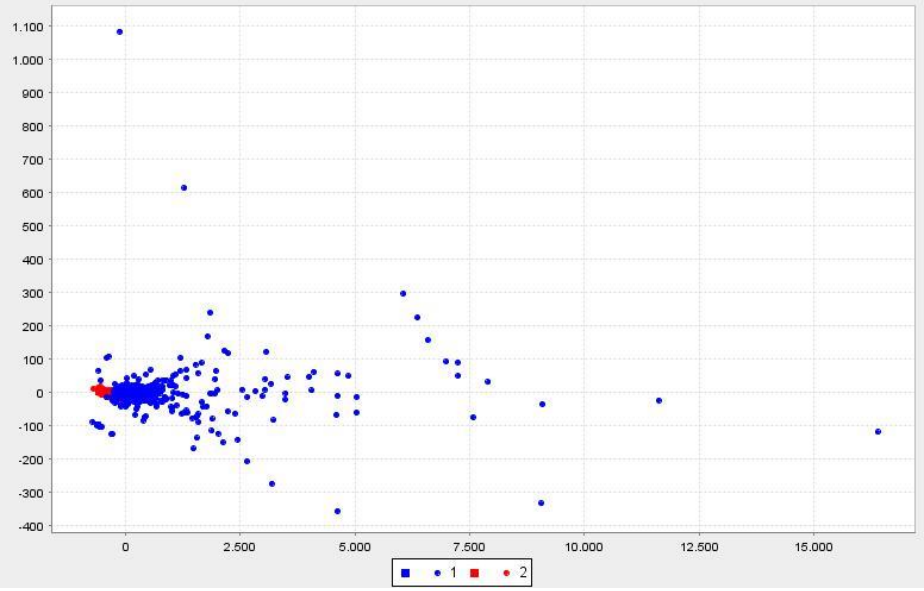
Experimentos	**Algoritmo RCP: Número de *Clusters***	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie Beni (-)**
1º Experimento	**2**	***0.782852***	***0.350546***	***1.456981***
2º Experimento	**3**	0.552086	0.721911	4.100578
3º Experimento	**4**	0.460067	0.957109	1924.4879

Fonte: Do Autor.

5.1.4.2.4 Base de dados de Bacias Hidrográficas

O resultado obtido pela clusterização foi de: 426 elementos (44,65%) pertencentes ao primeiro cluster e 528 elementos (55,35%) referentes ao cluster dois (figura 32).

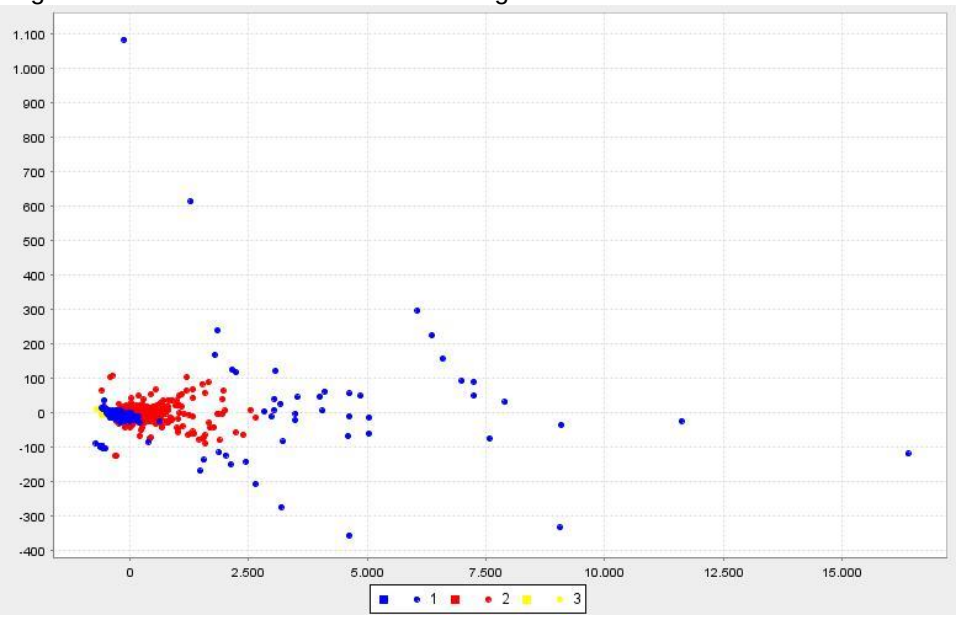
Figura 31 - Resultado RCP: Bacias Hidrográficas com 2 *clusters*



Fonte: Do Autor.

A clusterização resultante para o segundo experimento foi: cluster um 306 elementos (32,08%), cluster dois 289 (30,29%) e o terceiro cluster 359 elementos (37,63%) (figura 33).

Figura 32 - Resultado RCP: Bacias Hidrográficas com 3 *clusters*



Fonte: Do Autor.

Para o último experimento o resultado obtido do particionamento ficou da seguinte forma: 381 elementos (39,94%) para o *cluster* um, zero elementos pertencentes ao *cluster* dois, 214 elementos (22,43%) pertinentes ao *cluster* três e o *cluster* quatro 359 elementos (37,63%), porém o algoritmo para este

experimento não gerou um gráfico para visualizar o mesmo, deste modo não foi possível inseri-lo no texto.

O resultado das medidas com relação a base de dados das bacias hidrográficas é mostrado pela tabela 19, na qual observa-se que todas as medidas indicam como dois o número de *clusters* presentes nesta base de dados.

Tabela 19 - Medidas de qualidade dos experimentos RCP nas Bacias Hidrográficas

Experimentos	**Algoritmo RCP: Número de *Clusters***	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie Beni (-)**
1º Experimento	**2**	***0.699701***	***0.461217***	***12.197569***
2º Experimento	**3**	0.623158	0.647054	22.106230
3º Experimento	**4**	0.623158	0.647065	NaN

Fonte: Do Autor.

5.1.4.3 Aplicação do Algoritmo URCP

O algoritmo URCP assim como o RCP foi desenvolvido por Hichem Frigui e Raghu Krishnapuram no ano de 1996, porém o URCP é um algoritmo que não possui o parâmetro de entrada da quantidade de *cluster.* Portanto, não tem-se a necessidade do usuário saber o número exato de *clusters* presente nos dados, somente é informado o número máximo de *clusters* para o algoritmo considerar (FRIGUI; KRISHAPURAM, 1996, tradução nossa).

Além do número máximo de *clusters* como parâmetro de entrada, no URCP acrescenta-se a abrangência da função de peso[6], no qual todos os parâmetros de entrada empregados nos experimentos desta pesquisa são apresentados na tabela 20.

Tabela 20 - Escolha dos Parâmetros de Entrada URCP

Algoritmo URCP	**Bases de Dados**			
	Iris	***Wine***	***Breast Cancer Wisconsin***	***Bacias Hidrográficas***
1 º Experimento				
Quantidade de Clusters	4	4	4	4
Quantidade de Iterações	3	2	2	2
Taxa de Erro	0,3	0,3	0,3	0,3
Abrangência de Peso	12	12	12	12

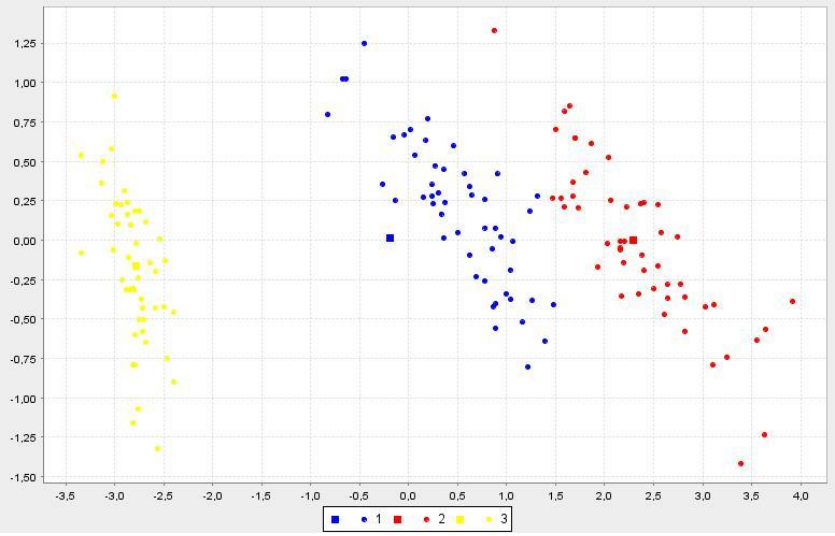
Fonte: Do Autor.

[6] A abrangência da função de peso é estipulada com relação aos dados, onde bases de dados com ruídos deve-se minimizar o valor da abrangência da função de peso e do contrário maximizá-lo.

5.1.4.3.1 Base de dados Iris

Como resultado do experimento de clusterização pelo algoritmo URCP nos dados da Íris encontraram-se três *clusters* que o próprio algoritmo fornece como sendo o ideal. Mesmo aplicando-se mais de uma vez e aumentando o parâmetro da quantidade de *clusters*, o número de *clusters* gerados pelo algoritmo se manteve igual. Com relação aos *clusters* gerados, cada um contém 50 elementos correspondentes a 33,33% dos dados (figura 34).

Figura 33 - Resultado URCP: Iris



Fonte: Do Autor.

O valor das medidas de qualidade encontradas pelo particionamento da base da *Iris* no URCP é mostrado pela tabela 18, onde o CP e CE indicam uma boa clusterização e o Xie-Beni obteve um valor alto indicando uma clusterização ruim. A coluna de experimentos foi retirada da tabela, pois foi realizado apenas um experimento, ao fato do algoritmo gerar automático o número de *cluster.*

Tabela 21 - Medidas de qualidade do experimento URCP na *Iris*

Algoritmo URCP: Número de *Clusters*	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie Beni (-)**
3	0.781311	0.548732	9.894755

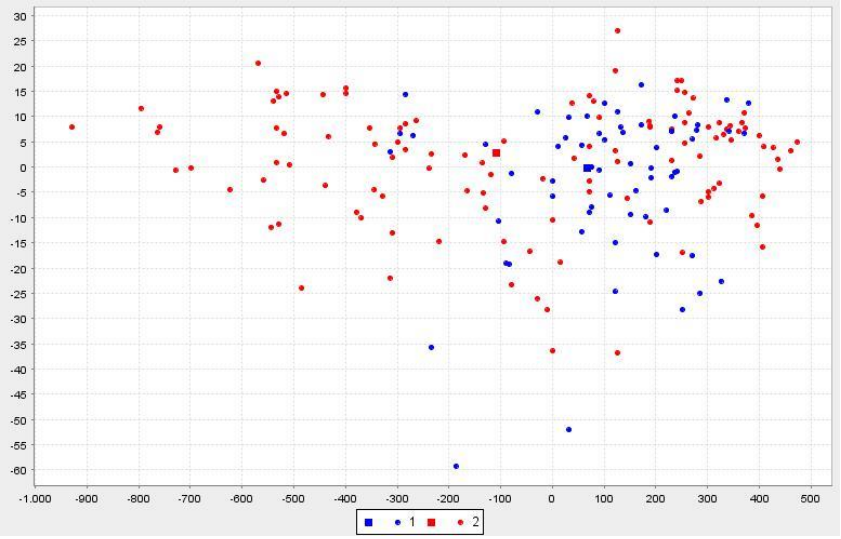
Fonte: Do Autor.

5.1.4.3.2 Base de dados Wine

Na aplicação da base *Wine*, no algoritmo URCP, identificou-se o número ideal de *clusters* como dois, sendo o *cluster* um composto por 113

elementos (63,48%) e o *cluster* dois por 65 elementos (36,52% dos dados) (figura 35).

Figura 34 - Resultado URCP: *Wine* – 1º Experimento



Fonte: Do Autor.

Os valores das medidas de qualidade na base de dados *Wine* pelo algoritmo URCP podem ser observadas na tabela 22.

Tabela 22 - Medidas de qualidade URCP: *Wine*

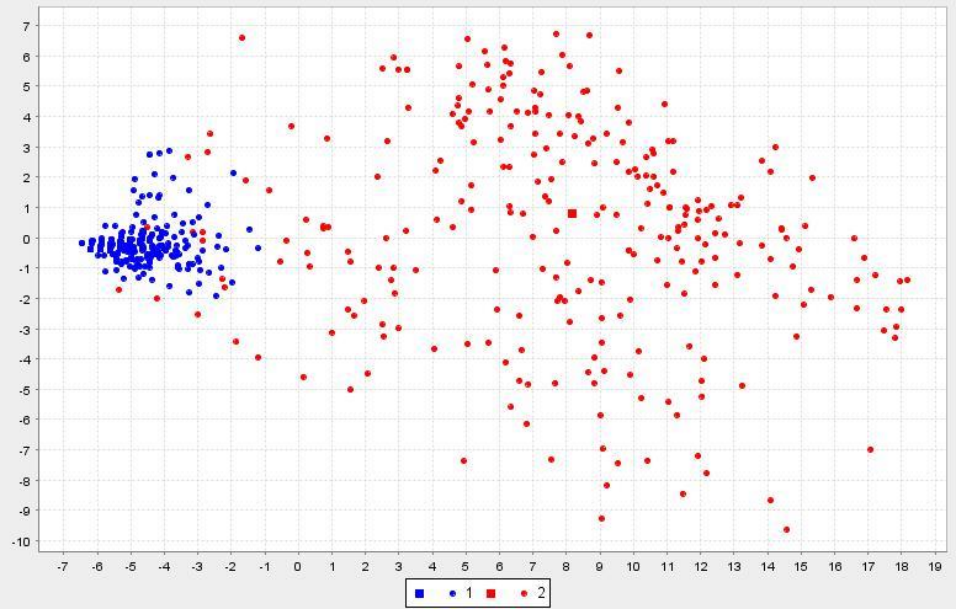
Algoritmo URCP: Número de *Clusters*	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie Beni (-)**
2	0.570524	0.618745	0.025760

Fonte: Do Autor.

5.1.4.3.3 Base de dados Breast Cancer Wisconsin

Na execução da base de dados BCW no algoritmo URCP, o resultado da clusterização, foi de dois *clusters*, sendo o primeiro com 414 elementos (60,61%) e o segundo com 269 elementos (39,39% dos dados) (figura 36).

Figura 35 – Resultado URCP: BCW – 1º Experimento



Fonte: Do Autor.

Os resultados das medidas de qualidade da base BCW, no algoritmo URCP, são mostrados na tabela 20, sendo que todas as medidas encontraram valores bons para a clusterização, indicando um bom particionamento.

Tabela 23 - Medidas de qualidade dos experimentos URCP na BCW

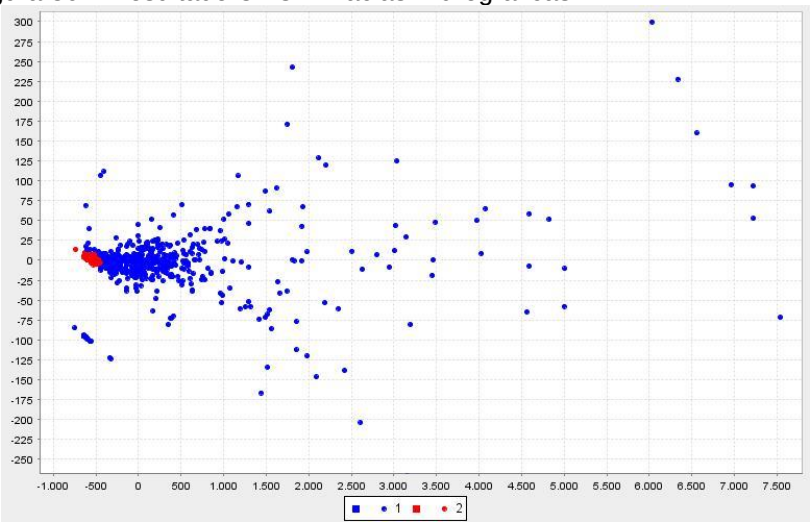
Algoritmo URCP: Número de *Clusters*	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie Beni (-)**
2	0.788365	0.341522	1.434037

Fonte: Do Autor.

5.1.4.3.4 Base de dados de Bacias Hidrográficas

Como resultado da base de dados das bacias hidrográficas, a clusterização obteve dois *clusters*: o primeiro com 531 elementos (55,66% dos dados) e o segundo com 423 elementos (44,34%) (figura 37).

Figura 36 – Resultado URCP: Bacias Hidrográficas



Fonte: Do Autor.

A tabela 24 corresponde aos resultados das medidas de qualidade pela base de dados no algoritmo URCP, no qual as medidas CP e CE indicaram um bom particionamento e o Xie-Beni o contrário.

Tabela 24 - Medidas de qualidade URCP: Bacias Hidrográficas

Algoritmo URCP: Número de *Clusters*	**Coeficiente de Partição (+)**	**Coeficiente de Entropia (-)**	**Xie Beni (-)**
2	0.742586	0.400693	35.964715

Fonte: Do Autor.

5.1.4.4 Aplicação do Algoritmo DBSCAN

O algoritmo DBSCAN foi proposto por Martin Ester, Hans-Peter Kriegel, Jörg Sander e Xiaowei Xu, no ano de 1996, no qual, os *clusters* são formados por um conjunto de pontos conectados por densidade.

Os parâmetros de entrada do algoritmo são: raio da ε-vizinhança de um ponto; número mínimo de pontos e a distância (Euclidiana, Euclidiana Normalizada e *Manhattan*), sendo estas mostradas pela tabela 25 que traz os parâmetros usados em cada experimento.

Neste caso, como o algoritmo DBSCAN gera os *clusters* de forma automática foi realizado apenas um experimento pois não é possível definir o parâmetro quantidade de *clusters*. As medidas de qualidade utilizadas neste algoritmo foram Dunn e C-Index, sendo o Dunn a medida maximizadora e C-Index a medida minimizadora.

