

**LIGA DE ENSINO DO RIO GRANDE DO NORTE
FACULDADE NATALENSE PARA O DESENVOLVIMENTO
DO RIO GRANDE DO NORTE**

SINARA DA ROCHA MARTINS

CLUSTERIZAÇÃO DE DADOS ATRAVÉS DA LÓGICA FUZZY

CLUSTERIZAÇÃO DE DADOS ATRAVÉS DA LÓGICA FUZZY

Monografia apresentada ao Curso de Graduação em Sistemas de Informação da Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte como requisito obrigatório para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Thaís Rosendo do Amaral, Msc.

NATAL / RN.

2008

SINARA DA ROCHA MARTINS

CLUSTERIZAÇÃO DE DADOS ATRAVÉS DA LÓGICA FUZZY

Monografia apresentada ao curso de Bacharelado em Sistemas de Informação da Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte e aprovada em _____ de _____ de _____ pela banca examinadora composta

CLUSTERIZAÇÃO DE DADOS ATRAVÉS DA LÓGICA FUZZY

Ytalo Rosendo do Amaral, Ms.
Ytalo Rosendo do Amaral, Ms.

Monografia apresentada ao Curso de Graduação em Sistemas de Informação da Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte como requisito obrigatório para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Ytalo Rosendo do Amaral, Ms.

Orientador: Ytalo Rosendo do Amaral (Ms.)
Monografia (Bacharelado em Sistemas de Informação) – Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte.

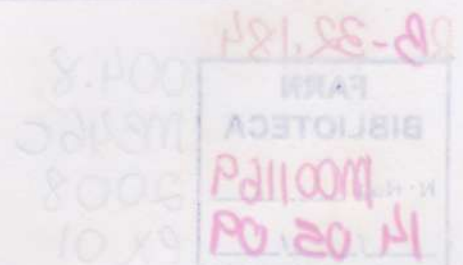
1. Lógica Fuzzy – Monografia. 2. Inteligência Artificial – Monografia. 3. Mineração de Dados – Monografia. 4. Clusters – Monografia. 1. Título.

CDU 004.8

RNT/ARNBC

NATAL / RN.

2008



CLUSTERIZAÇÃO DE DADOS ATRAVÉS DA LÓGICA FUZZY

Monografia apresentada ao Curso de
Graduação em Sistemas de Informação da
Faculdade Natalense para o
Desenvolvimento do Rio Grande do Norte
como requisito obrigatório para obtenção
do grau de Bacharel em Sistemas de Informação.

Martins, Sinara da Rocha.

Clusterização de dados através da lógica Fuzzy/ Sinara da Rocha Martins.

– Natal: FARN, 2008.

58f.

Orientador: Ytalo Rosendo do Amaral, Ms.

Orientador: Ytalo Rosendo do Amaral (Msc.)

Monografia (Bacharelado em Sistemas de Informação) – Faculdade
Natalense para o Desenvolvimento do Rio Grande do Norte.

1. Lógica Fuzzy – Monografia. 2. Inteligência Artificial – Monografia.
Mineração de Dados – Monografia. 4. Clusters – Monografia. I.Título.

RN/FARN/BC

CDU 004.8



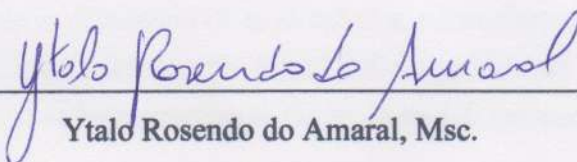
004.8
M346C
2008
ex 01

NATAL/RN
2008

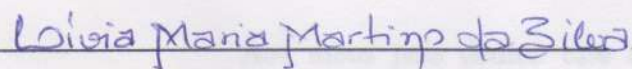
SINARA DA ROCHA MARTINS

CLUSTERIZAÇÃO DE DADOS ATRAVÉS DA LÓGICA FUZZY

Monografia apresentada ao curso de Bacharelado em Sistemas de Informação da Liga de Ensino do Rio Grande do Norte – Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte e, aprovada em _____ de _____ de _____ pela banca examinadora composta pelos seguintes membros:

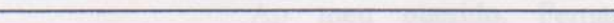

Ytalo Rosendo do Amaral, Msc.

Orientador



Livia Maria Martins da Silva, Msc.

Membro Examinador


Bartira Paraguaçu Falcão Dantas Rocha, Msc.

Membro Examinador

AGRADECIMENTOS

Este trabalho foi construído com muito entusiasmo e dedicação. Por isso, quero agradecer a todos com as mais sinceras palavras. Com o término deste trabalho, quero atingir minhas metas no presente, como também desejo alcançar minhas metas para o futuro, dando continuidade na pesquisa sobre o tema aqui pesquisado.

Assim, agradeço a todos os meus professores, pelos ensinamentos ministrados e, pelo estímulo aos desafios acadêmicos que deles recebi.

Agradeço em especial ao meu Professor e Orientador Ytalo Rosendo do Amaral, que me ajudou no aprofundamento do tema escolhido para esta pesquisa e que, necessariamente, distinguiu os erros e acertos deste trabalho, com sua atenção e vasta experiência, sempre corrigindo e aprimorando minha pesquisa. Sem ele eu não teria recebido meu primeiro prêmio para ingresso de iniciação científica.

Agradeço a minha professora e coordenadora Livia Maria Martins, pelo bom exemplo a seguir e, pela destreza de seus conhecimentos e mim profícuos.

Agradeço à Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte, pelo seu suporte físico, pela legitimidade de suas ações e, particularmente nas pessoas de seus Diretores, Professores e Funcionários, pela prontidão e suporte que me ofereceram para o desenvolvimento e conclusão desta oportunidade produtiva e valerosa que é a produção de conhecimento.

E enfim, agradeço a todos que, de alguma forma, contribuíram na elaboração deste trabalho.

Aos meus pais dedico este trabalho, com todo carinho, como resultado de meu esforço e também do apoio incondicional que deles recebi em minha aprendizagem, sempre ouvindo de minha mãe que sem estudo o homem nada é.

Ao meu querido, Fernando, pelo apoio e paciência constante.

Ao meu irmão Igor, aos meus amigos, e a todos aqueles que fazem parte significativa da minha vida.

AGRADECIMENTOS

Este trabalho foi construído com muito entusiasmo e dedicação. Por isso, venho agradecer a todos com as mais sinceras palavras. Com o término deste trabalho, anseio atingir minhas metas no presente, como também desejo alcançar minhas metas num futuro próximo, dando continuidade na pesquisa sobre o tema aqui pesquisado.

Assim, agradeço a todos os meus professores, pelos ensinamentos ministrados e, pelo estímulo aos desafios acadêmicos que deles recebi.

Agradeço em especial ao meu Professor e Orientador Ytalo Rosendo do Amaral, que me ajudou no aprofundamento do tema escolhido para esta pesquisa e que, necessariamente, distinguiu os erros e acertos deste trabalho, com sua atenção e vasta experiência, sempre corrigindo e aprimorando minha pesquisa. Sem ele eu não teria recebido meu primeiro prêmio num congresso de iniciação científica.

Agradeço a minha professora e coordenadora Livia Maria Martins, pelo bom exemplo a seguir e, pela destreza de seus conhecimentos a mim proferidos.

Agradeço à Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte, pelo seu suporte físico, pela legitimidade de suas ações e, particularmente nas pessoas de seus Diretores, Professores e Funcionários, pela prontidão e suporte que me ofereceram para o desenvolvimento e conclusão desta oportunidade produtiva e valorosa que é a produção de conhecimento.

E enfim, agradeço a todos que, de alguma forma, contribuíram na elaboração deste trabalho.

"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original."

Albert Einstein

RESUMO

Atualmente arquivar dados por meio computacional é uma necessidade para instituições que possuem dados em papel sendo ultrapassados, mas além do armazenamento é necessário um mecanismo que também efetue a leitura, a organização e a extração de informação significativa desses dados, de forma que haja um aproveitamento dessa informação para obtenção de conhecimento válido em possíveis tomadas de decisão. Nesse contexto, a caracterização de dados através da lógica fuzzy, é viável e tem sido aplicada, devido, em grande parte, a naturalidade em que ela agrupa conjuntos de dados e, a proximidade desse processo com o modo de raciocínio humano, graças à lógica fuzzy. A caracterização de dados é uma técnica aplicada a diversos áreas como mineração de dados, estatística, reconhecimento de padrões e processamento de imagens. A caracterização com fuzzy para partição de dados, pode gerar uma das partes, com grau de similaridade e dissimilaridade, identificando o centro de cada agrupamento e, a qual grupo cada elemento pertence. Neste trabalho é proposta um procedimento de segmentação baseado no algoritmo de clusterização Fuzzy c-means, para segmentar dois conjuntos de dados bidimensionais, que representam o salário líquido de 119 funcionários de uma empresa. Para melhoramento dos resultados, utilizou-se a distância euclidiana como método de similaridade entre os indivíduos e os clusters. Constatando-se de uma pesquisa aplicada com metodologia qualitativa, os resultados obtidos demonstram que o uso do algoritmo Fuzzy c-means é possível extrair informações preciosas dos dados. E assim, de acordo com a quantidade de clusters, dificulta-se ou facilita-se, a percepção de variáveis significativas das informações de cada indivíduo.

Palavras-chave: Mineração de dados. Lógica Fuzzy. Clusterização. Fuzzy C-means.

"A mente que se abre a uma nova idéia jamais voltará ao seu tamanho original."

Albert Einstein

RESUMO

Atualmente arquivar dados por meio computacional é uma necessidade clara. Instituições que arquivam dados em papel estão ultrapassadas. Mas além do arquivamento é necessário um mecanismo que também efetue a leitura, a organização e a extração de informação significativa desses dados, de forma que haja um aproveitamento dessa informação para aplicação do conhecimento valioso em possíveis tomadas de decisão. Nesse contexto, a clusterização de dados através da lógica fuzzy, é viavelmente aplicada, devido, em grande parte, a naturalidade em que ela agrupa conjunto de dados e, a proximidade desse processo com o modelo de raciocínio humano, graças à lógica fuzzy. A clusterização de dados é uma técnica aplicada a diversas áreas como mineração de dados, estatística, reconhecimentos de padrões e processamento de imagens. A clusterização com fuzzy gera partições de dados, próximas uma das outras, com graus de similaridade e dissimilaridade, identificando o centro de cada agrupamento e, a qual grupo cada elemento pertence. Neste trabalho é proposto um procedimento de segmentação baseado no algoritmo de clusterização Fuzzy c-means, para segmentar dois conjuntos de dados bidimensionais, que representam o salário líquido de 119 funcionários de uma empresa. Para melhoramento dos resultados, utilizou-se a distância euclidiana como método de similaridade entre os indivíduos e os clusters. Constituindo-se de uma pesquisa aplicada com metodologia qualitativa, os resultados obtidos demonstram que através do algoritmo Fuzzy c-means é possível extrair informações preciosas dos dados. E que, de acordo com a quantidade de clusters, dificulta-se ou facilita-se, a percepção de variações significativas das informações de cada indivíduo.

Palavras-chave: Mineração de dados. Lógica Fuzzy. Clusterização. Fuzzy C-means.