Tabela 25 - Parâmetros de entrada do algoritmo DBSCAN

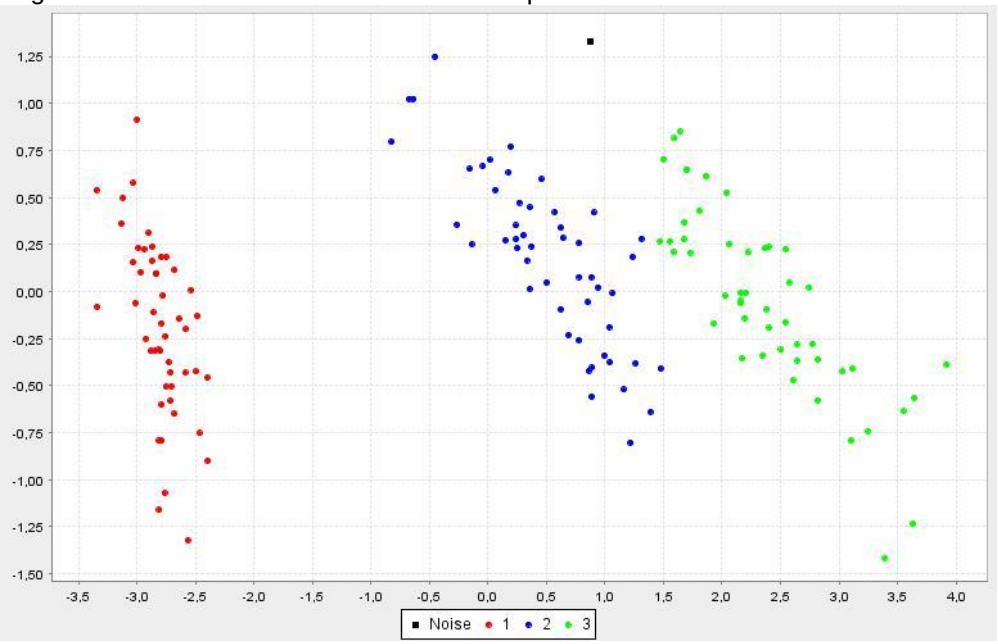
Algoritmo DBSCAN	**Bases de Dados**			
	Iris	***Wine***	***Breast Cancer Wisconsin***	***Bacias Hidrográficas***
Raio da ε-vizinhança	0,9	-	0,9	0,9
Num. Mínimo pontos	4	-	4	4
Distância	Euclidiana	-	Euclidiana Normalizada	Euclidiana Normalizada

Fonte: Do Autor.

5.1.4.4.1 Base de dados Iris

O resultado da clusterização pelo DBSCAN ficou do seguinte modo: *cluster* um e *cluster* dois cada um com 50 elementos (33,56% cada um) e o terceiro *cluster* com 49 elementos (32,89%), 1 registro da base de dados foi classificado como ruído, o gráfico do particionamento pode ser visto pela figura 38.

Figura 37 - Resultado DBSCAN: Iris – 1º Experimento



Fonte: Do Autor.

As medidas de qualidade com relação a base da *Iris* os resultados são mostrados pela tabela 22, onde o C-Index mostra o quão compacto é cada *cluster* gerado pelo algoritmo neste caso obteve-se valores bons. O valor de Dunn representa uma clusterização ruim.

Tabela 26 - Medidas de qualidade DBSCAN: *Iris*

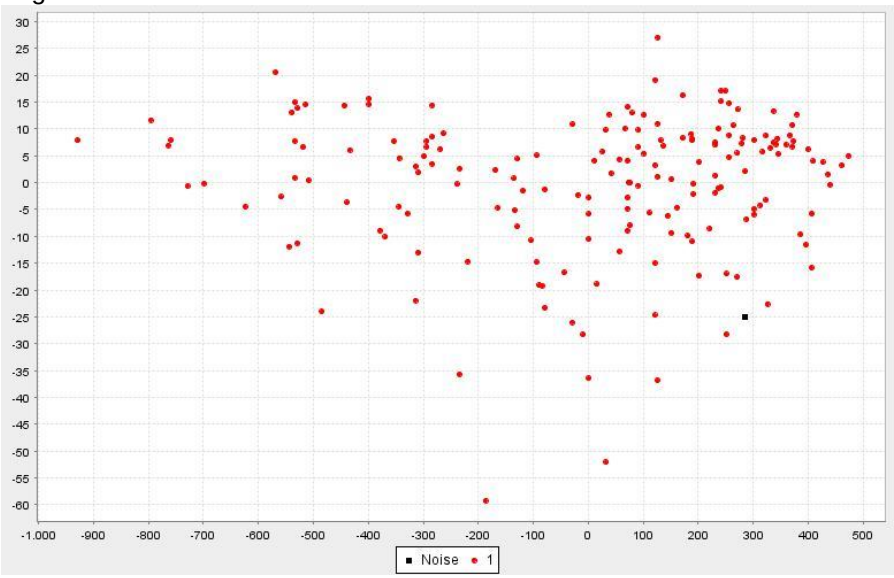
Algoritmo DBSCAN: Número do *Cluster*	**Dunn (+)**	**C-Index (-)**
1	0.344837	0.047277
2	0.344837	0.103562
3	0.344837	0.130308

Fonte: Do Autor.

5.1.4.4.2 Base de dados Wine

Na aplicação da base *Wine* no DBSCAN, mesmo alterando os parâmetros de entrada e utilizando as três distâncias, não foi possível com nenhuma destas clusterizar a base de dados, pois em todos os casos o algoritmo reconhecia 177 registros (dos 178 existentes) como sendo um *cluster* e o registro faltante como ruído, como pode ser observado pela figura 39. As medidas de qualidade neste caso não puderam ser coletadas pelo fato do algoritmo não ter conseguido clusterizar a base de dados, pois as medidas necessariamente precisam de pelo menos dois *clusters* gerados para serem executadas corretamente.

Figura 38 - Resultado DBSCAN: *Wine*

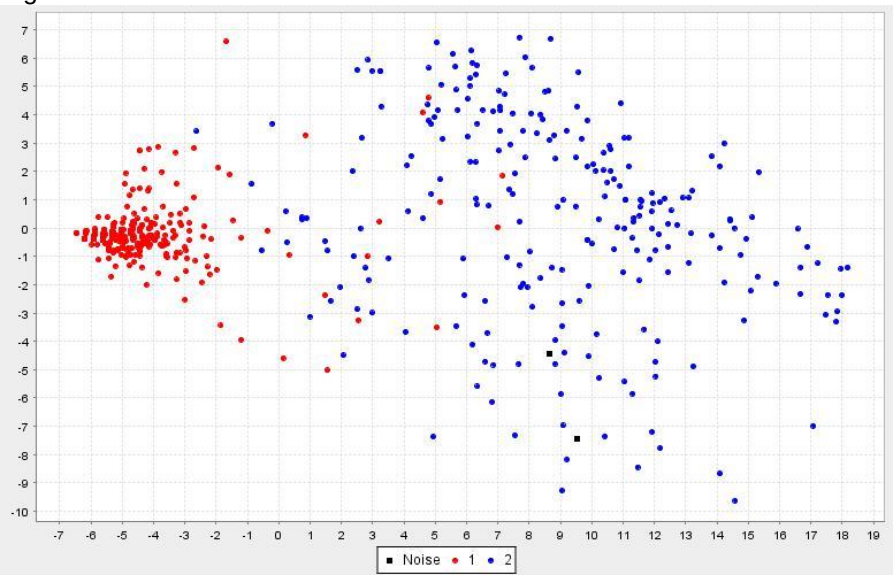


Fonte: Do Autor.

5.1.4.4.3 Base de dados Breast Cancer Wisconsin

Como resultado da aplicação da base BCW a clusterização resultante da base de dados foi de dois *clusters*, sendo o primeiro composto por 444 elementos (65,20%) e o segundo por 237 correspondentes a 34,80% (figura 40).

Figura 39 - Resultado DBSCAN: BCW



Fonte: Do Autor.

Os valores das medidas de qualidade são mostrados pela tabela 27 que mostra *clusters* compactos pelo C-Index e uma clusterização ruim pelo Dunn.

Tabela 27 - Medidas de qualidade dos experimentos DBSCAN na BCW

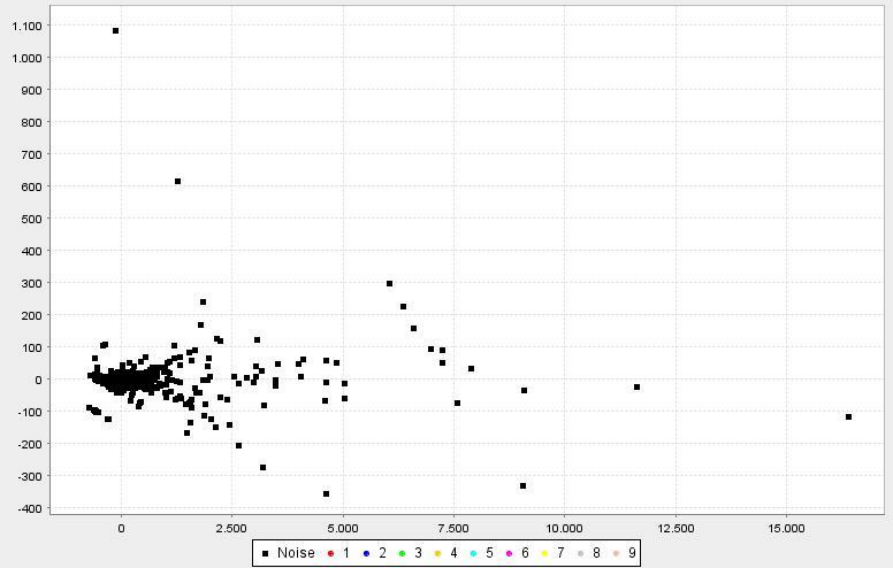
Algoritmo DBSCAN: Número do *Cluster*	**Dunn (+)**	**C-Index (-)**
1	0.402570	0.014563
2	0.402570	0.468115

Fonte: Do Autor.

5.1.4.4.4 Base de dados de Bacias Hidrográficas

O DBSCAN encontrou nove *clusters* para a base de dados das bacias hidrográficas, porém o algoritmo considerou apenas 130 registros dos 954 atributos existentes na base, a figura 41 mostra a clusterização, porém a mesma não é válida, pois não houve um aproveitamento de todos os atributos do conjunto de dados.

Figura 40 - Resultado DBSCAN: Bacias hidrográficas – 1º Elemento



Fonte: Do Autor.

5.1.4.5 Aplicação do Algoritmo ABC

O algoritmo *Ant Based Clustering* (ABC) é uma modificação do algoritmo SACA (proposto por Lumer e Faieta em 1994) proposta por Handl (2003).

Os parâmetros de entrada do algoritmo ABC são: número de formigas, total de iterações, memória (definido como padrão: 2) , alfa α (definido como padrão: 0,7), kp e KD (definidos padrão como: 0.1 e 0.3), sendo estes mostrados pela tabela 28.

Tabela 28 - Parâmetros de entrada algoritmo ABC

Algoritmo ABC	**Bases de Dados**			
	Iris	***Wine***	***Breast Cancer Wisconsin***	***Bacias Hidrográficas***
Num. Iterações	20000	11000	-	10000
Num. Formigas	30	10	-	50
Tamanha da memória das formigas	10	10	-	10

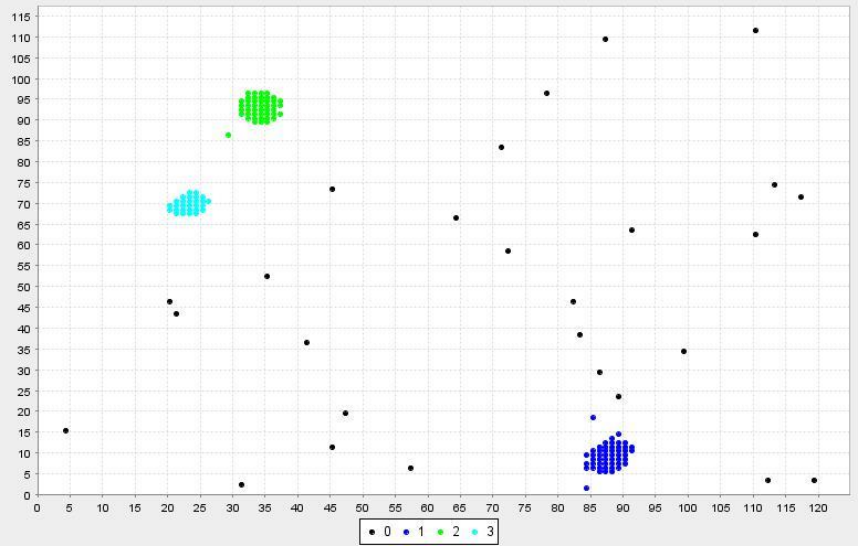
Fonte: Do Autor.

5.1.4.5.1 Base de dados Iris

Como resultado da clusterização encontraram-se três *clusters*, sendo 50 elementos (33,33%) do *cluster* um, 45 elementos (30,00%) pertinentes ao

cluster dois e 28 elementos (18,67%) do *cluster* três, 27 registros foram identificados pelo algoritmo como sendo pontos ruidosos (figura 43).

Figura 41 - Resultado ABC: *Iris*



Fonte: Do Autor.

Pela tabela 29 é possível visualizar que o C-Index indica *clusters* compactos, porém o Dunn indica uma clusterização ruim.

Tabela 29 - Medidas de qualidade dos experimentos ABC na *Iris*

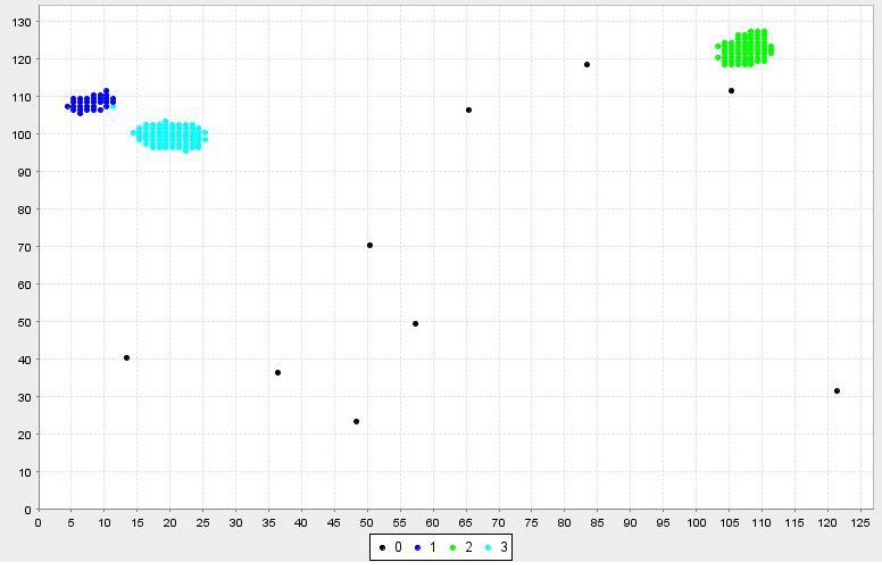
Algoritmo ABC: Número do *Cluster*	**Dunn (+)**	**C-Index (-)**
1	0.012217	0.198707
2	0.012217	0.270436
3	0.012217	0.227525

Fonte: Do Autor.

5.1.4.5.2 Base de dados Wine

Como resultado da clusterização da base de dados *Wine*, encontraram-se três *clusters*, sendo 30 elementos (16,85%) do *cluster* um, 66 elementos (37,08%) pertinentes ao *cluster* dois e 71 elementos (39,89%) do *cluster* três, 11 registros (06,18%) foram identificados pelo algoritmo como sendo pontos ruidosos (figura 44). Os resultados das medidas de qualidade são mostrados pela tabela 30 que mostra um bom particionamento dos dados, tendo em vista os bons resultados das medidas.

Figura 42 - Resultado ABC: *Wine*



Fonte: Do Autor.

Tabela 30 - Medidas de qualidade ABC: *Wine*

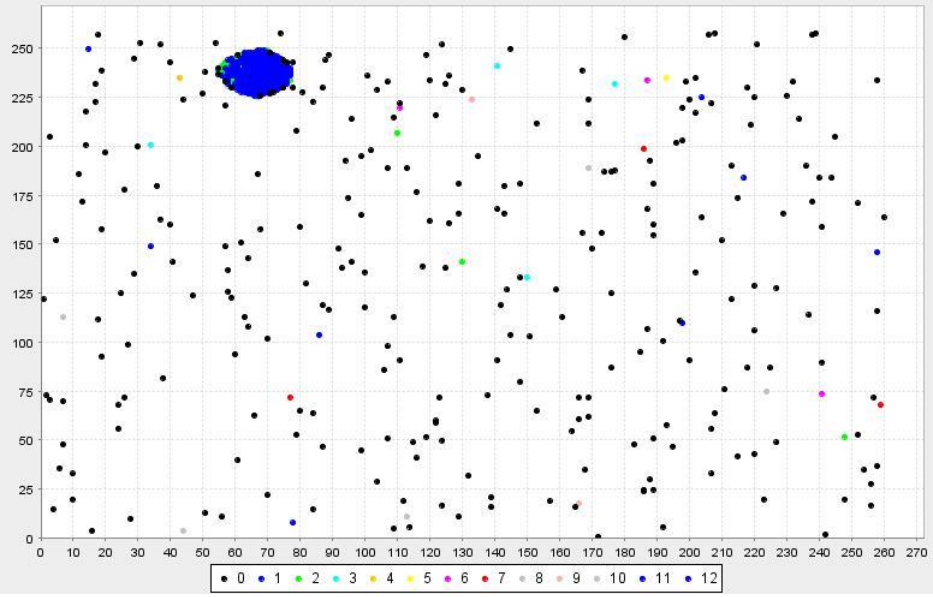
Algoritmo ABC: Número do *Cluster*	**Dunn (+)**	**C-Index (-)**
1	3.612698	0.277684
2	3.612698	0.426757
3	3.612698	0.422466

Fonte: Do Autor.

5.1.4.5.3 Base de dados Breast Cancer Wisconsin

O ABC com relação à base BCW não conseguiu um bom resultado, pois o algoritmo encontrou 12 *clusters* (figura 45) e dos 683 registros da base de dados ele considerou 268 como sendo ruídos. Deste modo as medidas não obtiveram bons resultados e foram desconsideradas.