ABSTRACT

Currently archive data via computer is a clear need. Institutions that store data on paper are overcome. But beyond filing is a necessary mechanism to also make the reading, the organization and extract meaningful information from these data, so that there is a use that information for application of knowledge valuable in possible decision-making. In this context, clustering the data through the fuzzy logic, is feasible to apply, because, largely, the naturalness where her groups data sets and, the proximity of this process with the model of human reasoning, because the fuzzy logic. The clustering in the data is a technique applied to various fields as data-mining, statistics, pattern recognition and processing of images. The clustering with fuzzy generates partitions data, close together, with degrees of similarity and dissimilarity, identifying the center of each grouping and, to which group each element belongs. This paper proposed a procedure for targeting based on the algorithm of clustering Fuzzy c-means, to target two sets of two-dimensional data, representing the net salary of 119 employees of a company. To improve results, using the Euclidean distance as a method of similarity between individuals and the clusters. Being is an applied research with qualitative methodology, where the results show that by Fuzzy c-means algorithm you can extract valuable information from data. And that, according to the number of clusters, is hampered or facilitated, the perception of significant variations of the information of each individual.

Key Words: Data mining. Fuzzy Logic. Clustering. Fuzzy C-means.

LISTA DE TABELAS

TABELA 1 – Porcentagem da mutação dos elementos com 5 clusters.51

TABELA 2 – Porcentagem da mutação dos elementos com 3 clusters.53

QUADRO 3 – Localização dos 3 centróides dos dados do mês de setembro.42

QUADRO 4 – Dados de entrada do mês de outubro.44

QUADRO 5 – Localização dos 5 centróides dos dados do mês de outubro.45

QUADRO 6 – Localização dos 3 centróides dos dados do mês de setembro.47

QUADRO 7 – Localização dos 3 centróides dos dados do mês de outubro.48

QUADRO 8 – Mutação dos elementos com 5 clusters.51

QUADRO 9 – Funcionários que apresentaram alto grau de variação salarial quando clusterizados com 5 clusters.52

QUADRO 10 – Mutação do elemento com 3 clusters.53

LISTA DE QUADROS

QUADRO 1 – Função de pertinência para os conjuntos do exemplo para “jovem”.	30
QUADRO 2 – Dados de entrada do mês de setembro.	41
QUADRO 3 – Localização dos 5 centróides dos dados do mês de setembro.	42
QUADRO 4 – Dados de entrada do mês de outubro.	44
QUADRO 5 – Localização dos 5 centróides dos dados do mês de outubro.	45
QUADRO 6 – Localização dos 3 centróides dos dados do mês de setembro.	47
QUADRO 7 – Localização dos 3 centróides dos dados do mês de outubro.	48
QUADRO 8 – Mutação dos elementos com 5 clusters.	51
QUADRO 9 – Funcionários que apresentaram alto grau de variação salarial quando clusterizados com 5 clusters.	52
QUADRO 10 - Mutação do elemento com 3 clusters.	53
FIGURA 1 – Função de pertinência e complemento dos conjuntos difusos A e B.	13
FIGURA 2 – Sistema Lógico Fuzzy.	34
FIGURA 3 – Exemplo de clusterização <i>very x hard</i>	36
FIGURA 4 – Clusters Setembro (5 Centróides dos dados).	42
FIGURA 5 – Clusters Outubro (5 Centróides dos dados).	45
FIGURA 6 – Clusters Setembro (3 Centróides dos dados).	47
FIGURA 7 – Clusters Outubro (3 Centróides dos dados).	49

LISTA DE ILUSTRAÇÕES

FIGURA 1 – Classificação das galáxias.	19
FIGURA 2 – Clusterização de um conjunto de dados. (a) Dados originais, (b) Divisão em dois clusters, (c) Divisão em três clusters e (d) divisão em cinco clusters. Com cada cluster sendo representado com uma cor diferente.	22
FIGURA 3 – Etapas do processo de clusterização.	23
FIGURA 4 – Exemplo de algoritmo de clusterização de dados hierárquico aglomerativo e divisivo	26
FIGURA 5 – Clusterização de um conjunto de dados em 3 grupos baseadas no algoritmo K-means.	27
FIGURA 6 – Função de pertinência e operação de união dos conjuntos difusos A e B.	32
FIGURA 7 – Função de pertinência e operação de interseção dos conjuntos difusos A e B.	33
FIGURA 8 – Função de pertinência e complemento dos conjuntos difusos A e B.	33
FIGURA 9 – Sistema Lógico Fuzzy.	34
FIGURA 10 – Exemplo de clusterização <i>fuzzy x hard</i>	36
FIGURA 11 – Clusters Setembro (5 Centróides dos dados).	42
FIGURA 12 – Clusters Outubro (5 Centróides dos dados).	45
FIGURA 13 – Clusters Setembro (3 Centróides dos dados).	47
FIGURA 14 – Clusters Outubro (3 Centróides dos dados).	49

SUMÁRIO

1 INTRODUÇÃO	14
1.1 PROBLEMA	14
1.2 OBJETIVOS	14
1.2.1 Objetivos Gerais	14
1.2.2 Objetivos Específicos	15
1.3 JUSTIFICATIVA	15
2 REFERENCIAL TEÓRICO	17
2.1 DATA MINING	17
2.1.1 Tarefas da Mineração de Dados	18
2.1.1.1 Classificação	18
2.1.1.2 Estimativa	19
2.1.1.3 Previsão	20
2.1.1.4 Associação ou análise de cesta de supermercado	20
2.1.1.5 Descrição	21
2.1.1.6 Segmentação ou clustering	21
2.2 CLUSTERS	21
2.2.1 Etapas do Processo de Clusterização	23
2.2.1.1 Pré-processamento e seleção de variáveis	24
2.2.1.2 Medidas de similaridade	24
2.2.1.3 Algoritmos de clusterização	24
2.2.1.4 Validação e análise dos resultados	25
2.2.2 Métodos de Clusterização	25
2.2.2.1 Métodos hierárquicos	25
2.2.2.2 Métodos particionais	26
2.2.2.2.1 K-means (KM)	27
2.3 LÓGICA FUZZY	28
2.3.1 Conjuntos Fuzzy	29
2.3.1.1 Operações dos conjuntos fuzzy	31
2.3.1.1.1 União	31
2.3.1.1.2 Interseção	32
2.3.1.1.3 Complemento	33
2.3.2 Regras Fuzzy	34

2.3.3 A Estrutura do Sistema Fuzzy	34
2.3.3.1 Fuzzyficação	35
2.3.3.2 Inferência	35
2.3.3.3 Desfuzzyficação	35
2.4 COMPLEXO FUZZY CLUSTERING	36
2.4.1 Fuzzy C-Means (FCM)	37
3 METODOLOGIA	39
4 ANÁLISE DE RESULTADOS	40
4.1 CLUSTERIZAÇÃO COM 5 CLUSTERS	40
4.1.1 Clusterização com 5 Clusters dos Salários do Mês de Setembro	40
4.1.2 Clusterização com 5 Clusters dos Salários do Mês de Outubro	43
4.2 CLUSTERIZAÇÃO COM 3 CLUSTERS	46
4.2.1 Clusterização com 3 Clusters dos Salários do Mês de Setembro	46
4.2.2 Clusterização com 3 Clusters dos Salários do Mês de Outubro	48
5 CONCLUSÃO	50
5.1 COMPARAÇÃO DOS CLUSTERS: SETEMBRO X OUTUBRO	50
5.1.1 Comparação dos Dois Meses com 5 Clusters	50
5.1.1 Comparação dos Dois Meses com 3 Clusters	52
5.2 COMPARAÇÃO DOS PROCEDIMENTOS: 5 CLUSTERS X 3 CLUSTERS	53
5.3 CONCLUSÃO SOBRE O ALGORITMO FUZZY C-MEANS	54
6 REFERÊNCIAS	55

produzidos em diversos campos e setores econômicos, e com isso, necessitando-se de mecanismos que além de classificar a tarefa de organização dos dados, extraia informações significativas dos mesmos de forma a aproveitar o conhecimento obtido para posteriormente escolher um auxílio de apoio diagnóstico. É partindo desta situação, como usufruir de um algoritmo para realizar adequadamente esta tarefa de organizar e extrair informações significativas dos dados?

1.2 OBJETIVOS

1.2.1 Objetivos Gerais

1 INTRODUÇÃO

É fato que hoje o mundo encontra-se num processo de informatização onde dados arquivados em fichas de papel é considerado um episódio retrogrado. Dados precisam ser conservados e organizados, de maneira que tenham durabilidade e que exibam informações significativas. O foco de estudo dessa pesquisa se constitui na utilização de dados clusterizados (agrupados), por meio da lógica fuzzy.

Diante dos fundamentos articulados no trabalho, ampliar-se-á uma visão englobando data mining, clusterização e lógica fuzzy, visto que não é plausível um entendimento solitário desses temas, e sim que, precisa-se entender todo o contexto, para que se possa ter uma boa articulação da clusterização de dados com fuzzy.

Nesse contexto, o desafio aqui proposto pode ser resumido com duas questões básicas: uma boa forma para se extrair conhecimento estratégico e, devido o emprego da lógica fuzzy, chegar o mais próximo possível da realidade.

1.1 PROBLEMA

Através de avanços tecnológicos, cresce diariamente a quantidade de dados produzidos em diversos campos e setores mundiais, e com isso, necessita-se de mecanismos que além de efetuar a tarefa de organização desses dados, extraia informação significativa dos mesmos de forma a aproveitar o conhecimento obtido para posteriormente efetivar um auxílio de apoio diagnóstico. E partindo desta situação, como usufruir de um algoritmo para realizar adequadamente esta tarefa de organizar e extrair informação significativa dos dados?

1.2 OBJETIVOS

1.2.1 Objetivos Gerais

O objetivo do presente trabalho é utilizar uma das técnicas do Data Mining denominada Clusterização para detectar padrões de comportamentos válidos, empregando o mecanismo da lógica fuzzy para melhor organização dos dados.

1.2.2 Objetivos Específicos

- I. Pré-processamento e seleção de dados a serem usados na pesquisa;
- II. Selecionar a estrutura para extração dos dados;
- III. Identificar o algoritmo para o processo de clusterização;
- IV. Infundir o processo de Data Mining para clusterizar os dados aplicando a lógica fuzzy como método de similaridade;
- V. Validação e análise do conhecimento adquirido.
- VI. Aplicar o conhecimento adquirido como ferramenta para tomada de decisão.

1.3 JUSTIFICATIVA

Estamos em um mundo sistematicamente informatizado, onde a computação numérica está na maior parte dos equipamentos eletrônicos usados diariamente, e por trás dessas máquinas computacionais provêm códigos, programação e uma lógica abraçando tudo isso para que tarefas sejam corretamente executadas.

A luz desses pressupostos, programas cada vez mais estruturados vêm gerando uma crescente quantidade de dados, que precisam ser armazenados organizadamente, possibilitando que através dessa organização, possa se ter uma leitura e interpretação de características contida nesse corpus, que muitas vezes estão ocultas sob a enorme coleção de dados armazenados.

O processo de Data Mining consiste em recuperar aquelas informações, antes camufladas nos dados, aonde a técnica de clusterização agrupará os dados segundo os parâmetros estabelecidos pela lógica fuzzy, ajudando numa maior análise das informações obtidas.