Figura 43 - Resultado ABC: BCW

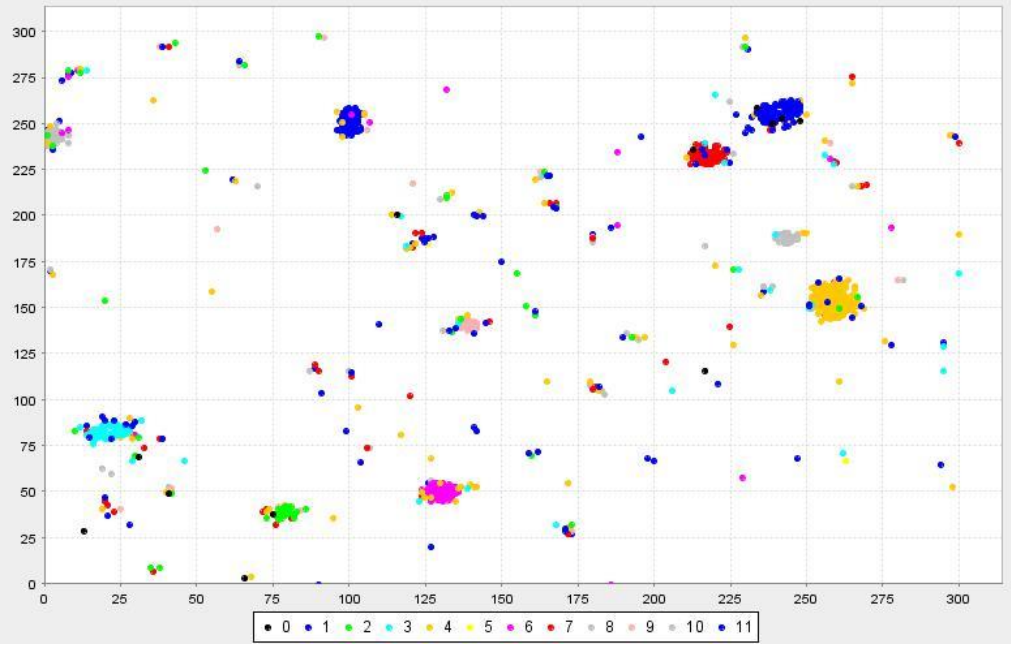


Fonte: Do Autor.

5.1.4.5.4 Base de dados de Bacias Hidrográficas

Na base de dados das bacias hidrográficas houve o mesmo problema que na base de dados BCW, pois o ABC encontrou 12 *clusters* (figura 46) inviabilizando a interpretação dos mesmos e das medidas de qualidade.

Figura 44 - Resultado ABC: Bacias hidrográficas



Fonte: Do Autor.

5.1.4.6 Aplicação do Algoritmo SACA

Segundo Lumer e Faieta (1994), o algoritmo SACA assim como no algoritmo ACA (DENEUBOURN,1991), as formigas e os objetos (conjunto de dados) são espalhados em uma grade bidimensional.

Os parâmetros de entrada do algoritmo são: número de iterações, número de formigas e tamanho da memória das formigas, os parâmetros utilizados são mostrados na tabela 31.

Tabela 31 - Parâmetros de entrada algoritmo SACA

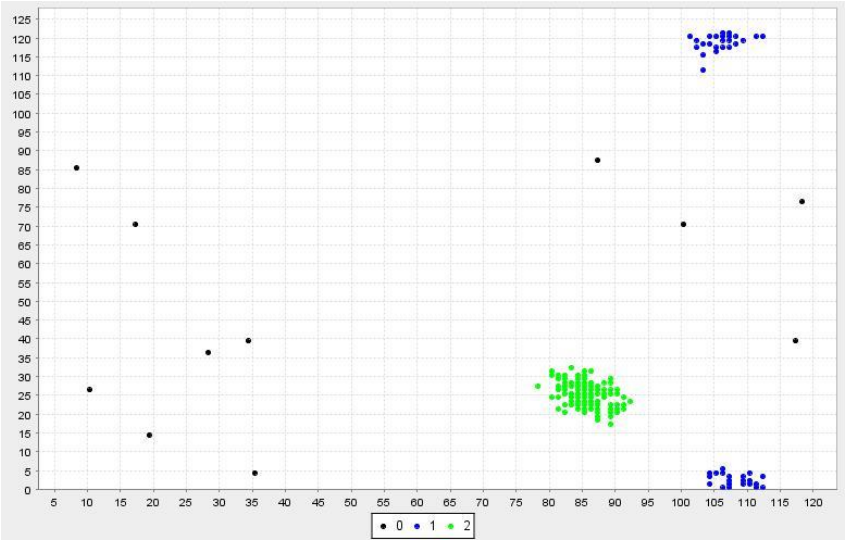
Algoritmo SACA	**Bases de Dados**			
	Iris	***Wine***	***Breast Cancer Wisconsin***	***Bacias Hidrográficas***
Num. Iterações	2000	2000	-	2500
Num. Formigas	60	60	-	60
Tamanha da memória das formigas	30	30	-	30

Fonte: Do Autor.

5.1.4.6.1 Base de dados Iris

Como resultado da clusterização da base de dados *Iris*, encontraram-se dois *clusters*, sendo 51 elementos (34,00%) do *cluster* um e 87 elementos (58,00%) pertinentes ao *cluster* dois, 12 registros (08,00%) foram identificados pelo algoritmo como sendo pontos ruidosos (figura 47).

Figura 45 - Resultado SACA: *Iris*



Fonte: Do Autor.

Os resultados das medidas de qualidade são mostrados na tabela 32, onde observa-se o C-Index indicando *clusters* bem compactos e o Dunn apresentou um valor ruim.

Tabela 32 - Medidas de qualidade SACA: *Iris*

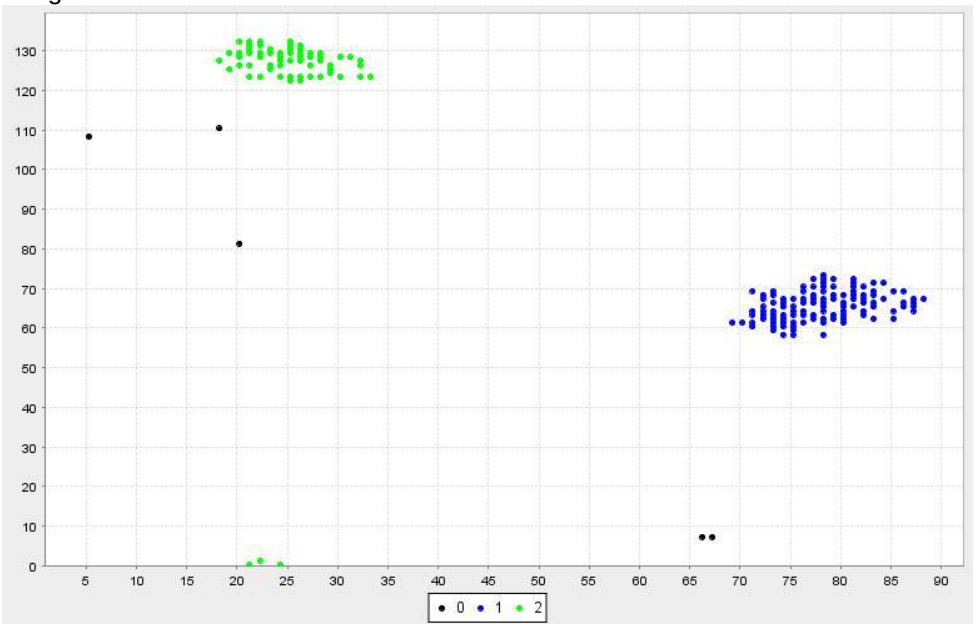
Algoritmo SACA: Número do *Cluster*	**Dunn (+)**	**C-Index (-)**
1	0.023212	0.069286
2	0.023212	0.151758

Fonte: Do Autor.

5.1.4.6.2 Base de dados Wine

O resultado do particionamento da base de dados *Wine* foi de dois *clusters*, sendo 88 elementos (49,44%) do *cluster* um e 67 elementos (37,64%) pertinentes ao *cluster* dois, 23 registros (12,92%) foram classificados pelo algoritmo como pontos ruidosos (figura 48). Os resultados das medidas de qualidade são mostrados pela tabela 33 que mostra um bom particionamento da base de dados pelas duas medidas.

Figura 46 - Resultado SACA: *Wine*



Fonte: Do Autor.

Tabela 33 - Medidas de qualidade dos experimentos SACA na *Wine*

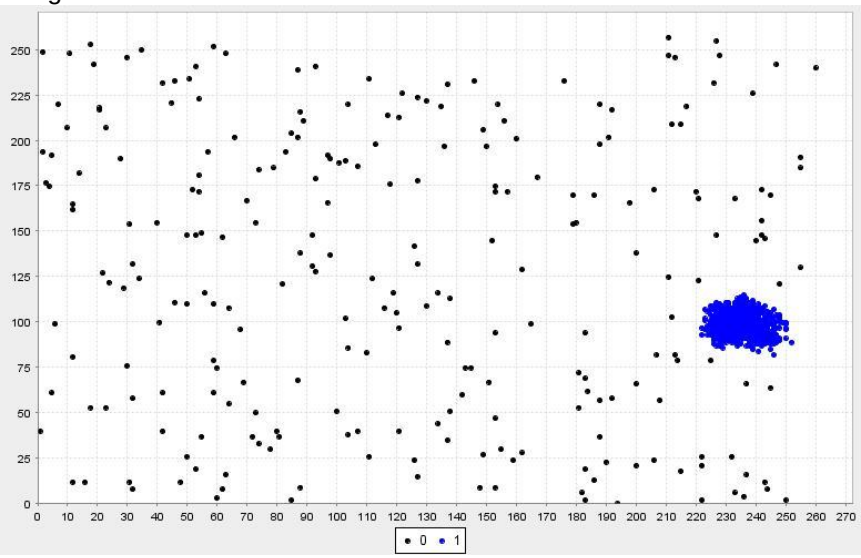
Algoritmo SACA: Número de *Clusters*	**Dunn (+)**	**C-Index (-)**
1	3.469666	0.345134
2	3.469666	0.247756

Fonte: Do Autor.

5.1.4.6.3 Base de dados Breast Cancer Wisconsin

O SACA com relação a base BCW não conseguiu um bom resultado da clusterização, pois o algoritmo encontrou um *cluster* (figura 49) com 434 elementos e dos 683 registros da base de dados, ele considerou 249 como sendo ruídos. Deste modo as medidas não obtiveram bons resultados e foram desconsideradas.

Figura 47 - Resultado SACA: BCW

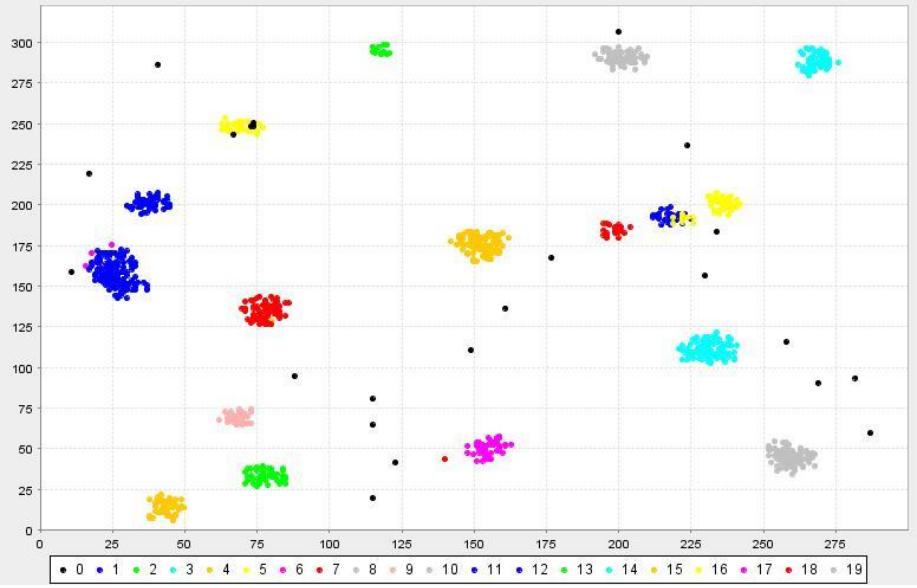


Fonte: Do Autor.

5.1.4.6.4 Base de dados de Bacias Hidrográficas

O resultado do particionamento da base das bacias hidrográficas no SACA foi de 18 *clusters* (figura 50), porém, apesar da grande quantidade de *clusters* gerados, os valores das medidas de qualidade foram aceitáveis, pois a Dunn obteve o valor de 1.790367, e os valores de C-Index podem ser visto pela tabela 34.

Figura 48 - Resultado SACA: Bacias hidrográficas



Fonte: Do Autor.

Tabela 34 - SACA: Valores Dunn bacias hidrográficas

Algoritmo SACA: Número do *Cluster*	**Dunn (+)**	**C-Index (-)**
1	1.790367	0.064043
2	1.790367	0.028695
3	1.790367	0.049455
4	1.790367	0.072870
5	1.790367	0.033138
6	1.790367	0.030368
7	1.790367	0.093781
8	1.790367	0.040057
9	1.790367	0.046253
10	1.790367	0.050909
11	1.790367	0.030192
12	1.790367	0.059692
13	1.790367	0.027998
14	1.790367	0.030640
15	1.790367	0.033738
16	1.790367	0.039323
17	1.790367	0.096229
18	1.790367	0.025066

Fonte: Do Autor.

5.1.4.7 Aplicação do Algoritmo A²CA

Os parâmetros de entrada usados na aplicação do algoritmo A²CA podem ser visualizados na tabela 35.

Tabela 35 - Parâmetros de entrada algoritmo A²CA

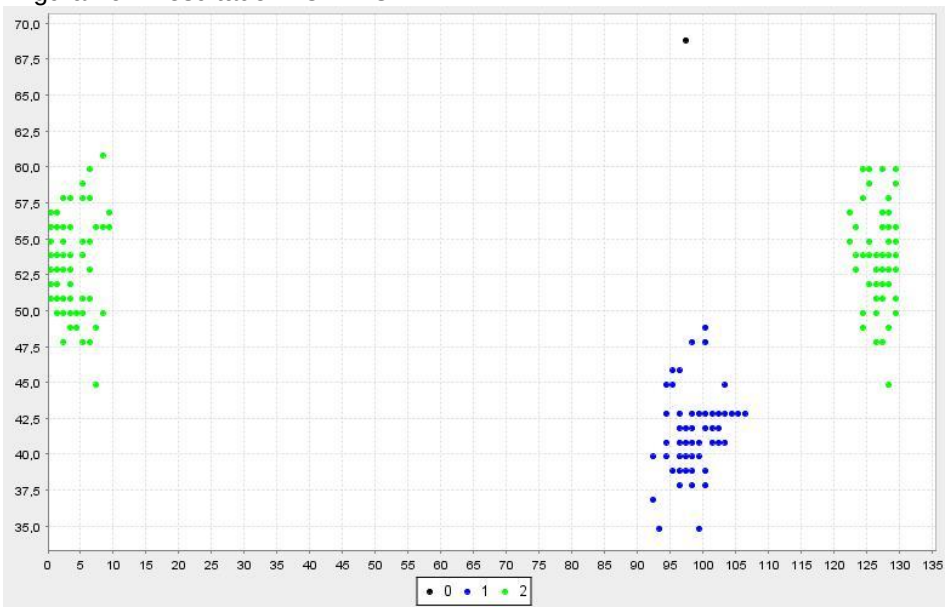
Algoritmo A²CA	**Bases de Dados**			
	Iris	***Wine***	***Breast Cancer Wisconsin***	***Bacias Hidrográficas***
1 º Experimento				
Num. Iterações	2500	3000	3000	3000
Num. Formigas	100	40	100	40
Tamanha da memória das formigas	20	20	20	30

Fonte: Do Autor.

5.1.4.7.1 Base de dados Iris

O resultado da aplicação do A²CA na Iris foi de dois *clusters*, sendo 99 elementos (66,00%) do *cluster* um e 50 elementos (33,33%) pertinentes ao *cluster* dois, apenas um registro foi classificado pelo algoritmo como ruído (figura 51). Os resultados das medidas de qualidade são mostrados pela tabela 36, onde pode-se observar que o Dunn obteve um valor ruim e o C-Index apontou cluster compactos.

Figura 49 - Resultado A²CA: *Iris*



Fonte: Do Autor.

Tabela 36 - Medidas de qualidade A²CA: *Iris*

Algoritmo SACA: Número de *Clusters*	**Dunn (+)**	**C-Index (-)**

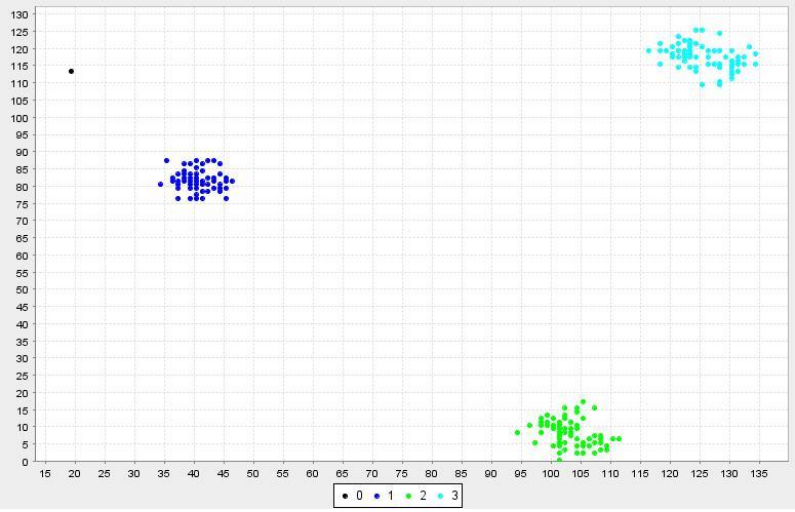
1	0.053226	0.032541
2	0.053226	0.164694

Fonte: Do Autor.