Este estudo é de grande importância para sociedade, pois avanços nessa área podem

adquirir benefícios em um meio comum. Como ilustração de um desses avanços pode-se citar a prevenção de desastres naturais, que através da leitura dos dados históricos de uma determinada região, analisa-se o grau de pertinência aonde possivelmente podem acontecer terremotos, secas, inundações, geadas, etc. (CAVALCANTI JÚNIOR, 2006).

A lógica Fuzzy na clusterização entra como intensificadora de respostas, visto que ela pode gerar saídas de dados possivelmente precisas, sem a necessidade de entradas específicas.

Já do ponto de vista econômico, o emprego da mineração de dados é relevante porque, em meio a inúmeras reflexões, através dos seus resultados é possível (KAESTNER, 2008):

- Detectar e gerenciar fraudes em serviços de cartões de crédito, telefonia celular, e convênios de saúde, pela construção de modelos de pessoas com comportamento fraudulento e uso de data mining para identificar instâncias similares;

- Apontar para dirigentes do comércio eletrônico, produtos que potencialmente podem ser comprados, por determinado grupo de clientes, através da análise de compras antecessoras; entre outras situações.

Inexoravelmente, este projeto visa o emprego de uma aplicação algorítmica, que tem por finalidade, analisar e utilizar o conhecimento extraído, para apoiar algum processo de tomada de decisão.

Indústrias, governos, corporações e instituições científicas utilizam as ferramentas e técnicas empregadas pelo abrangente e emergente campo da Descoberta do Conhecimento, para análise automática e inteligente, dos seus inensos repositórios de banco de dados. Mineração de Dados é a fase responsável do Knowledge Discovery in Database, para selecionar os métodos que serão utilizados para procurar e reconhecer padrões nos dados, sob diferentes perspectivas, seguidos de uma busca efetiva de interesse particular de representação, onde sequencialmente será ajustado um algoritmo, para melhor articulação da tarefa proposta (SILVA, 2004).

Segundo a bem elaborada definição difundida por Fayyad et al. (apud MELLO, SILVA, SOUZA, 2007), Knowledge Discovery in Database, em português: Descoberta de Conhecimento em Banco de Dados, representa: "... o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

2 REFERENCIAL TEÓRICO

2.1 DATA MINING

Data Mining, ou Mineração de Dados, é um dos estudos da computação já incorporado em diversas empresas e setores comerciais, sejam tanto da área tecnológica, como de várias outras áreas em âmbito mundial. Com a geração cada vez maior de informações, é essencial tirar o máximo de vantagem possível desse volume de dados, aproveitando a tecnologia para o descobrimento de conhecimentos na armazenagem de elementos (NAVEGA, 2002).

Freqüentemente, existem informações "escondidas" nos dados que não são facilmente evidentes, analistas humanos poderiam demorar semanas para descobrir informações úteis que máquinas programadas descubrem em poucas horas (TAN; STEINBACH; KUMAR, 2004).

O processo de Data Mining possui diversos algoritmos, diversos deles recentes, que processam e encontram padrões novos e válidos. Mas para que esses padrões sejam consideráveis "valiosos" ainda é essencial a presença de analistas humanos, visto que ainda não temos uma solução eficaz para determinar se um padrão é interessante ou não. Diante de tal circunstância é fato que o Data Mining ainda necessita de uma potente interação com analistas humanos, que determinam, num ultimo processo, o valor dos padrões encontrados, bem como, a condução e o direcionamento para a exploração dos dados, pontos chaves que não podem faltar num projeto com perspectiva de sucesso (NAVEGA, 2002).

Indústrias, governos, corporações e institutos científicos utilizam as ferramentas e técnicas empregadas pelo abrangente e emergente campo da Descoberta do Conhecimento, para análise automática e inteligente, dos seus imensos repositórios de banco de dados. Mineração de Dados é a fase responsável do Knowledge Discovery in Database, para selecionar os métodos que serão utilizados para procurar e reconhecer padrões nos dados, sob diferentes perspectivas, seguidos de uma busca efetiva de interesse particular de representação, onde seqüencialmente será ajustado um algoritmo, para melhor articulação da tarefa proposta (SILVA, 2004).

Segundo a bem elaborada definição difundida por Fayyad et al. (apud MELLO; SILVA; SOUZA, 2007), Knowledge Discovery in Database, em português: Descoberta de Conhecimento em Banco de Dados, representa: "... o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

O processo de KDD inclui: limpeza; integração; seleção e transformação dos dados; avaliação dos padrões; apresentação do conhecimento e o Data Mining, que por sua vez é o coração do KDD (KAESTNER, 2008).

Foram elaboradas várias técnicas de Data Mining no passado, sempre com o intuito de extrair informações dos dados. Data Mining é a combinação de diferentes técnicas de sucesso comprovado, como inteligência artificial, estatística e banco de dados. (FERRO; LEE, 2001).

2.1.1 Tarefas da Mineração de Dados

Segundo Silveira (2003) “tarefa” no contexto da mineração de dados, corresponde a um tipo de problema a ser solucionado na descoberta do conhecimento. Aonde essas tarefas podem ser efetuadas sozinhas ou podem se combinar entre si para uma resolução completa. A escolha da tarefa mais apropriada depende do tipo de aplicação a ser desenvolvida.

As tarefas da mineração de dados podem ser classificadas como preditivas (conhecidas também como previsíveis), aonde algumas variáveis podem ser usadas para prever valores desconhecidos ou futuros de outras variáveis, e como descritivas quando descrevem conceitos ou conjuntos de dados relevantes ao homem (TAN; STEINBACH; KUMAR, 2004).

Harrison (apud GONÇALVES, 2001), identificou seis tarefas de mineração de dados. São elas: classificação, estimativa, previsão, associação ou análise de cesta de supermercado, descrição e segmentação, também conhecida como clustering.

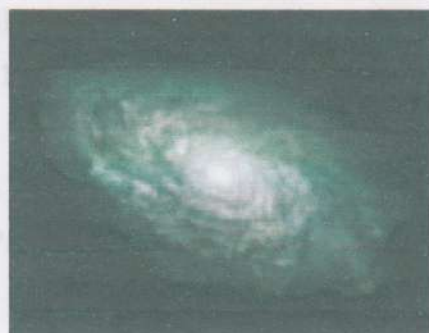
Serão descritas a seguir cada uma dessas tarefas.

2.1.1.1 Classificação

Segundo Amorim (2006) a classificação é uma das técnicas mais utilizadas, pois é uma das tarefas mais realizadas pelo homem, para a compreensão do ambiente em que se vive. O ser humano está sempre classificando tudo ao seu redor em classes, como colegas de trabalho, amigos e familiares, assim a cada classe podem ser atribuídos diferentes tipos de tratamentos pessoais.

Segundo Tan; Steinbach; Kumar (2004) a classificação é uma técnica generalizada que

engloba muitas e variadas aplicações. Exemplos incluem detectar spams de mensagens eletrônicas com base no cabeçalho e no conteúdo da mensagem, categorizar as células malignas ou benignas baseada nos resultados de imagens de ressonâncias magnéticas, e de classificação de galáxias com base em suas formas (ver figura 1).



(A) A spiral galaxy.



(B) An elliptical galaxy.

Figura 1: Classificação das galáxias.

Fonte: TAN, STEINBACH e KUMAR (2004).

Segundo Berry; Linoff (apud GONÇALVES, 2001) a classificação consiste em examinar um objeto e ligá-lo a uma classe pré-definida, aonde serão representados por registros no banco de dados, e atualizados de forma que cada campo receba um código da classe, sempre que houver o mandato da classificação. Concretizando assim, um método previsível tratado com valores discretos.

Como exemplos sucintos de algumas das tarefas realizadas pela classificação, pode-se citar: a separação de pedidos de créditos em baixo, médio e alto risco; o esclarecimento de pedidos de seguros fraudulentos; e a identificação a respeito da forma de tratamento que um paciente em potencial estaria mais apto de receber, baseando-se em classes de pacientes que já receberam algum tipo de tratamento médico ocorrido com sucesso (SILVEIRA, 2003).

2.1.1.2 Estimativa

A estimativa é uma tarefa usada para atribuir valores a variáveis contínuas desconhecidas, como o saldo de um cartão de crédito; ao contrário da classificação que trata de resultados discretos, a estimativa sempre tratará de resultados contínuos, mas pode ser usada como uma das tarefas da classificação, convencionando-se que os intervalos dos valores

contínuos correspondem a diferentes classes. Estimativa é um aprendizado que direciona um item de dado, para uma variável de predição real estimada (FAYYAD apud SILVEIRA, 2003).

São Exemplos de Estimativa, segundo Harrison, (apud GONÇALVES, 2001):

- Estimar o número de filhos em uma família;
- Estimar a renda total de uma família;
- Estimar valores em tempo de vida de um cliente;

2.1.1.4 Segmentação ou clustering

2.1.1.3 Previsão

A previsão tem os mesmos comportamentos da classificação e da estimativa, sobressaindo-se pelo fato de prever valores futuros, como quais clientes sairão nos próximos meses ou a previsão de quais assinantes de telefone usariam um serviço adicional (HARRISON apud GONÇALVES, 2001).

A única forma de saber se uma previsão ocorrerá com sucesso é esperar o acontecimento, e então se verifica o grau de acerto ou de erro da realizada previsão (AMORIM, 2006).

2.1.1.4 Associação ou análise de cesta de supermercado

O objetivo principal desta tarefa é associar registros de dados que estão ou devem estar relacionados, ou seja, definir a probabilidade de determinados itens serem obtidos em conjunto numa mesma transação (SILVEIRA, 2003).

A associação pode ser avaliada como uma tarefa bem definida, determinística e relativamente simples, visto que não envolve predição, assim como a tarefa da classificação (FREITAS apud SILVEIRA, 2003).

2.1.1.5 Descrição

Conforme Berry; Linoff (apud GONÇALVES, 2001) às vezes o objetivo da mineração de dados é apenas a descrição do que está ocorrendo em um banco de dados complexo, para aumentar o entendimento sobre um determinado grupo de instâncias (pessoas, produtos, processos, etc.) que produziram dados.

2.1.1.6 Segmentação ou clustering

A segmentação dos dados consiste em um agrupamento, onde se pretende colocar os itens de dados com características semelhantes em um mesmo grupo, denominado de cluster, enquanto que itens de dados com características diferentes sejam alocados em outros clusters. Ou seja, a finalidade da segmentação, também conhecida como clusterização ou agrupamento, é identificar grupos e a partir disto formar agrupamentos (ALMEIDA; SOUZA, 2007).

O clustering pode ser definido como uma forma de aprendizagem não supervisionada, aonde não há classes pré-definidas. A clusterização é tipicamente usada como ferramenta para análise da distribuição dos dados e como pré-processamento para outros métodos (KAESTNER, 2008).

Devido ao fato da tarefa de clusterização ser o objeto de estudo deste trabalho, ela será melhor detalhada no próximo item.

2.2 CLUSTERS

As técnicas de clusterização de dados vêm sendo cada vez mais pesquisadas e utilizadas, principalmente em grandes empresas que produzem uma alta quantidade de dados, e necessitam de informações estratégicas em prol da melhoria da mesma (VALE, 2005).

A clusterização divide os dados em grupos, chamados clusters, que sejam significativos e/ou úteis. Se a meta é atingir grupos significativos, então os grupos capturam a estrutura natural dos dados. Sobretudo em alguns casos, a clusterização é um bom ponto de partida para outros fins, como outras tarefas da mineração de dados (ex: sumarização,

classificação, estimativa, previsão). Há muito tempo a clusterização tem desempenhado um papel importante numa ampla variedade de campos, como psicologia, biologia, estatística, reconhecimento de padrões, recuperação de informação, máquinas de aprendizagem, e mineração de dados (TAN; STEINBACH; KUMAR, 2004).