5.1.4.7.2 Base de dados Wine

O resultado da clusterização da base de dados *Wine* foi de três *clusters*, sendo 54 elementos do *cluster* um, 62 elementos pertinentes ao *cluster* dois e 61 do terceiro *cluster*, apenas um registro foi classificado pelo algoritmo como ruído (figura 52). Os resultados das medidas de qualidade são mostrados na tabela 37, que indica uma boa clusterização pelos valores das medidas.

Figura 50 - Resultado A²CA: *Wine*



Fonte: Do Autor.

Tabela 37 - Medidas de qualidade A²CA: *Wine*

Algoritmo A²CA: Número de *Clusters*	**Dunn (+)**	**C-Index (-)**
1	3.955415	0.416503
2	3.955415	0.271009
3	3.955415	0.289912

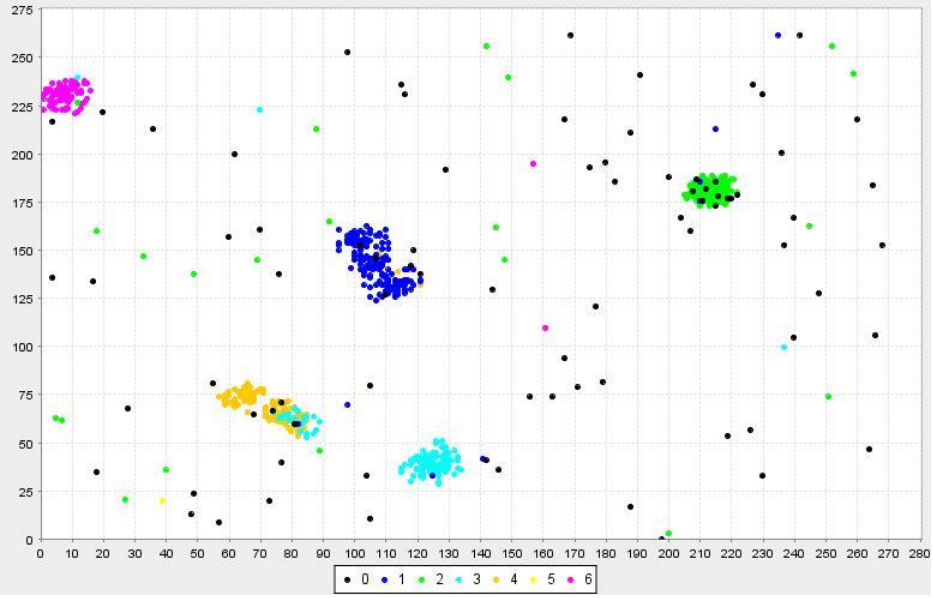
Fonte: Do Autor.

5.1.4.7.3 Base de dados Breast Cancer Wisconsin

O resultado da clusterização da base de dados BCW foi de seis *clusters. O cluster* um possui 160 elementos (23,43%); *clusters* dois 146 (21,38%); *cluster* três 111 elementos (16,25%); *cluster* quatro 111 elementos (16,25%); *cluster* cinco 1 elemento (00,15%) e *cluster* seis 73 elemento (10,69%); 81 registros (11,86%) foram classificados pelo algoritmo como ruído (figura 53).

Os resultados das medidas de qualidade são mostrados na tabela 38, sendo possível identificar *clusters* compactos pela medida C-Index e uma clusterização ruim pelo Dunn.

Figura 51 - Resultado A²CA: BCW



Fonte: Do Autor.

Tabela 38 - Medidas de qualidade A²CA: BCW

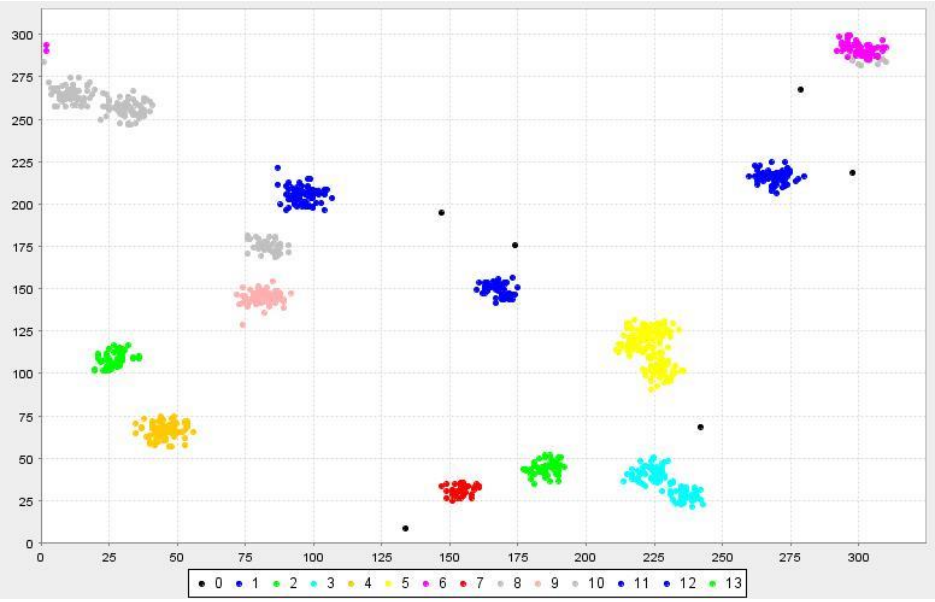
Algoritmo SACA: Número de *Clusters*	**Dunn (+)**	**C-Index (-)**
1	0.012730	0.114613
2	0.012730	0.205437
3	0.012730	0.108499
4	0.012730	0.114379
5	0.012730	NaN
6	0.012730	0.120600

Fonte: Do Autor.

5.1.4.7.4 Base de dados de Bacias Hidrográficas

O resultado do clusterização da base de dados das bacias hidrográficas foi de 13 *clusters* (figura 54) e seis elementos o algoritmo identificou como ruído, o resultado dos *clusters* e das medidas de qualidade são observados na tabela 39, a qual mostra que as duas medidas indicaram uma boa clusterização.

Figura 52 - Resultado A²CA: Bacias Hidrográficas



Fonte: Do Autor.

Tabela 39 - Medidas de qualidade A²CA na Bacias hidrográficas

Algoritmo SACA: Número de *Clusters*	**Dunn (+)**	**C-Index (-)**
1	7.006652	0.099156
2	7.006652	0.026325
3	7.006652	0.088237
4	7.006652	0.076949
5	7.006652	0.098553
6	7.006652	0.041632
7	7.006652	0.018795
8	7.006652	0.077167
9	7.006652	0.051601
10	7.006652	0.048161
11	7.006652	0.048576
12	7.006652	0.039351
13	7.006652	0.029857

Fonte: Do Autor.

Finalizada a etapa de aplicação dos setes algoritmos de clusterização da Shell Orion nas quatro bases de dados definidas para a realização desse estudo, iniciou-se a etapa de análise a fim de se compreender os resultados obtidos.

5.2 RESULTADOS OBTIDOS

Após a aplicação dos sete algoritmos de clusterização, realizou-se a análise dos resultados obtidos pela clusterização de cada uma das bases de dados empregadas na execução dos experimentos, observando-se o funcionamento do algoritmo, a interpretação dos valores gerados pelas medidas de qualidade e a comparação com os resultados de pesquisas realizadas em âmbito acadêmico por outros autores.

5.2.1 Resultados da Clusterização pelos Algoritmos

Todos os algoritmos foram executados com as quatro bases de dados, porém, alguns não obtiveram resultados satisfatórios quanto a separação ou a coesão dos *clusters* gerados, um dos fatores, foi o fato de algumas bases de dados possuírem *clusters* muito próximos afetando ou o desempenho dos algoritmos ou das medidas consideradas.

5.2.1.2 Algoritmos versus base de dados *Iris*

A base de dados da *Iris* possui três *clusters*, correspondentes aos tipos da flor (*Virginica*, *Versicolor* e *Setosa*), porém em alguns trabalhos, as medidas de qualidade encontram dois *clusters* como sendo o ideal, isso se deve ao fato dos *clusters* correspondentes a Virginica e Versicolor serem muito próximos, quase sobrepostos. Porém, o número ideal de *clusters* para o conjunto de dados da *Iris* perante a literatura é aceito tanto dois, quanto três (ZAHID; LIMOURI; ESSAID, 1999, tradução nossa).

No algoritmo FCM nos três experimentos onde foi escolhida a quantidade de *clusters* em 2, 3 e 4, pode-se observar que os melhores valores das medidas CP, CE e XB encontraram-se na primeira execução que resultou em dois *clusters*. Portanto, conforme os testes realizados nesta pesquisa as medidas de qualidade apontam que para a base de dados da Íris dois é o número ideal de *clusters*.

Quando executado no algoritmo RCP, as medidas CP e CE conseguiram identificar três *clusters* na base de dados, enquanto a XB só dois *clusters*.

Pelo algoritmo URCP não foi possível definir a quantidade de *clusters* como no RCP, pois ele identifica-os de forma automática, porém o algoritmo conseguiu particionar corretamente em três *clusters*, o que foi confirmado pelas medidas CP e CE que indicam um bom valor, já a XB trouxe um valor ruim, reafirmando o resultado do RCP de dois *clusters* como o número ideal.

No algoritmo DBSCAN também não se pode inserir a quantidade de *clusters*, portanto o resultado gerado pelo algoritmo foi de três *clusters*. Os valores obtidos pelo C-Index para cada *cluster*, mostrou que os mesmos foram bem separados, entretanto o valor do Dunn foi ruim, demonstrando que a medida não considera três o número ideal de *cluster*.

O algoritmo *Ant Based Clustering* separou corretamente os dados em três *clusters*, porém rotulou 27 elementos dos 150, como dados ruidosos o que não foi bom, porém mesmo assim o C-Index obteve valores bons para todos os *clusters* e o Dunn novamente não encontrou um valor bom, assim como no algoritmo DBSCAN.

Pelo algoritmo SACA foram encontrados dois *clusters* para a base de dados da *Iris* e assim como nos exemplos anteriores, apenas o C-Index trouxe valores bons para o particionamento, o valor encontrado pelo Dunn indicou uma clusterização muito ruim, próxima de zero.

No algoritmo A²CA o resultado encontrado foi de dois *clusters* e a mesma situação anterior foi encontrada para as medidas C-Index e Dunn.

Osresultados mostrados anteriormente, traz-se uma tabela com os valores das medidas dos três primeiros algoritmos (tabela 36) e abaixo dela outra mostrando os resultados dos quatro algoritmos faltantes, para ambas as tabelas os valores ruins correspondem a cor vermelha (tabela 37).

Tabela 40 – *Iris*: Resultado das medidas de qualidade algoritmos FCM, RCP e URCP

Alg.	2 Clusters			3 Clusters			4 Clusters		
	CP	CE	XB	CP	CE	XB	CP	CE	XB
FCM	0,6824	0,4936	0,2120	0,54481	0,7880	0,34001	0,4302	1,0649	0,3153
RCP	0.6518	0.5213	0.1051	0.9496	0.1217	3.4356	0.4287	1.054.7	0.2741
URCP	X	X	X	0.7813	0.5487	9,8947	X	X	X

Fonte: Do Autor.

Tabela 41 – *Iris*: Resultados Medidas de qualidade algoritmos DBSCAN, ABC, SACA, A²CA

Algoritmos	Nº do Cluster	Dunn	C-Index
DBSCAN	1	0.344837	0.047277
	2	0.344837	0.103562
	3	0.344837	0.130308
ABC	1	0.012217	0.198707
	2	0.012217	0.270436
	3	0.012217	0.227525
SACA	1	0.023212	0.069286
	2	0.023212	0.151758
A²CA	1	0.053226	0.032541
	2	0.053226	0.164694

Fonte: Do Autor.

5.2.1.3 Algoritmos versus Base de dados *Wine*

A base de dados *Wine*, é amplamente usada em trabalhos na área de avaliação de *cluster*, mesmo a literatura indicando a presença de três *clusters* é uma base que pode gerar números inferiores ou superiores, pois segundo Saitta, Raphael e Smith (2007, tradução nossa) é uma base de dados com 13 dimensões.

No algoritmo FCM pode-se observar que os valores das medidas CP, CE e XB ambas identificam dois *clusters* como sendo o ideal. Aplicando-se o algoritmo RCP, as medidas CP e CE identificaram dois *clusters* na base de dados, e só o XB conseguiu identificar três *clusters*.

O algoritmo URCP particionou os dados em dois *clusters*, porém as medidas CP e CE não obtiveram valores bons, indicando que dois não é o número de *clusters* na base de dados. A medida XB trouxe um valor bom, errando quanto ao número de clusters.

O algoritmo DBSCAN não conseguiu clusterizar os dados da base Wine, pois identificou apenas um *cluster* no conjunto de dados.

No algoritmo *Ant Based Clustering* os resultados foram bons, pois todas as medidas obtiveram valores bons, confirmando o número de *clusters* encontrado pelo algoritmo que foi de três.

Já pelo algoritmo SACA foi encontrado dois *clusters* para a base de dados *Wine* e os valores das medidas indicaram o correto particionamento dos dados, mesmo sendo dois o número de *cluster*.

No algoritmo A²CA o resultado encontrado foi de três *clusters*, sendo que as três medidas indicaram um bom particionamento dos dados.

Nas tabelas 38 e 39 pode-se observar os valores encontrados pelas medidas de qualidade em relação ao conjunto de dados Wine.

Tabela 42 - *Wine*: Resultado das medidas de qualidade algoritmos FCM, RCP e URCP

Alg.	2 *Clusters*			3 *Clusters*			4 *Clusters*		
	CP	CE	XB	CP	CE	XB	CP	CE	XB
FCM	0,7152	0,4456	0,1974	0.5723	0.7456	0.2477	0.5070	0.9344	0.2136
RCP	0.5039	0.6891	0.1432	0.3850	0.1217	1.0180	0.2668	1.3488	0.03495
URCP	0.7883	0.3415	1.4340	X	X	X	X	X	X

Fonte: Do Autor.

Tabela 43 - *Wine*: Resultados Medidas de qualidade algoritmos DBSCAN, ABC, SACA, A²CA

Algoritmos	Nº do *Cluster*	Dunn	C-Index
DBSCAN	X	X	X
ABC	1	3.612698	0.277684
	2	3.612698	0.426757
	3	3.612698	0.422466
SACA	1	3.469666	0.345134
	2	3.469666	0.247756
A²CA	1	3.955415	0.416503
	2	3.955415	0.271009
	3	3.955415	0.289912

Fonte: Do Autor.

5.2.1.4 Algoritmos versus base de dados *Breast Cancer Wisconsin*

A base de dados BCW é uma base com dados médicos e assim como as anteriores, também é muito usada por pesquisadores da área, sendo composta por dois *clusters*, porém, mesmo os centros dos *clusters* sendo distantes, nota-se que há uma certa sobreposição dos dois *clusters*.

Aplicando-se o algoritmo FCM percebe-se que as medidas CP e CE indicam o particionamento correto em dois *clusters* e o XB em três.

No algoritmo RCP, as medidas CP e CE identificaram dois *clusters* na base de dados, e o XB não conseguiu um bom resultado em nenhuma

clusterização. Já com o algoritmo URCP tanto o algoritmo, quanto as medidas indicaram dois *clusters* como sendo o ideal.

O algoritmo DBSCAN particionou os dados em dois *clusters*, a medida C-Index mostrou bom resultado, porém a Dunn gerou um valor ruim.

Os algoritmos ABC e SACA não conseguiram clusterizar corretamente esta base de dados. Enquanto o algoritmo A²CA, mesmo gerando seis *clusters*, os mesmos foram considerados bem separados pelo C-Index, e o contrário pela medida Dunn.

As tabelas 38 e 39 disponibilizam os valores encontrados pelas medidas de qualidade em relação ao conjunto de dados *Breast Cancer Wisconsin*.

Tabela 44 - BCW: Resultado das medidas de qualidade algoritmos FCM, RCP e URCP

Alg.	2 *Clusters*			3 *Clusters*			4 *Clusters*		
	CP	CE	XB	CP	CE	XB	CP	CE	XB
FCM	0.6537	0.5256	0.2909	0.4814	0.8896	0.2125	0.3268	1.2188	5.6621
RCP	0.7828	0.3505	1.4569	0.5520	0.7219	4.1005	0.4600	0.9571	1924.48
URCP	0.7883	0.3415	1.4340	X	X	X	X	X	X

Fonte: Do Autor.

Tabela 45 - BCW:Resultados Medidas de qualidade algoritmos DBSCAN, ABC, SACA, A²CA

Algoritmos	Nº do Cluster	Dunn	C-Index
DBSCAN	1	0.402570	0.014563
	2	0.402570	0.468115
ABC	X	X	X
SACA	X	X	X
A²CA	1	0.012730	0.114613
	2	0.012730	0.205437
	3	0.012730	0.108499
	4	0.012730	0.114379
	5	0.012730	NaN
	6	0.012730	0.120600

Fonte: Do Autor.

5.2.1.5 Algoritmos versus base de dados das bacias Hidrográficas

Os dados desta base de dados foram aplicados em alguns trabalhos do Grupo de Inteligência Computacional Aplicada, porém em nenhum deles chegou a

conclusão do número ideal de *clusters*, cada trabalho resulta em uma clusterização diferente.

Na execução do algoritmo FCM as três medidas obtiveram os melhores resultados com dois *clusters*, indicando como sendo o número ideal. A mesma resposta foi encontrada pelas medidas nos algoritmos RCP e URCP.

No algoritmo DBSCAN e ABC o particionamento gerado foi ruim, ou não interpretável com relação a base de dados das bacias hidrográficas, então foram desconsiderados.

O algoritmo SACA resultou em 18 *clusters*, mesmo com um número alto de *clusters* as medidas indicaram um bom particionamento. No algoritmo A²CA o resultado encontrado foi 12 *clusters* e as duas medidas indicaram um excelente particionamento dos dados.

Nas tabelas 42 e 43 pode-se observar os valores encontrados pelas medidas de qualidade em relação a base de dados das bacias hidrográficas.

Tabela 46 - Bacias Hidrográfias: Resultados das medidas de qualidade algoritmos FCM, RCP e URCP

Alg.	2 *Clusters*			3 *Clusters*			4 *Clusters*		
	CP	CE	XB	CP	CE	XB	CP	CE	XB
FCM	0.7658	0.3831	0.1627	0.6826	0.5619	0.2059	0.64424	0.6821	0.0949
RCP	0.6997	0.4612	12.1975	0.6231	0.6470	22.1062	X	X	X
URCP	0.7425	0.4006	35.9647	X	X	X	X	X	X

Fonte: Do Autor.

Tabela 47 - Bacias Hidrográficas: Resultados Medidas de qualidade algoritmos DBSCAN, ABC, SACA, A²CA

Algoritmos	Nº do *Cluster*	Dunn	C-Index
DBSCAN	X	X	X
ABC	X	X	X
SACA	18	1.790367	0.0250 - 0.0962
A²CA	13	7.0066	0.0187 - 0.0991

Fonte: Do Autor.

5.2.1.6 Comparação de Resultados

A fim de comparar os resultados encontrados pelas medidas aplicadas no presente trabalho, com os valores encontrados por outros autores com relação a cada base de dados e que medidas encontraram esses valores, tem-se a tabela 48. Por esta tabela é possível visualizar os valores encontrados pelas medidas

aplicadas em outros trabalhos quanto ao número de *clusters*, também é possível observar que o presente trabalho é o que mais aplica algoritmos de clusterização em relação aos demais.

Tabela 48 - Resultados das medidas de qualidade por outros autores

(continua)

Autor	Algoritmos	*Iris*					*Breast Cancer Wisconsin*					*Wine*				
		CP	CE	XB	CI	DN	CP	CE	XB	CI	DN	CP	CE	XB	CI	DN
Bandyopadhyay e Maulik (2001)	VGA-Clustering	-	-	-	-	2	-	-	-	-	-	-	-	-	-	2
Boudraa (1999)	FCM	4	4	4	-	-	-	-	-	-	-	-	-	-	-	-
Ding e Harrison (2007)	K-Means	-	-	-	3	2	-	-	-	3	2	-	-	-	-	-
	PAM	-	-	-	3	2	-	-	-	4	2	-	-	-	-	-
	SOM	-	-	-	5	2	-	-	-	5	2	-	-	-	-	-
	Neuro-Gas	-	-	-	3	3	-	-	-	7	2	-	-	-	-	-
	Hierarchical	-	-	-	3	2	-	-	-	8	2	-	-	-	-	-
Fernandèz e Corbacho (2010)	Algoritmo Genético	-	-	-	-	2	-	-	-	-	2	-	-	-	-	3
Hu et al (2011)	FCM	2	2	2	-	-	-	-	-	-	-	2	2	2	-	-
Kim, Lee e Lee (2004)	FCM	2	2	2	-	-	-	-	-	-	-	-	-	-	-	-
Kim e Ramakrishma (2005)	K-Means	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-
Maulik e Bandyopadhyay (2002)	K-Means	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-
	Single Linkage	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-
	Simulated Annealing (AS)	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-
Saad; Alimi, (2012)	FCM	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-
Mok et al. (2012)	FCM	2	2	2	-	-	2	2	2	-	-	2	2	3	-	-
Pakhira, Bandyopadhyay e Maulik (2004)	FCM	-	-	2	-	2	-	-	2	-	2	-	-	-	-	-

Tabela 49 - Resultados das medidas de qualidade por outros autores

(conclusão)

Autor	Algoritmos	*Iris*					*Breast Cancer Wisconsin*					Wine				
		CP	CE	XB	CI	DN	CP	CE	XB	CI	DN	CP	CE	XB	CI	DN
Saitta, Raphael e Smith (2007)	K-Means	-	-	-	-	2	-	-	-	-	8	-	-	-	-	-
Silva e Gomide (2004)	Participatory Learning (AP)	3	3	3	-	-	-	-	-	-		-	-	-	-	-
Tasdemir e Merényi (2011)	K-Means	-	-	-	-	2	-	-	-	-	2					2
Wang e Zhang, (2007)	FCM	2	2	-	-	-	2	2	-	-	-	2	2	-	-	-
Zahid, Limourie e Essaid (1999)	FCM	2	2	2	-	-	-	-	-	-	-	-	-	-	-	-
Zhang et al (2008)	FCM	2	2	2	-	-	2	2	2	-	-	2	2	3	-	-
Palhano (2013)	FCM	2	2	2	-	-	2	2	3	-	-	2	2	2	-	-
	RCP	3	3	2	-	-	2	2	2	-	-	2	2	4	-	-
	URCP	3	3	3	-	-	2	2	2	-	-	2	2	-	-	-
	DBSCAN	-	-	-	-	3	-	-	-	-	2	-	-	-	-	-
	ABC	-	-	-	-	3	-	-	-	-	-	-	-	-	2	2
	SACA	-	-	-	-	2	-	-	-	-	-	-	-	-	2	2
	A²CA	-	-	-	-	2	-	-	-		6	-	-	-	3	3

Fonte: Do Autor.

Pela tabela 48 também pode-se visualizar que os valores encontrados pelas medidas de qualidade neste trabalho, são muito parecidos com o de outros autores, confirmando assim a validade dos valores encontrados e a relevância do presente trabalho na área de medidas de qualidade na tarefa de clusterização.

Observa-se na aplicação dos sete algoritmos que nem todos conseguiram atender os requisitos mínimos da clusterização, mostrado pelo Capítulo 2, pois quando tinham-se bases de dados com: muitas dimensões (*Iris* e *Wine*), algum tipo de ruído (bacias hidrográficas), ou com muitos elementos ou atributos (bacias hidrográficas e *breast cancer Wisconsin*) alguns algoritmos, como foi o caso do DBSCAN, ABC e SACA não conseguiram particionar corretamente os *clusters*.

Com relação as cinco medidas de qualidade aplicadas pode-se afirmar que as medidas que melhor mostraram o número ideal de *clusters* e se estes eram compactos e bem separados, sem contrariedades foram o Coeficiente de Partição, Coeficiente de Partição Entrópica e C-Index. Já as medidas que mais

foram arbitrárias em seus resultados foram a Xie-Beni e Dunn, o Dunn justifica-se pelo fato de ser uma medida muito sensível a ruídos.

6 CONCLUSÃO

Os avanços dos computadores com relação a desempenho e custo acessível possibilitaram que o data mining pudesse ganhar cada vez mais destaque na análise e processamento de dados, possibilitando trazer padrões novos e desconhecidos nas bases de dados.

Considerando isso, esta pesquisa fundamentou-se em entender conceitos como o de clusterização que é uma das tarefas mais importantes do data mining, e sobre medidas de qualidade que são métricas para avaliar o particionamento gerado quanto a coesão e separação dos *clusters*.

Dentre as dificuldades encontradas no decorrer da pesquisa destacam-se a falta de trabalhos na área de aplicação de medidas de qualidade e a escolha das bases de dados que fossem adequadas para executar os algoritmos e o entendimento do funcionamento de medidas de qualidade na clusterização, porém todas as dificuldades foram superadas e os objetivos alcançados.

Durante o trabalho foi realizada uma pesquisa das medidas de qualidade existentes e também uma revisão da literatura quanto as aplicações das medidas, no qual tornou-se visível quais medidas, algoritmos e bases de dados são mais utilizadas.

No trabalho empregaram-se sete algoritmos de clusterização da Shell Orion, sendo cada um executado com quatro bases de dados diferentes. O particionamento realizado pelos algoritmos para cada base de dados foram analisados por meio de medidas de qualidade, chegando-se a conclusão de que a maioria dos algoritmos da Shell Orion conseguiram identificar corretamente os *clusters* quando a base de dados era pequena, porém com bases de dados maiores como foi o caso das bacias hidrográficas e *Wine*, alguns algoritmos não conseguiram agrupar de forma correta os dados.

Com relação as medidas de qualidade observou-se que elas são muito relativas, pois nem sempre uma medida que funciona corretamente em uma base de dados ou algoritmo específico se comportará deste mesmo modo com outra base de dados ou algoritmo. Destaca-se também a vulnerabilidade das medidas com relação a ruídos, bases de dados com *clusters* sobrepostos, e bases de dados com múltiplas dimensões.

Com isso, conclui-se que medidas de qualidade na clusterização é uma área relativamente nova e ainda está em constantes estudos e buscando cada vez mais novas medidas que se comportem melhor a uma determinada aplicação, pois não existe uma medida universal que consiga atender a todo algoritmo ou base de dados, porém é a única forma de auxiliar o usuário quanto ao número ideal de *clusters* e a medir a qualidade dos *clusters* gerados.

Ao final, são descritas algumas sugestões de trabalhos futuros, visando dar continuidade ao desenvolvimento do projeto da Shell Orion Data Mining Engine:

a) implementar mais medidas de qualidade na Shell Orion, pois por meio deste trabalho identificaram-se várias medidas de qualidade existentes e que poderiam ser implementadas e agregadas aos algoritmos;
b) realizar uma revisão sistemática mais abrangente da que foi feita neste trabalho, quanto ao uso das medidas de qualidade;
c) comparar os resultados das medidas obtidas dos algoritmos da Shell Orion com outras ferramentas de *data mining;*
d) empregar testes estatísticos para análise dos resultados obtidos pelas medidas de qualidade aplicadas aos algoritmos de clusterização, no qual um teste que poderia ser aplicado neste contexto é o *Inter Rater Agreement*.
e) agregar na Orion uma parte heurística para identificação do melhor algoritmo de clusterização a partir da avaliação das medidas de qualidade de forma automática, onde o usuário só escolhe a base de dados e a Shell Orion indica o algoritmo que melhor clusteriza os dados com base nos melhores valores das medidas de qualidade.

REFERÊNCIAS

AGRAWAL, R., GEHRKE, J., GUNOPULOS, D., et al., 1998, “Automatic Subspace Clustering on High Dimensional Data for Data Mining Applications”, In: Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 94-105, Seattle, Washington, USA, June.

ALVES, Luciano. **Eficiência de métodos de agrupamento de dados na modelagem nebulosa Takagi-Sugeno.** 2007. 143 f. Dissertação (Mestre em Engenharia de Produção e Sistemas) – Pontifica Universidade Católica do Paraná, Curitiba, 2007.

ANKERST, M., BREUNIG, M., M., KRIEGEL, H.-P., et al., 1999, “OPTICS: Ordering Points to Identify the Clustering Structure”, In: Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 49-60, Philadelphia, PA, USA, June.

ANSARI, Zahid et al. A Comparative Study of Mining Web Usage Patterns Using Variants of k-Means Clustering Algorithm. International Journal of Computer Science and Information Technologies, Vol. 2 (4), July 2011, pp. 1407-1413.

APPEL, Ana Paula. **Métodos para o pré-processamento e mineração de grandes volumes de dados multidimensionais e redes complexas.** 2010. 179f. Tese (Doutorado em Ciências da Computação e Matemática Computacional)- Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo, São Carlos, 2010.

ARBELAITZ, Olatz et al. An extensive comparative study of cluster validity indices, *Pattern Recognition*, Volume 46, Issue 1, January 2013, pp. 243-256.

BAARSCH, Jonathan; CEBELI, M. Emre. Investigation of Internal Validity Measures for K-Means Clustering. Proceedings of the International, Multi Conference of Engineers and Computer Scientists, Vol 1, pp. 471-476, Mar. 2012.

BACKER, E.; JAIN, Anil K., A Clustering Performance Measure Based on Fuzzy Set Decomposition, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.PAMI-3, no.1, pp.66,75, Jan. 1981

BAKER, Frank B.; HUBERT, Lawrence J. Measuring the Power of Hierarchical Cluster Analysis. *Journal of the American Statistical Association*, Vol. 70, No. 349, pp. 31-38, Mar.1975

BALL, G. H.; HALL, D. J. (1965) ISODATA, A novel method of data analysis and pattern classification. Tech. Rep. NTIS No. AD 699616, Stanford Research Institute, Menlo Park, California.

BANDYOPADHYAY, S.; MAULIK, U. Nonparametric genetic clustering: comparison of validity indices, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol.31, no.1, pp.120,125, Feb 2001

BANDYOPADHYAY, S.; SAHA, S. A Point Symmetry-Based Clustering Technique for Automatic Evolution of Clusters. *IEEE Transactions on Knowledge and Data*

Engineering, vol.20, no.11, pp.1441-1457, Nov. 2008.

BATISTA, Gustavo Enrique de Almeida Prado. **Pré-processamento de dados em aprendizado de máquina supervisionado.** 2003. 231f. Tese (Doutorado em Ciências da Computação e Matemática Computacional)- Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo, São Carlos, 2003.

BEZDEK, J. C.; PAL, N. R. An index of topological preservation for feature extraction. *Pattern Recognition*, Vol. 28, pp. 381-391, 1995.

BEZDEK, James et al. Fuzzy models and algorithms for pattern recognition and image processing. New York: Springer, 2005.

BEZDEK, J. C. **Pattern recognition in handbook of fuzzy computation**, IOP Publishing Ltd., Boston, MA, 1998.

BEZDEK, James. C. Cluster Validity with Fuzzy Sets. Journal of Cybernetics, Vol. 3, Iss. 3, 1973.

BEZDEK, James. C. Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology, Volume 1, Issue 1, pp 57-71, 1974.

BISQUERRA ALZINA, Rafael; CASTELLÃ SARRIERA, Jorge; MARTÍNEZ, Francesc. **Introdução á estatística:** enfoque informático com o pacote estatístico SPSS. Porto Alegre: Artmed, 2004. pp. 255.

BORGES, Ederson. **Um novo algoritmo imunológico artificial para agrupamento de dados.** 2010. 66 f. Dissertação (Mestrado Engenharia Elétrica) Escola de Engenharia, Universidade Presbiteriana Mackenzie, São Paulo, 2010.

BORTOLOTTO, Leandro Sehnem. **O Método de Redes Neurais pelo Algoritmo Kohonen para Clusterização na Shell Orion Data Mining Engine**. 2007. Trabalho de Conclusão de Curso - Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2007.

BOSCARIOLI, Clodis. **Análise de agrupamento baseada na topologia dos dados e em mapas auto- organizáveis.** 2008. 118 f. Tese (Doutor em Sistemas Eletrônicos) - Escola politécnica da Universidade de São Paulo, São Paulo, 2008.

BOUDRAA, A. Dynamic estimation of number of clusters in data sets, *Electronics Letters* , vol.35, no.19, pp.1606,1608, 16 Sep 1999

BOUGUESSA, M.; WANG, S.R. A new efficient validity index for fuzzy clustering, in: Proc. *Third Internat. Conf. on Machine Learning and Cybernetics*, Shanghai, pp. 26–29 August 2004.

BOZKIR, A. Selman; SEZER, Ebru Akcapinar. FUAT–A fuzzy clustering analysis tool, *Expert Systems with Applications*, Volume 40, Issue 3, 15 February 2013, pp. 842-849,

CABENA, Peter et al. **Discovering data mining:** from concept to implementation**.**

New Jersey: Prentice Hall, 1998.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis, *Communications in Statistics - Theory and Methods*, 3: 1, pp. 1-27, 1974.

CARLANTONIO, Lando Mendonça di. **Novas metodologias para clusterização de dados**. 2001. 148p. Dissertação (Mestrado) - Departamento de Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2001.

CASAGRANDE, Diego Paz. **O Módulo da Técnica de Associação pelo Algoritmo Apriori no desenvolvimento da Shell de Data Mining Orion**. 2005. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2005.

CASSETTARI JUNIOR, José Márcio. **O Método de Lógica Fuzzy pelo Algoritmo Gustafson-Kessel na Tarefa de Clusterização da Shell Orion Data Mining Engine.** 2008. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina. 2008.

CHEN, M.Y.; LINKENS, D.A. Rule-base self-generation and simplification for data-driven fuzzy models, *Fuzzy Sets and Systems*, 142 (2004) 243–265.

CHOU, C. H.; SU, M. C.; LAI, E. A New Cluster Validity Measure and Its Application to Image Compression. Pattern Analysis and Applications, Vol. 7, No. 2, pp. 205-220, 2004.

CHO, S. B.; YOO, S.H. Fuzzy Bayesian validation for cluster analysis of yeast cell-cycle data, *Pattern Recognition* (2005).

CLIFFORD, H. T.; STEPHENSON, W. **An introduction to numerical classification**. Academic Press, New York, 1975.

COGGINS, James M., JAIN, Anil K. A spatial filtering approach to texture analysis, Pattern Recognition Letters, Volume 3, Issue 3, May 1985, Pages 195-203

COSTA, Alceu Ferraz. **Mineração de imagens médicas utilizando características de forma.** 2012. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012.

CROTTI JUNIOR, A. et al. Os algoritmos Fuzzy C-Means, Robust C-Prototypes e Unsupervised Robust C-Prototypes aplicados à base de dados das Bacias Hidrográficas da Região Carbonífera de Criciúma.. In: **VI Congresso Sul Brasileiro de Computação (VI SULCOMP)**, 2012, Criciúma, SC. Anais do Sulcomp. Criciúma, SC: UNESC, 2012. v. 6.

CROTTI JUNIOR, Ademar. **O Método de Lógica Fuzzy pelos algoritmos Robust C-Prototypes e Unsupervised Robust C-Prototypes para a Tarefa de Clusterização na Shell Orion Data Mining Engine.** 2010. Trabalho de Conclusão

de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina. 2010.

CRUZ, Armando; CORTEZ, Paulo. Data Mining via Redes Neuronais Artificiais e Máquinas de Vectores de Suporte. Tékhne [online]. 2009, n.12, pp. 99-118.

DATTA, Susmita; DATTA, Somnath . Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, Volume 19, Issue 4, pp. 459-466, 2002.

DAVE, Rajesh N. Validating fuzzy partitions obtained through c-shells clustering, *Pattern Recognition Letters*, Volume 17, Issue 6, 15 May 1996, Pages 613-623

DAVIES, David L.; BOULDIN, Donald W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.PAMI-1, no.2, pp. 224,227, April 1979.

DIAS, Madalena Maria. **Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados.** – Florianópolis, 2001. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina – Programa de Pós-graduação em Engenharia de Produção, 2001. 212 p.

DIMITRIADOU, Evgenia; DOLNIČAR, Sara; WEINGESSEL, Andreas. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, March 2002, Volume 67, Issue 1, pp. 137-159.