A tarefa de clusterizar procura encontrar grupos de dados similares entre si. Portanto um agrupamento é um conjunto de elementos que possuem características semelhantes, que os distingam dos elementos de outros clusters, podendo revelar informações importantes a partir dessa aglomeração (VALE, 2005).

A figura 2 exibe um conjunto de dados antes da clusterização, e gradativamente a separação desse conjunto em grupos distintos, onde cada cor corresponde a um grupo. Na figura 2(a) os dados estão todos em um mesmo grupo, na figura 2(b) os dados são separados em dois clusters, na figura 2(c) em três clusters e na figura 2(d) em cinco clusters (OLIVEIRA, 2008).

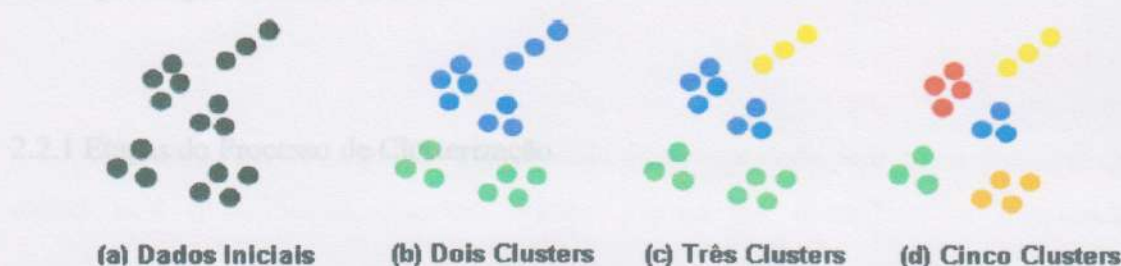


Figura 2: Clusterização de um conjunto de dados. (a) Dados originais, (b) Divisão em dois clusters, (c) Divisão em três clusters e (d) divisão em cinco clusters. Com cada cluster sendo representado com uma cor diferente. Fonte: OLIVEIRA (2008).

Assim, Viglioni (2007), disserta que o objetivo da clusterização é aumentar a similaridade “intracluster” (dos elementos internos do cluster), e diminuir a similaridade “intercluster” (entre os elementos externos pertencentes aos outros clusters). A clusterização necessita identificar automaticamente os rótulos desconhecidos, diferente da classificação, que possui os rótulos já previamente definidos, e é mediante a essa identificação automática que a clusterização é denominada de indução não-supervisionada (forma de aprendizagem, na inteligência artificial, adquirida por meio de observações e descobertas).

Segundo Berkhin (apud VALE, 2005), os itens considerados no processo de agrupamento são:

- Tipo de atributos que o algoritmo opera;
- Escalabilidade para grandes conjuntos de dados;
- Habilidade de trabalhar com um grande número de variáveis;

- Habilidade de encontrar agrupamentos de forma irregular;
- Tratar valores discrepantes;
- Tempo de execução;
- Dependência de ordem dos dados;
- Classificação;
- Segurança no conhecimento;
- Interpretabilidade dos resultados.

No entendimento de Viglioni (2007), o processo de clusterização exige que o usuário determine a quantidade de grupos a serem analisados. Pois assim, através do embasamento da quantidade de grupos, é que os registros de dados são separados de tal forma cujos registros similares fiquem nos mesmos grupos e registros diferentes fiquem em grupos distintos. Depois da formação desses grupos, pode-se fazer uma análise dos elementos que compõe cada conjunto, achando as características comuns aos seus elementos, e assim, criando um rótulo para representar cada cluster.

2.2.1 Etapas do Processo de Clusterização

Segundo Jain et al. (apud OLIVEIRA, 2008), existem muitos fatores que devem ser considerados para que a eficiência da clusterização seja assegurada, como: coleção dos dados, seleção de variáveis, medidas de similaridade, algoritmo de clusterização, validação e análise dos resultados obtidos, demonstrados na figura 3.

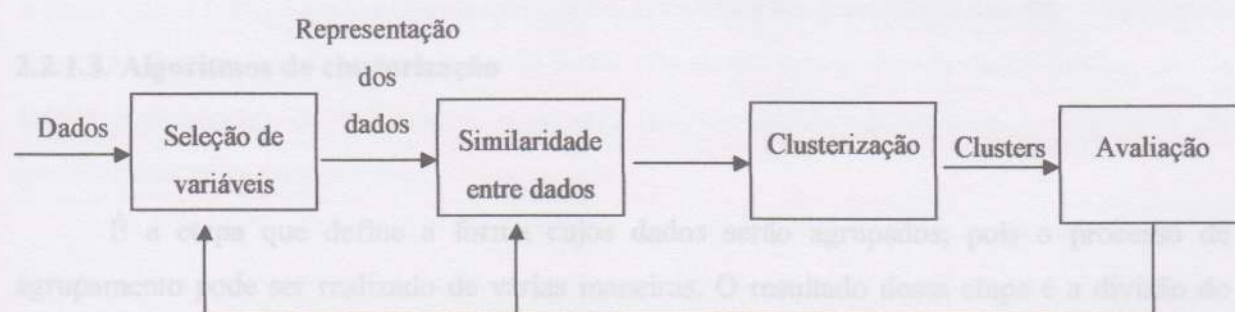


Figura 3: Etapas do processo de clusterização.

Fonte: OLIVEIRA (2008).