DING, Yunfei; HARRISON, Robert F. Relational visual cluster validity (RVCV) *Pattern Recogn*. *Lett.,*November, 2007, Vol. 28, N. 15, nov 2007, pp. 2071-2079.

DINIZ, Carlos Alberto R.; LOUZADA NETO, Francisco. **Data Mining:** uma introdução. São Paulo: Associação Brasileira de Estatística, 2000. 123 p.

DUARTE, Fernando Jorge Ferreira. **Otimização da Combinação de Agrupamentos baseado na Acumulação de Provas pesadas por Índices de Validação e com uso de Amostragem.** 2008. 355 f. Tese (Doutor Electrotécnica e Computadores) - Departamento de Engenharias Área de Engenharia, Universidade de Trás-os-Montes e Alto Douro, Vila Real, 2008.

DUBES, Richard C. How many clusters are best? - An experiment, *Pattern Recognition*, Volume 20, Issue 6, 1987, pp. 645-663.

DUBES, Richard; JAIN, Anil K. Validity studies in clustering methodologies, *Pattern Recognition*, Volume 11, Issue 4, 1979, pp. 235-254.

DUNN, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters', *Cybernetics and Systems*, 3: 3, pp. 32-57, 1973.

DUNN, J. C. Some Recent Investigations of a New Fuzzy Partitioning Algorithm and

its Application to Pattern Classification Problems. *Journal of Cybernetics*, Vol. 4, Iss. 2, 1974.

ESOGBUE, A.O.; LIU, Baoding. Cluster validity for fuzzy criterion clustering, *Computers & Mathematics with Applications*, Volume 37, Issues 11–12, June 1999, pp. 95-100.

ESTER, Martin et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference On Knowledge Discovery And Data Mining (KDD-96), 2. 1996, Portland. **Proceedings…** . Portland: AAAI Press, 1996 p. 226-231.

FACELI, Katti; CARVALHO, André C. P. L. F. de; SOUTTO, Marcilio C. P. Validação de algoritmos de agrupamento. ICMC - Universidade de São Paulo, São Carlos, 2005.

FACELLI, Katti. **Um framework para análise de agrupamento baseado na combinação multi-objetivo de algoritmos de agrupamento.** 2006. 183f. Tese (Doutorado em Ciências da Computação e Matemática Computacional)- Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo, São Carlos, 2006.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic.From data mining to knowledge discovery in databases. AI Magazine, Providence, v.17, n. 3, pp. 37-54, autumn 1996.

FERNÁNDEZ, Luis F. L.; CORBACHO, Fernando. Normality-based validation for crisp clustering, *Pattern Recognition*, Volume 43, Issue 3, March 2010, pp. 782-795.

FIGUEIREDO, Ana Reis de. **Mineração de Dados Citométricos:** Obtenção de Conhecimento de Padrões Celulares para Otimização de Processos Biotecnológicos. Tese (doutorado) – UFRJ/COPPE/ Programa de Engenharia Civil, 2010. 97 p. Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

FRANCO, Cláudia Rita de. **Novos Métodos de Classificação Nebulosa e Validação de Categorias e suas Aplicações a Problemas de Reconhecimento de Padrões**. 2002. 133 f. Dissertação (Mestrado) - IM-NCE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2002.

FRANCO, Rita de, C.; VIDAL, Silva, L.; CRUZ, A.J de Oliveira. A validity measure for hard and fuzzy clustering derived from Fisher's linear discriminant, *Proceedings of the IEEE International Conference on Fuzzy Systems, 2002. FUZZ-IEEE'02.*, vol.2, no., pp.1493,1498, 2002.

FRIGUI, Hichem; KRISHNAPURAM, Raghu. **A Robust Algorithm for Automatic Extraction of an Unknown Number of Clusters from Noisy Data.** Pattern Recognition Letters 17. p. 1223-1232, 1996.

FUKUYAMA, Y.; SUGENO, M. A new method of choosing the number of clusters for the fuzzy C-means method. Proc. 5th Fuzzy Syst. Symp., pp. 247-250, 1989 (in

japanese).

GAN, Goujan; MA, Chaoqun; WU, Jianhong. **Data clustering**: theory, algorithms and applications. Philadelphia: SIAM, 2007.

GATH, I.; GEVA, A.B. Unsupervised optimal fuzzy clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.11, no.7, pp.773-780, Jul 1989.

GAVA, Éverton Marangoni. **O Método de Densidade pelo Algoritmo Density-Based Spatial Clustering of Applications with Noise (DBSCAN) na Tarefa de Clusterização da Shell Orion Data Mining Engine.** 2011. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina. 2011.

GEVA, Amir B. et al. A comparison of cluster validity criteria for a mixture of normal distributed data, Pattern Recognition Letters, Volume 21, Issues 6–7, June 2000, pp. 511-529.

GHELLERE, Samuel Lodetti**. Algoritmo Standard Ant Clustering Algorithm na Tarefa de Clusterização da Shell Orion Data Mining Engine**. 2012. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina. 2012.

GOEBEL, M.; GRUENWALD, L. A Survey of Data Mining and Knowledge Discovery Software Tools. ACM SIGKDD Explorations, New York, v. 1, no. 1, p. 20-33, June. 1999.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data mining**: uma guia prático: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005.

GONZALES, Luis Fernando Planella. **Uma abordagem para mineração de dados e visualização de resultados em imagens batimétricas.** Porto Alegre, 2012. 73 f. Diss. (Mestrado) – Fac. de Informática, PUCRS, 2012.

GUILLET, Fabrice; HAMILTON, Howard J. **Quality measures in data mining**. Chichester: Springer, 2010. xiv, 313 p.

GURRUTXAGA, Ibai et al. SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recognition*, Vol. 43, Issue 10, pp. 3364-3373, Oct 2010.

GTA - Grupo Técnico de Assessoramento (DNPM, CPRM, SIECESC, FATMA, MPF). **Terceiro Relatório de Monitoramento dos Indicadores Ambientais.** Versão 03. Revisão Final. Criciúma, 2009.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering algorithms and validity measures. Tutorial paper in the Proceedings of the SSDBM Conference, 2001.

HALKIDI, M.; VAZIRGIANNIS, M. Managing uncertainty and quality in the classification process. In the Proceedings of SETN Conference, April 2002, Thessaloniki, Greece

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment using multi representatives. Poster paper in the Proceedings of SETN Conference, April 2002, Thessaloniki, Greece

HALKIDI, M.; VAZIRGIANNIS, M. Clustering Validity Assessment: Finding the optimal partitioning of a data set. In the Proceedings of ICDM Conference, California, USA, November 2001

HALKIDI, M.; VAZIRGIANNIS, M.; BATISTAKIS, I. Quality scheme assessment in the clustering process. Proceedings of the 4th PKDD Conference, Lyon, September 2000.

HALKIDI, Maria; BATISTAKIS, Yannis; VAZIRGIANNIS, Michalis. "On clustering validation techniques." In: JIIS 17, 2-3. 2001. pp. 107-145.

HAN, J., KAMBER, M., **Data Mining Concepts and Techniques**. Morgan Kaufman Publishers, San Francisco, USA, 2006.

HANDL, J., **Ant-based methods for tasks of clustering and topographic mapping: extensions, analysis and comparison with alternative techniques.** Master's Thesis, Universität, Erlangen-Nürnberg, Erlangen, Germany, 2003.

HANDL, Julia; KNOWLES, Joshua; KELL, Douglas B. Computational cluster validation in post-genomic data analysis, *Bioinformatics*, August 2005, Vol. 21, N. 15,Aug 2005, pp. 3201-3212.

HARTIGAN, J. Asymptotic distribution for clustering criteria. *Annls. Statistics*, 1978, 6: pp. 117-131.

HASHIMOTO, Wataru; NAKAMURA, Tetsuya; MIYAMOTO, Sadaaki. Comparison and Evaluation of Different Cluster Validity Measures Including Their Kernelization *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol.13, No.3 pp. 204-209, 2009.

HAYKIN, Simon. **Redes neurais**: princípios e técnicas. 2. ed. Porto Alegre: Bookman, 2000.

HE, Yong et al. The optimal cluster number of FCM in complex economic systems, *IEEE International Conference on Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence).* vol., no., pp. 766-770, 1-6 June 2008.

HU, Yating. A robust cluster validity index for fuzzy c-means clustering, *Transportation, Mechanical, and Electrical Engineering (TMEE), 2011 International Conference on*, vol., no., pp. 448,451, 16-18 Dec. 2011.

HUBERT, Lawrence J.; LEVIN. Joel R. A General Statistical Framework for Assessing Categorical Clustering in Free Recall. Psychological Bulletin, Vol. 83 (6), pp. 1072-1080, 1976.

JAIN, Anil K.; DUBES, Richard C. **Algorithms for clustering data.** Englewood Cliffs: Prentice Hall, 1988.

JAIN, Anil K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys (CSUR)*, New York, pp. 264-323, 1999.

JOLLIFFE, Ian T.; PHILIPP, Andreas, Some recent developments in cluster analysis, Physics and Chemistry of the Earth, Parts A/B/C, Volume 35, Issues 9–12, pp. 309-315, 2010.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis**. Prentice Hall, 1982.

KANTARDZIC, M. **Data mining:** Concepts, models, methods, and algorithms. New York, NY: Segunda Edição, John Wiley and Sons, 2011.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data:** an introduction to cluster analysis. New York: Wiley, 1990.

KRZANOWSKI, W.; LAI, Y. A criterion for determining the number of groups in a dataset using sum of squares clustering . *Bioinformatics,* 1985, 44: pp. 23-34.

KIM, Dae-Won; LEE, Kwang H.; LEE, Doheon. On cluster validity index for estimation of the optimal number of fuzzy clusters, *Pattern Recognition*, Volume 37, Issue 10, October 2004, pp. 2009-2025.

KIM, Minho; RAMAKRISHNA, R.S. New indices for cluster validity assessment, *Pattern Recognition Letters*, Volume 26, Issue 15, November 2005, pp. 2353-2363.

KIM, Minho; YOO, Hyunjin; RAMAKRISHNA, R. S. Cluster Validation for High Dimensional Datasets. *Artificial Intelligence: Methodology, Systems, and Applications.* In: 11th International Conference, AIMSA 2004, Varna, Bulgaria, September 2-4, 2004. Proceedings, pp. 178-187, 2004.

KIM, Young-Il et al. A cluster validation index for GK cluster analysis based on relative degree of sharing, *Information Sciences*, Volume 168, Issues 1–4, 2004, pp. 225-242.

KOVÁCS, Ferenc; LEGÁNY, Csaba; BABOS, Attila. Cluster validity measurement techniques. *In: Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, Series AIKED'06, 2006, Madrid, Spain, pp. 388-393.

KWON, S.H. Cluster validity index for fuzzy clustering, *Electronics Letters* , vol.34, no.22, pp.2176, 2177, 29 Oct 1998.

LAGO-FERNÁNDEZ Luis F.; CORBACHO, Fernando. Normality-based validation for crisp clustering, *Pattern Recognition*, Volume 43, Issue 3, March 2010, pp. 782-795.

LAROSE, Daniel T. **Discovering knowledge in data: an introduction to data mining.** Hoboken: Wiley-Interscience, 2005.

LARSEN, B.; AONE, C. Fast and effective text mining using linear time document clustering. *In Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 1999, pp. 16-22.

LAURO, André Luís. **Agrupamento de dados utilizando algoritmo de colônia de formigas**. 2008. Dissertação (Mestrado em Engenharia Civil) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, 2008.

LINDEN, Ricardo. **Algoritmos Genéticos.** Rio de Janeiro: Brasport, 2006.

MARTINS, Denis Piazza. **O Algoritmo de Particionamento K-means na Tarefa de Clusterização da Shell Orion Data Mining Engine. 2007**. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2007.

MARTINS, Eduardo Siqueira. **Aplicação do processo de conhecimento em base de dados a metadados textuais de infraestruturas de dados espaciais**. 2012. 78 p. Dissertação (Mestrado) – Universidade Federal de Viçosa, Viçosa, MG, 2012.

MARY, Angel Latha S.; KUMAR, K.R. Shankar. Density Based Dynamic Data Clustering Algorithm based on Incremental Dataset- *Journal of Computer Science* 8 (5): pp. 656-664, 2012.

MAULIK, U.; BANDYOPADHYAY, S. Performance evaluation of some clustering algorithms and validity indices, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no.12, pp.1650-1654, Dec 2002.

MELANDA, E.A. **Pós-processamento em regras de associação**. 2005. 225f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, São Carlos, 2005.

MILLIGAN, Glenn W.; COOPER, Martha C. An examination of procedures for determining the Number of clusters in a data set. Psychometrika, Colombus, pp. 159-179. Jun. 1985.

MILLIGAN, G. W. A monte carlo study of thirty internal criterion measures for cluster analysis, Psychometrika, 46:187-199, 1981.

MITRA, S.; PAL, S.K. Fuzzy multi-layer perceptron, inferencing and rule generation, *IEEE Trans. Neural Networks*, 6 (1995), pp. 51–63.

MOK, P.Y. et al. A robust adaptive clustering analysis method for automatic identification of clusters, *Pattern Recognition*, Volume 45, Issue 8, August 2012, pp.

3017-3033.

MONDARDO, Ricardo Lineburger. **O algoritmo C4.5 na tarefa de classificação da Shell Orion Data Mining Engine.** 2009. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina. 2009.

MORAES, D. R. S.; EVSUKOFF, A. G.; EBECKEN, N. F. F. Comparação entre métodos de agrupamentos de dados em atributos sísmicos. In: Congresso Brasileiro de P&D em Petróleo e Gás, 3., 2005, Salvador. **Anais**... Salvador: Instituto Brasileiro de Petróleo e Gás, 2005. P. 1-6.

MOUTINHO, Adriano M.; SILVA, Viviane S.R. Aplicação do Algoritmo de Categorização FCM e avaliação das Medidas de Validação ICC e CS, 2002.

NALDI, Murilo Coelho**. Técnicas de combinação para agrupamento centralizado e distribuído de dados**. 2011. 245f. Tese (Doutorado em Ciências da Computação e Matemática Computacional)- Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo, São Carlos, 2011.

NOVASKI, Marcio. **O Teorema de Probabilidade pelo Algoritmo Naive Bayes para a Tarefa de Classificação na Shell Orion Data Mining Engine.** 2012. Trabalho de Conclusão de Curso – Curso e Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina. 2012.

OLIVEIRA, José Valente de; PEDRYCZ, Witold. **Advances in Fuzzy Clustering and Its Applications**. England: John Wiley & Sons, 2007.

PAKHIRA, Malay K.; BANDYOPADHYAY, Sanghamitra; MAULIK, Ujjwal. Validity index for crisp and fuzzy clusters. Pattern Recognition, Volume 37, Issue 3, pp. 487-501, March 2004.

PAL, N.R.; BISWAS, J. Cluster validation using graph theoretic concepts. Pattern Recognition, Vol. 30, Issue 6, pp. 847-857, June 1997.

PASTA, Arquelau. **Aplicação da técnica de Data Mining na base de dados do ambiente de gestão educacional:** um estudo de caso de uma instituição de ensino superior de Blumenau, SC. 2011. 153 f. Dissertação (Mestrado Ciência da Computação Aplicada)- Universidade do Vale do Itajaí, São José, 2011.

PATI, Soumen Kr.; DAS, Asit kr. Cluster analysis of microarray data based on similarity Measurement. *International Journal of Bioinformatics Research*, Vol. 3, Issue 2, 2011, pp. 207-213.

PELEGRIN, Diana Colombo**. A Tarefa de Classificação e o Algoritmo ID3 para Indução de Árvores de Decisão na Shell de Data Mining Orion**. 2005. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2005.

PEREGO, Daniel. **O Método de Lógica Fuzzy pelo algoritmo GATH-GEVA na tarefa de clusterização da Shell Orion Data Mining Engine.** 2009. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina. 2009.

PIRES, Fábio Antero. **Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde**. 2011. Tese (Doutorado em Cardiologia) - Faculdade de Medicina, Universidade de São Paulo, São Paulo, 2011.

PYLE, Dorian. **Data preparation for data mining**. San Francisco: Morgan Kaufmann, 1999.

QUISPE, Germán Carlos Santos. **Reconhecimento de padrões em sensores integrados.** 2005. Tese (Doutorado em Microeletrônica) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2005.

RABELO, Emerson. **Avaliação de Técnicas de visualização para mineração de dados.** 2007. 104 f. Dissertação (Mestrado em Ciência da Computação - Universidade Estadual de Maringá. Maringá, 2007.

RAFTERY, A. A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society.* 48(2), 1986, pp. 249-250.

RAIMUNDO, Lidiane Rosso. **O Algoritmo CART na Tarefa de Classificação da Shell Orion Data Mining Engine**. 2007. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2007.

RAJASEKHAR, B. et al. Quality of Cluster Index Based on Study of Decision Tree. *International Journal of Research in Computer Science*, 2 (1): pp. 39-43, December 2011.

RENDON L. Eréndira et al. Niva: A Robust Cluster Validity. *2 th. WSEAS Int. Conf. Scientific Computation and Soft Computing*, Crete, Greece, 2002, pp. 209-213.

RENDÓN, Eréndira et al. Internal versus External cluster validation indexes. International Journal Of Computers And Communications, Issue 1, Vol. 5, pp. 27-33, 2011.

REZAEE, Babak, A cluster validity index for fuzzy clustering, *Fuzzy Sets and Systems*, Volume 161, Issue 23, 1 December 2010, pp. 3014-3025.

REZENDE, Solange Oliveira. **Sistemas inteligentes:** fundamentos e aplicações. Barueri: Manole, 2005.

RHEE, N.S.; OH, K.W. A validity measure for fuzzy clustering and its use in selecting optimal number of clusters, *in: IEEE Internat. Conf. on Fuzzy Systems*, Vol. 2, 1996, pp. 1020-1025.

RIVERA, F.F; ZAPATA, E.L; CARAZO, J.M. Cluster validity based on the hard tendency of the fuzzy classification, *Pattern Recognition Letters*, Volume 11, Issue 1, January 1990, pp. 7-12

ROMÃO, Wesley. **Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia.** 2002. Tese (Doutorado em Engenharia de Produção) Universidade Federal de Santa Catarina, 2002, Florianópolis

ROUBENS Marc. Pattern classification problems and fuzzy sets, *Fuzzy Sets and Systems*, Volume 1, Issue 4, October 1978, pp. 239-253.

ROUBENS, Marc. Fuzzy clustering algorithms and their cluster validity. European Journal of Operational Research, Volume 10, Issue 3, pp. 294-301, Julho 1982.

ROUSSEEUW, Peter J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, Vol. 20, pp. 53-65, Nov 1987.

SAAD, Mohamed Fadhel; ALIMI, Adel M. Validity Index and number of clusters. *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 1, No 3, January 2012.

SADR, A.; MOMTAZ, A. K. Application of Rosette Pattern for Clustering and Determining the Number of Cluster, Advances in Electrical and Computer Engineering, vol.11, no.3, pp.77-84, 2011.

SAHA, S.; BANDYOPADHYAY, S. Performance Evaluation of Some Symmetry Based Cluster Validity Indexes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* , vol.39, no.4, pp. 420-425, Julho 2009.

SAITTA, Sandro; RAPHAEL, Benny; SMITH, Ian F. C. A Bounded Index for Cluster Validity. *Machine Learning and Data Mining in Pattern Recognition*, Lecture Notes in Computer Science, Vol. 4571, pp. 174-187, 2007.

SANTOS, Scherer dos. **Bee Clustering:** um Algoritmo para Agrupamento de Dados Inspirado em Inteligência de Enxames. 2009. 89 f. Dissertação (Mestre em Computação)- Programa de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

SASSI, Renato José. **Uma arquitetura híbrida para descoberta de conhecimento em bases de dados:** teoria dos rough sets e redes neurais artificiais mapas auto organizáveis.2006. 169 f. Tese (Doutorado em Sistemas Eletrônicos) – Escola Politécnica, Universidade de São Paulo. São Paulo, 2006.

SCOTTI, Ana Paula. **O Método de Redes Neurais com Função de Ativação de Base Radial para a Tarefa de Classificação na Shell Orion Data Mining Engine.** 2010. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina. 2010.

SERRA, Laércio. **A Essência do Business Intelligence.** São Paulo: Berkeley, 2002.

SHANNON, C. E. A Mathematical Theory of Communication, *The Bell System Technical Journal*, 1948, pp. 379-423.

SHIM, Young; CHUNG, Jiwon; CHOI, In-Chan, "A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm," *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on* , vol.1, no., pp.199-204, 28-30 Nov. 2005.

SILVA, Inara Aparecida. **Descoberta de Conhecimento em Base de Dados de Monitoramento Ambiental para Avaliação da Qualidade da Água.** 2007. 119 f. Dissertação (Mestrado em Física e Meio Ambiente) - Instituto de Ciências Exatas e da Terra Programa de Pós-Graduação em Física e Meio Ambiente, Universidade Federal de Mato Grosso, Cuiabá, 2007.

SILVA, Leila R. S. da; GOMIDE, Fernando. Um estudo comparativo entre as funções de validação para agrupamento nebuloso de dados- IV Congresso Brasileiro de Computação, Itajaí, 2004.

SILVA, Marcelino Pereira dos Santos. **SKDQL Uma Linguagem Declarativa de Especificação de Consultas e Processos para Descoberta de Conhecimento em Bancos de Dados e sua Implementação.** 2002. 103 f. Dissertação (Mestre em Ciência da Computação) – Centro de Informática, Universidade Federal de Pernambuco, 2002

SIVANANDAM, S. N.; SUMATHI, S. **Introduction to data mining and its applications.** Berlim: Springer, 2006

SOMMERVILLE, Ian**. Engenharia de software.** São Paulo: Addison-Wesley, 2003

STREHL, A.; GHOSH, J. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research (JMLR)* 3, 2002, pp.583-617.

SUN, H.J.; WANG, S.G.; JIANG, Q.S. FCM-based model selection algorithms for determining the number of clusters, *Pattern Recognition* 37 (2004) pp. 2027–2037.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin; **Introdução ao Data Mining Mineração de dados**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

TANG, Y.G.; SUN, F.C.; SUN, Z.Q. Improved validation index for fuzzy clustering, in: *American Control Conf.*, June 8–10, 2005, Portland, OR, USA.

TASDEMIR, K.; MERENYI, E., "A Validity Index for Prototype-Based Clustering of Data Sets With Complex Cluster Structures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol.41, no.4, pp.1039-1053, Aug. 2011.

THEODORIDIS, Sergios; KOUTROUMBAS, Konstantinos. **Pattern Recognition**, 4th Edition, Printbook , Release 2008.

TSEKOURAS, George E.; SARIMVEIS, Haralambos. A new approach for measuring the validity of the fuzzy c-means algorithm*, Advances in Engineering Software*, Volume 35, Issues 8–9, August–September 2004, pp. 567-575.

VELMURUGAN, T.; SANTHANAM, T. Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points- *Journal of Computer Science* 6 (3): pp. 363-368, 2010.

WANG, Weina; ZHANG, Yunjie. On fuzzy cluster validity indices, *Fuzzy Sets and Systems*, Volume 158, Issue 19, 1 October 2007, pp. 2095-2117.

WINCK Ana Trindade. **3D-Tri: Um Algoritmo de Indução de Árvore de Regressão para Propriedades Tridimensionais: um estudo sobre dados de docagem molecular considerando a flexibilidade do receptor.** 2012. 108 f. Tese (Doutor em Ciência da Computação) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

WINDHAM, M. Cluster validity for the fuzzy c-means clustering algorithm, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 4, no. 4, pp. 357–363, 1982.

WINDHAM, Michael P. Cluster validity for fuzzy clustering algorithms, *Fuzzy Sets and Systems*, Volume 5, Issue 2, March 1981, pp. 177-185.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. **Data mining practical machine learning tools and techniques**. 3. ed. Burlington: Morgan Kaufmann, 2011.

WU, Kuo-Lung; YANG, Miin-Shen, HSIEH, June-Nan. Robust cluster validity indexes, Pattern Recognition, Volume 42, Issue 11, November 2009, Pages 2541-2550.

WU, Kuo-Lung; YANG, Miin-Shen. A cluster validity index for fuzzy clustering, *Pattern Recognition* Letters, Volume 26, Issue 9, 1 July 2005, Pages 1275-1291.

XIE, Xuanli Lisa; BENI, G. A validity measure for fuzzy clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.13, no.8, pp.841-847, Aug 1991.

YANG, Miin-Shen; LAI, Chien-Yo. A Robust Automatic Merging Possibilistic Clustering Method, *IEEE Transactions on Fuzzy Systems,* vol.19, no.1, pp. 26-41, Feb. 2011.

YANG, Miin-Shen; WU, Kuo-Lung, A new validity index for fuzzy clustering, *The 10th IEEE International Conference on Fuzzy Systems, 2001.* vol.1, no., pp. 89-92, 2001.

YANG, Xulei; CAO, Aize; SONG, Qing. A New Cluster Validity for Data Clustering, *Neural Processing Letters*. June 2006, Volume 23, Issue 3, pp. 325-344.

ZAHID, N.; LIMOURI, M.; ESSAID, A. A new cluster-validity for fuzzy clustering, *Pattern Recognition*, Volume 32, Issue 7, July 1999, pp. 1089-1097.

ŽALIK, Krista Rizman; ŽALIK, Borut. Validity index for clusters of different sizes and densities. *Pattern Recognition Letters*, Volume 32, Issue 2, pp. 221-234, January 2011.

ZHANG, Yunjie et al. A cluster validity index for fuzzy clustering, *Information Sciences*, Volume 178, Issue 4, 15 February 2008, Pages 1205-1218

ZHAO, Qinpei; XU, Mantao; FRÄNTI, Pasi. Sum-of-Squares Based Cluster Validity Index and Significance Analysis. *Adaptive and Natural Computing Algorithms.* In: 9th International Conference, ICANNGA 2009, Kuopio, Finland, April, 2009, pp. 313-322, 2009.

APÊNDICE A

Medidas de Qualidade Aplicadas aos Algoritmos de Clusterização da Shell Orion Data Mining Engine

Maicon B. Palhano[1], Merisandra C. de Mattos Garcia[1]

[1]Acadêmico do Curso de Ciência da Computação Unidade Acadêmica de Ciências, Engenharias e Tecnologias – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma – SC – Brasil

[2]Professora do Curso de Ciência da Computação – Unidade Acadêmica de Ciências, Engenharias e Tecnologias – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma– SC - Brasil

{maiconpalhano, mem}@unesc.net

Resumo. *A evolução da tecnologia da informação possibilitou que grandes quantidades de dados pudessem ser processadas e armazenadas em bases de dados, porém quanto maior a quantidade, maior a complexidade de análise. O data mining surge como uma opção para analisar esses grandes conjunto de dados de forma automática, buscando padrões e relações nos dados. O data mining é composto por tarefas e métodos, onde uma dessas tarefas é a de clusterização que gera clusters de dados baseados na semelhança entre os elementos. Porém os clusters gerados por estes algoritmos devem ser avaliados por medidas de qualidade. Este artigo demonstra a aplicação de sete algoritmos de clusterização em quatro bases de dados, a fim de analisar as partições geradas por meio de cinco medidas de qualidade.*

1. Introdução

O expansivo crescimento dos bancos de dados empresariais, governamentais e científicos acarretou a incapacidade humana de interpretar e analisar essas informações, de modo que fossem necessários novos métodos e ferramentas capazes de analisar e extrair informações úteis a partir desses dados armazenados [Fayyad, Piatetsky-Shapiro e Smith 1996]. Sendo assim, de acordo com Sassi (2006) o data mining explora repositórios de dados à procura de padrões e relacionamentos implícitos, encontrando dados que possam prever tendências ou comportamentos futuros.

O *data mining* tem sido aplicado em diversas áreas, com a finalidade, tanto de descrever características do passado, quanto predizer tendências futuras, dentre essas, destacam-se áreas como: finanças, energia, saúde, recursos hídricos, genética e biologia [Han e Kamber 2006].

A etapa de data mining é composta por tarefas e métodos, ambos possuindo suas particularidades, onde a escolha adequada de cada um irá depender do conhecimento acerca do domínio de aplicação do conjunto de dados utilizado [Kantardzic 2011].

Dentre as tarefas, tem-se a classificação, associação, sumarização, regressão, previsão de séries temporais e clusterização [Goldschmidt e Passos 2005].

A clusterização, segundo Han e Kamber (2006) reúne um conjunto de dados em grupos de objetos similares, formando *clusters*, buscando maximizar a semelhança entre objetos do mesmo *cluster* e minimizar a semelhança entre *clusters* [Larose 2005].

Clusterização é uma tarefa onde não há grupos pré-definidos e exemplos que mostrem se os *clusters* encontrados pelos algoritmos são totalmente válidos [Halkidi et al 2002].

Mediante isso, a fim de se identificar a qualidade dos resultados gerados por algoritmos de clusterização quanto a coesão e separação dos *clusters*, deve-se empregar medidas qualidade voltadas a essa tarefa de data mining [Gan, Ma e Wu 2007].

Esta pesquisa enfatiza uma revisão de literatura na área de medidas de qualidade, destacando-se aquelas aplicadas na clusterização, bem como a análise dos resultados gerados pelas medidas de qualidade Coeficiente de Partição, Coeficiente de Partição Entrópica, Xie-Beni, Dunn e C-Index nos algoritmos *Fuzzy C-Means*, *Robust C-Prototypes* (RCP), *Unsupervised Robust C-Prototypes* (URCP), *Density-based Spatial Clustering of Applications With Noise* (DBSCAN), *Ant Based Clustering* (ABC), *Adaptative Ant-Clustering Algorithm* (A²CA) e *Standard Ant Clustering Algorithm* (SACA) da tarefa de clusterização da ferramenta *Shell Orion Data Mining Engine*, avaliando os resultados das suas partições, empregando-se quatro bases de dados.

2. Clusterização

De forma geral, a maioria dos autores [Rezende 2005, Mary e Kumar 2012, Larose 2005] trazem definições muito parecidas com relação à clusterização, onde geralmente, sustentam, como sendo a busca por *clusters* em uma base de dados de forma automática, por meio da similaridade entre os dados, sendo uma tarefa não-supervisionada e descritiva.

Rezaee et al (1998) salientam que como a clusterização é um processo não-supervisionado, onde não se tem conhecimento a priori, é inevitável o uso de algum tipo de avaliação das partições geradas, como medidas de qualidade.

3. Medidas de Qualidade na clusterização

O particionamento de um conjunto de dados resultante da aplicação de um algoritmo de clusterização pode ser tradicional, com valores {0,1} ou fuzzy, com graus de pertinência variando entre [0,1] [Silva e Gomide 2004].

A divisão de um conjunto de dados gerada por algoritmos de clusterização, geralmente não leva em consideração se essa estrutura realmente existe, tornando assim, importante a busca em saber se esses padrões encontrados são válidos, ou se apenas um particionamento sem significado foi gerado pelo algoritmo.

Após a execução dos algoritmos de clusterização, é o momento em que podem surgir algumas dúvidas aos usuários relacionadas ao particionamento, como por exemplo: a) quantos clusters possuem o conjunto de dados; b) os resultados representam os dados reais; c) a maneira usada foi a melhor para particionar o conjunto de dados [Halkidi, Batistakis e Vazirgiannis 2001].

Considerando isso, tem-se a necessidade de métodos capazes de responder a estas perguntas, qualificando o particionamento gerado pelos algoritmos de clusterização, de modo que se possa, além de avaliar o particionamento dos dados, identificar a quantidade exata de *clusters* presentes em um conjunto de dados, dados estes, que geralmente são não rotulados.

Geralmente, essas medidas na clusterização, são definidas pela combinação da coesão e separabilidade dos *clusters*, no qual, coesão refere-se as medidas de proximidade dos conjuntos, como por exemplo, a variância; e separabilidade indica o quão distintos são

os dois *clusters*, como a distância entre objetos de dois *clusters*, por exemplo [Rendón et al 2011].

Após o entendimento das medidas de qualidade na clusterização, realizou-se uma pesquisa na literatura das principais medidas existentes, do período de 1973 a 2013: Dunn (1973), Variance Ratio Criterion (VRC) (1974), Coeficiente de Partição Entrópica (PE) Coeficiente de Partição (CP) (1974), Gamma (1975), C-Index (1976), Performance da Nebulosidade (1978), Davies-Bouldin (DB) (1979), Silhouette (1987), Fukuyama-Sugeno (FS) (1989), Xie-Beni (1991), Variações do Davies-Bouldin e Dunn, baseados em teoria dos grafos (1997), Separação e Dispersão (SD) (2000), S_Dbw (2001), Separação da Partição (2001), Contraste entre classes (2002), CS (2004), PBM (2004), Davies-Bouldin* (2005), Score Function (2007), Sym (2008), Distância Baseada em Pontos Simétricos (2009), COP (2010), Negentropy increment (2010), SV (2010) e OS (2010).

Uma das dificuldades encontradas em trabalhos relacionados à validação de clusters é a necessidade de um framework para interpretação de uma medida. Considerando-se que o valor final encontrado por uma medida seja, por exemplo, o número 10, qual o significado desse valor, ele é bom, ruim ou aceitável [Tan, Steinbach e Kumar 2005].

4. Análise das Medidas de Qualidade Aplicadas aos Algoritmos de Clusterização da Shell Orion Data Mining Engine

A pesquisa desenvolvida consiste na análise de medidas de qualidade para validação dos resultados de algoritmos de clusterização da Shell Orion Data Mining Engine, tendo em vista a importância destas medidas para a avaliação dos resultados gerados por esses algoritmos.

No decorrer do trabalho foi identificada a necessidade de se fazer uma revisão de literatura com relação a aplicação dessas medidas, como por exemplo, quantificar quais medidas são mais usadas, quais algoritmos são mais aplicados nessas pesquisas e quais bases de dados são comumente usadas pelos autores. Sendo assim, a pergunta a qual esta revisão de literatura destina-se a responder é: Quais medidas, algoritmos e bases de dados são comumente usados com relação as medidas de qualidade na tarefa de clusterização?

A fim de responder essa pergunta, a primeira etapa da revisão de literatura consistiu na definição de palavras-chave a serem empregadas para a busca destes trabalhos em periódicos e artigos de conferências internacionais, sendo estas: índice de validação, medidas de validação, medidas de qualidade na clusterização, métricas de validação e análise de cluster, posteriormente a isso buscou-se em repositórios de trabalhos científicos renomados e utilizados em muitas áreas de estudos da computação, sendo estes: *ScienceDirect*, *Association for Computing Machinery* (ACM), *Institute of Eletrical and Eletronics Engineers* (IEEE), *Scirus*, *SpringerLink*, Periódicos CAPES e *Scielo*.

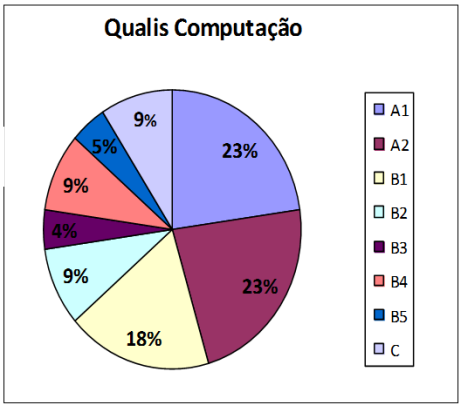
A pesquisa realizada nos periódicos geraram centenas de artigos, porém os que não condiziam com o propósito da pesquisa foram descartados durante o processo.

Outro fator considerado na pesquisa, foi a identificação dos periódicos que fossem cientificamente relevantes, para isto empregou-se a medida por meio da classificação do Qualis (disponível pelo sistema online WebQualis[7]) com relação as áreas de Computação e Engenharias IV. Na área de Computação, como pode ser visto pela figura 1 (a), os periódicos e conferências com Qualis entre A1 e B1 totalizam 64 % do total, seguido de

[7] Sistema que fornece a classificação do Qualis de periódicos, por meio de um site de busca online, no qual é atualizado anualmente pelo CAPES: http://qualis.capes.gov.br/webqualis/principal.seam

27% entre os Qualis B2, B3 e B4 e C, e 9% referentes aos Qualis B3 e B5. Na área de Engenharia IV visível pela figura 1 (b), os Qualis A1, A2 e B1 somam 72% e os demais Qualis totalizam 34% das conferências e periódicos encontrados.

(a)



(b)

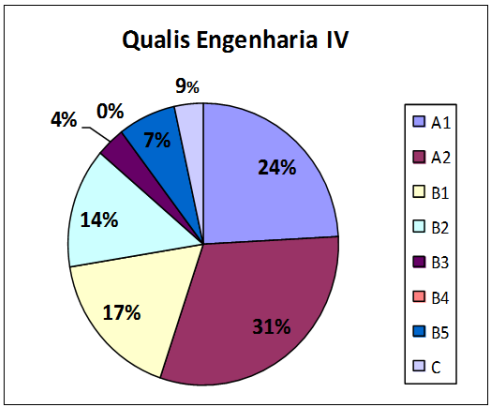


Figura 53. Qualis Computação e Engenharia IV

Pode-se observar pela revisão da literatura realizada, que grande parte dos trabalhos apresentados são internacionais, no qual, dos 38 trabalhos encontrados apenas dois são nacionais. Com isso, ressalta-se o quão pouco a área de medidas de qualidade na clusterização é explorada em pesquisas nacionais, destacando a contribuição significativa do presente trabalho na área de medidas de qualidade na clusterização.

Como o intuito desta revisão de literatura era responder quais medidas, algoritmos e bases de dados são mais aplicados, observou-se que os algoritmos que mais apareceram nos artigos foram: Fuzzy C-Means, K-Means e DBSCAN. As medidas de qualidade mais empregadas nos estudos apresentados foram: CP, CE, XB, Dunn, C-Index, DB, Silhouette, CH, PBM e Fukuyama-Sugeno. Com relação as bases de dados que mais foram utilizadas nos trabalhos, destacaram-se: Iris, Wine, Breast Cancer Wisconsin e Glass.

4.1. Bases de dados utilizadas

Para a aplicação dos algiritmos foram escolhidas as bases de dados *Iris*, *Wine*, *Breast Cancer Wisconsin* e das bacias hidrográficas.

As três primeiras são conjuntos de dados obtidos a partir do UCI Machine Learning Repository[8], no qual consiste em um repositório de dados reais administrado pela Universidade da Califórnia. A quarta base de dados utilizada no trabalho refere-se a dados locais da região de Criciúma. Essa base de dados possui registros das bacias hidrográficas afetadas por poluentes provenientes da extração do carvão mineral na região, a base foi fornecida pelo professor Dr. Carlyle Torres de Bezerra Menezes do curso de Engenharia Ambiental da UNESC.

4.1.1. Base de dados da *Iris*

A base de dados da Iris refere-se às flores da família Iridaceas, com predominância nas espécies Iris-Setosa, Iris-Versicolor e Iris-Virginica. O conjunto de dados é composto por 150 registros, distribuídos em 50 registros para cada espécie e constituída por quatro atributos.

4.1.2. Base de dados da *Breast Cancer Wisconsin*

Breast Cancer é um conjunto de dados que disponibiliza informações sobre pacientes com câncer de mama obtidas a partir dos Hospitais da Universidade de *Wisconsin*, sendo

[8] Repositório de dados disponível em: http://archive.ics.uci.edu/ml/

composta por 699 registros, porém foram retirados 16 registros que não estavam coerentes, restando 683 registros e composta por 10 atributos numéricos.

4.1.3. Base de dados da *Wine*

A base de dados *Wine* corresponde do resultado de análises químicas para determinar a origem de vinhos cultivados em uma região da Itália, porém provenientes de diferentes cultivadores de videiras, sendo composta por 178 registros e 13 atributos.

4.1.4. Base de dados das Bacias Hidrográficas

Esta base de dados corresponde a índices que trazem dados com relação a qualidade da água de três bacias hidrográficas da região de Criciúma. Os dados com indicadores ambientais, somam 1723 registros e 20 atributos.

4.2. Seleção das medidas de qualidade e algoritmos utilizados

Dentre os algoritmos de clusterização da Shell Orion, foram selecionados sete algoritmos para a aplicação das quatro bases de dados citadas anteriormente, sendo estes: FCM, RCP, URCP, DBSCAN, SACA, A²CA e ABC.

Com relação as medidas de qualidade analisadas, foram consideradas cinco medidas: Dunn, C-Index, Xie-Beni, Coeficiente de Partição e Coeficiente de Partição Entrópica, além da ampla aplicabilidade destas medidas em trabalhos da área, outro fator relevante foi o fato das mesmas já estarem implementadas nos algoritmos da Shell Orion[9], sendo os algoritmos FCM, RCP e URCP implementadas as medidas CP, CE e Xie-Beni e nos algoritmos DBSCAN, SACA, A²CA, ABC as medidas Dunn e C-Index.

A fim de uma melhor compreensão dos valores obtidos pelas medidas empregadas neste trabalho, as mesmas são descritas abaixo:

e) **Coeficiente de partição:** no intervalo de [1/C, 1], quanto mais próximo de um o valor da medida, mais compactos e bem separados serão os clusters;

f) **Coeficiente de partição entrópica:** esta medida é o contrário do coeficiente de partição, onde, quanto mais próximo de zero melhor será a qualidade dos clusters;

g) **Dunn:** esta medida encontra clusters compactos e separados quando valores maximizados são encontrados pela mesma;

h) **C-index:** diferentemente das medidas anteriores a C-Index avalia a coesão de cada cluster e não a separação de todos os cluster, onde no intervalo de [0,1], quanto mais próximo de zero, mais bem compacto é o cluster avaliado.

5 Resultados Obtidos

Todos os algoritmos foram executados com as quatro bases de dados, porém, alguns não obtiveram resultados satisfatórios quanto a separação ou a coesão dos clusters gerados, um dos fatores, foi o fato de algumas bases de dados possuírem clusters muito próximos afetando ou o desempenho dos algoritmos ou das medidas consideradas.

[9] Ferramenta acadêmica de *data mining* desenvolvida pelo Grupo de Pesquisa de Inteligência Computacional Aplicada da UNESC desde 2005, contém até o presente momento as tarefas de associação, classificação e clusterização.

5.1. Algoritmos versus base de dados *Iris*

A base de dados da *Iris* possui três *clusters*, porém dois *clusters* são muito próximos e quase sobrepostos, devido a isso alguns autores da literatura aceitam a presença de dois clusters ao invés de três. Pela tabela 1 é possível observar que CP e CE indicam a presença de três *clusters* nos algoritmos RCP e URCP, no algoritmo FCM obteve melhores resultados em dois clusters, com relação ao XB, o mesmo indicou a presença de dois clusters pelo algoritmo FCM e RCP, no algoritmo URCP encontrou um valor ruim.

2 Tabela 50 – *Iris*: Medidas de qualidade algoritmos FCM, RCP e URCP

Alg.	2 Clusters			3 Clusters			4 Clusters		
	CP	CE	XB	CP	CE	XB	CP	CE	XB
FCM	0,6824	0,4936	0,2120	0,54481	0,7880	0,34001	0,4302	1,0649	0,3153
RCP	0.6518	0.5213	0.1051	0.9496	0.1217	3.4356	0.4287	1.054.7	0.2741
URCP	X	X	X	0.7813	0.5487	9,8947	X	X	X

Pela tabela 2 pode-se visualizar que o Dunn indicou um valor ruim para todos os algoritmos, porém a C-Index indicou valores contrários apontando clusters compactos, tendo em vista que o C-Index avalia cada cluster de forma individual.

3 Tabela 51 – *Iris*: Medidas de qualidade algoritmos DBSCAN, ABC, SACA, A²CA

Algoritmos	Nº do Cluster	Dunn	C-Index
DBSCAN	1	0.344837	0.047277
	2	0.344837	0.103562
	3	0.344837	0.130308
ABC	1	0.012217	0.198707
	2	0.012217	0.270436
	3	0.012217	0.227525
SACA	1	0.023212	0.069286
	2	0.023212	0.151758
A²CA	1	0.053226	0.032541
	2	0.053226	0.164694

5.2. Algoritmos versus Base de dados *Wine*

A base de dados *Wine*, segundo a literatura possui três clusters, porém devido a sua alta dimensionalidade alguns algoritmos não a conseguem particioná-la corretamente. Na tabela 3 pode-se observar que todas as medidas obtiveram os melhores valores em dois clusters, indicando como tal o número ideal de clusters.

4 Tabela 52. *Wine:* Medidas de Qualidade algoritmos FCM, RCP e URCP

Alg.	2 *Clusters*			3 *Clusters*			4 *Clusters*		
	CP	CE	XB	CP	CE	XB	CP	CE	XB
FCM	0,7152	0,4456	0,1974	0.5723	0.7456	0.2477	0.5070	0.9344	0.2136
RCP	0.5039	0.6891	0.1432	0.3850	0.1217	1.0180	0.2668	1.3488	0.03495
URCP	0.7883	0.3415	1.4340	X	X	X	X	X	X

5 Pela tabela 4 observa-se que o Dunn mostrou uma boa clusterização, pois foi encontrado valores bons para todos os algoritmos, exceto o DBSCAN que não

conseguiu clusterizar a base da *Wine* e o C-Index indicou clusters compactos para todos os algoritmos exceto o DBSCAN, assim como o Dunn.

6 Tabela 53. *Wine*: Medidas de Qualidade algoritmos DBSCAN, ABC, SACA e A²CA

Algoritmos	Nº do *Cluster*	Dunn	C-Index
DBSCAN	X	X	X
ABC	1	3.612698	0.277684
	2	3.612698	0.426757
	3	3.612698	0.422466
SACA	1	3.469666	0.345134
	2	3.469666	0.247756
A²CA	1	3.955415	0.416503
	2	3.955415	0.271009
	3	3.955415	0.289912

5.3. Algoritmos versus base de dados *Breast Cancer Wisconsin* (BCW)

A base de dados BCW possui três clusters, porém os mesmos possuem uma certa sobreposição. Na tabela 5 observa-se que as medidas CP e CE em todos os casos mostraram dois clusters como o ideal, contudo o XB indicou três para o algoritmo FCM e um valor não tão bom para dois clusters dos algoritmos RCP e URCP.

7 Tabela 54. BCW: Medidas de Qualidade algoritmos FCM, RCP, URCP

Alg.	2 *Clusters*			3 *Clusters*			4 *Clusters*		
	CP	CE	XB	CP	CE	XB	CP	CE	XB
FCM	0.6537	0.5256	0.2909	0.4814	0.8896	0.2125	0.3268	1.2188	5.6621
RCP	0.7828	0.3505	1.4569	0.5520	0.7219	4.1005	0.4600	0.9571	1924.48
URCP	0.7883	0.3415	1.4340	X	X	X	X	X	X

Pela tabela 6 observa-se que os algoritmos ABC e SACA não conseguiram particionar corretamente os dados do BCW e consequentemente não foi possível identificar os valores das medidas nesses casos. O Dunn identificou valores ruins para DBSCAN e A²CA, porém o C-Index apontou clusters compactos neste caso.

8 Tabela 55. BCW: Medidas de Qualidade algoritmos DBSCAN, ABC, SACA e A²CA

Algoritmos	Nº do Cluster	Dunn	C-Index
DBSCAN	1	0.402570	0.014563
	2	0.402570	0.468115
ABC	X	X	X
SACA	X	X	X
A²CA	1	0.012730	0.114613
	2	0.012730	0.205437
	3	0.012730	0.108499
	4	0.012730	0.114379
	5	0.012730	NaN
	6	0.012730	0.120600

5.3. Algoritmos versus base de dados das Bacias Hidrográficas (BH)

Os dados desta base de dados foram aplicados em alguns trabalhos do Grupo de Inteligência Computacional Aplicada[10], porém em nenhum deles chegou a conclusão do número ideal de clusters, cada trabalho resulta em uma clusterização diferente. Na tabela 7 pode-se visualizar que as medidas CP, CE encontraram os melhores valores em todos os algoritmos em dois clusters, XB indicou a presença de quatro clusters pelo algoritmo FCM e valores ruins para os algoritmos RCP e URCP.

9 Tabela 56. BH: Medidas de Qualidade algoritmos FCM, RCP e URCP

Alg.	2 *Clusters*			3 *Clusters*			4 *Clusters*		
	CP	**CE**	**XB**	**CP**	**CE**	**XB**	**CP**	**CE**	**XB**
FCM	0.7658	0.3831	0.1627	0.6826	0.5619	0.2059	0.64424	0.6821	0.0949
RCP	0.6997	0.4612	12.1975	0.6231	0.6470	22.1062	X	X	X
URCP	0.7425	0.4006	35.9647	X	X	X	X	X	X

Pela tabela 8 observa-se que os algoritmos DBSCAN e o ABC não conseguiram clusterizar esta base de dados, com relação ao Dunn nos algoritmos SACA e A²CA obteve bons valores bem como o C-Index.

10 Tabela 57. BH: Medidas de Qualidade algoritmos DBSCAN, ABC, SACA e A²CA

Algoritmos	Nº do *Cluster*	Dunn	C-Index
DBSCAN	X	X	X
ABC	X	X	X
SACA	18	1.790367	0.0250 - 0.0962
A²CA	13	7.0066	0.0187 - 0.0991

Observa-se na aplicação dos sete algoritmos que nem todos conseguiram atender os requisitos mínimos da clusterização, mostrado pelo Capítulo 2, pois quando tinham-se bases de dados com: muitas dimensões (*Iris* e *Wine*), algum tipo de ruído (bacias hidrográficas), ou com muitos elementos ou atributos (bacias hidrográficas e *breast cancer Wisconsin*) alguns algoritmos, como foi o caso do DBSCAN, ABC e SACA não conseguiram particionar corretamente os clusters.

Com relação as cinco medidas de qualidade aplicadas pode-se afirmar que as medidas que melhor mostraram o número ideal de clusters e se estes eram compactos e bem separados, sem contrariedades foram o Coeficiente de Partição, Coeficiente de Partição Entrópica e C-Index. Já as medidas que mais foram arbitrárias em seus resultados foram a Xie-Beni e Dunn.

6. Considerações Finais

Este artigo mostrou uma pesquisa realizada das medidas de qualidade existentes e também uma revisão da literatura quanto as aplicações das medidas, no qual tornou-se visível quais medidas, algoritmos e bases de dados são mais utilizadas.

No trabalho empregaram-se sete algoritmos de clusterização da Shell Orion, sendo cada um executado com quatro bases de dados diferentes. O particionamento realizado pelos algoritmos para cada base de dados foram analisados por meio de medidas de qualidade, chegando-se a conclusão de que a maioria dos algoritmos da Shell Orion conseguiram

[10] Grupo de pesquisa do curso de Ciência da Computação da UNESC.

identificar corretamente os clusters quando a base de dados era pequena, porém com bases de dados maiores como foi o caso das bacias hidrográficas e *Wine*, alguns algoritmos não conseguiram agrupar de forma correta os dados.

Com relação as medidas de qualidade observou-se que estas são muito relativas, pois nem sempre uma medida que funciona corretamente em uma base de dados ou algoritmo específico se comportará deste mesmo modo com outra base de dados ou algoritmo. Destaca-se também a vulnerabilidade das medidas com relação a ruídos, bases de dados com clusters sobrepostos, e bases de dados com múltiplas dimensões.

Com isso, conclui-se que medidas de qualidade na clusterização é uma área relativamente nova e ainda está em constantes estudos e buscando cada vez mais novas medidas que se comportem melhor a uma determinada aplicação, pois não existe uma medida universal que consiga atender a todo algoritmo ou base de dados, porém é a única forma de auxiliar o usuário quanto ao número ideal de *clusters* e a medir a qualidade dos clusters gerados.

Referências

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic.From data mining to knowledge discovery in databases. AI Magazine, Providence, v.17, n. 3, pp. 37-54, autumn 1996.

GAN, G., MA, C. e WU, J. **Data clustering**: theory, algorithms and applications. Philadelphia: SIAM, 2007.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data mining**: uma guia prático: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment using multi representatives. Poster paper in the Proceedings of SETN Conference, April 2002, Thessaloniki, Greece

HALKIDI, M.; VAZIRGIANNIS, M. Clustering Validity Assessment: Finding the optimal partitioning of a data set. In the Proceedings of ICDM Conference, California, USA, November 2001 [Halkidi et al 2002].

HAN, J., KAMBER, M., **Data Mining Concepts and Techniques**. Morgan Kaufman Publishers, San Francisco, USA, 2006.

KANTARDZIC, M. **Data mining:** Concepts, models, methods, and algorithms. New York, NY: Segunda Edição, John Wiley and Sons, 2011.

LAROSE, Daniel T. **Discovering knowledge in data: an introduction to data mining.** Hoboken: Wiley-Interscience, 2005.

MARY, Angel Latha S.; KUMAR, K.R. Shankar. Density Based Dynamic Data Clustering Algorithm based on Incremental Dataset- *Journal of Computer Science* 8 (5): pp. 656-664, 2012.

RENDÓN, Eréndira et al. Internal versus External cluster validation indexes. International Journal Of Computers And Communications, Issue 1, Vol. 5, pp. 27-33, 2011.

REZAEE, Babak, A cluster validity index for fuzzy clustering, *Fuzzy Sets and Systems*, Volume 161, Issue 23, 1 December 2010, pp. 3014-3025.

REZENDE, Solange Oliveira. **Sistemas inteligentes:** fundamentos e aplicações. Barueri: Manole, 2005.

SASSI, Renato José. **Uma arquitetura híbrida para descoberta de conhecimento em bases de dados:** teoria dos rough sets e redes neurais artificiais mapas auto organizáveis.2006. 169 f. Tese (Doutorado em Sistemas Eletrônicos) – Escola Politécnica, Universidade de São Paulo. São Paulo, 2006.

SILVA, Leila R. S. da; GOMIDE, Fernando. Um estudo comparativo entre as funções de validação para agrupamento nebuloso de dados- IV Congresso Brasileiro de Computação, Itajaí, 2004.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin; **Introdução ao Data Mining Mineração de dados**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.