A seguir, serão explicitadas as etapas que formalizam o processo de clusterização:

2.2.1.1 Pré-processamento e seleção de variáveis

É a etapa onde as variáveis são identificadas, pode ocorrer à eliminação de algumas variáveis (como quando os valores da variável são iguais), e os dados coletados são formatados de uma forma que o algoritmo de clusterização consiga processá-los. Em um conjunto formado por n dados, o resultado será uma matriz $n \times d$, onde d é o número de atributos. Deste modo, um dado corresponde a um ponto no espaço d -dimensional, e a tarefa da clusterização incidi em identificar conjuntos de dados próximos a esse d -dimensional (JAIN; DUBES apud OLIVEIRA, 2008).

2.2.1.2 Medidas de similaridade

É indispensável empregar uma medida de similaridade para que a distância entre dois dados seja quantificada. Existem várias formas de quantificar a similaridade, ou dissimilaridade, dependendo da necessidade de cada clusterização. Normalmente a dissimilaridade é a forma mais empregada, e a Distância Euclidiana é a maneira mais comum de calcular essa dissimilação. Entre pares de dados, a medida de similaridade mais apropriada é fundamental para a clusterização dos dados (OLIVEIRA, 2008).

2.2.1.3. Algoritmos de clusterização

É a etapa que define a forma cujos dados serão agrupados, pois o processo de agrupamento pode ser realizado de várias maneiras. O resultado dessa etapa é a divisão do conjunto de dados inicial em clusters (OLIVEIRA, 2008).

No item 4.2.2.2.1 é apresentado o algoritmo de clusterização K-means, e no item 2.4.1.1 em especial, é apresentado o algoritmo de clusterização Fuzzy c-means, que foi colocado em prática no desenvolvimento deste trabalho.

2.2.1.4 Validação e análise dos resultados

É a etapa onde se avalia a qualidade dos clusters já encontrados, visto que eles são inicialmente desconhecidos. Dependendo da necessidade, a análise dos resultados pode vir a redefinir os atributos escolhidos e/ou a medida de similaridade. A validação pode ser feita com base em índices estatísticos ou através da comparação com outros algoritmos (OLIVEIRA, 2008).

2.2.2 Métodos de Clusterização

Conforme Vale (2005), os métodos de agrupamento podem ser definidos em duas grandes categorias: métodos hierárquicos e métodos particionais.

A seguir, esses métodos serão apresentados.

2.2.2.1 Métodos hierárquicos

Segundo Cavalcanti Júnior (2006) participam dos métodos hierárquicos: os algoritmos de clusterização que organizam um conjunto de dados, em uma estrutura hierárquica, de acordo com a proximidade entre os elementos. Os resultados geralmente são mostrados numa árvore binária ou um *dendograma*, onde a raiz representa o conjunto de dados inteiro, os nós folhas representam os indivíduos, e os nós intermediários representam a magnitude da proximidade entre os elementos.

Vale (2005) explica que dendograma é a representação gráfica que em forma de árvore, exhibe as estruturas dos agrupamentos, podendo ser visto a ordem em que os dados foram agrupados.

Os algoritmos hierárquicos são divididos em aglomerativos ou divisivos. Os algoritmos aglomerativos iniciam com cada indivíduo no conjunto de dados sendo representados por um único cluster, onde ocorre uma série de combinações entre os clusters iniciais, resultando que, depois de um determinado número de passos executados em todos os

indivíduos, os elementos similares estarão em um mesmo cluster (CAVALCANTI JÚNIOR, 2006).

Já os algoritmos divisivos atuam na direção oposta, iniciando com um conjunto de dados inteiro, que estarão em apenas um cluster, onde serão adotados vários procedimentos, que dividirão os clusters sucessivamente, até que cada indivíduo seja um único e centrado cluster (CAVALCANTI JÚNIOR, 2006).

Na figura 4, a seguir, Cavalcanti Júnior (2006), apresenta um exemplo de *dendograma* com algoritmos de clusterização hierárquicos aglomerativo e divisivo.

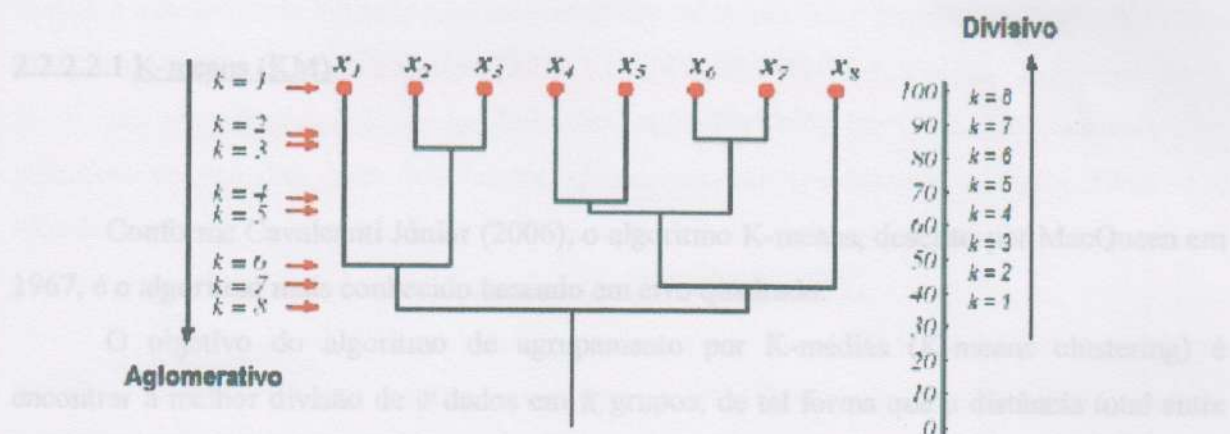


Figura 4: Exemplo de algoritmo de clusterização de dados hierárquico aglomerativo e divisivo.
Fonte: CAVALCANTI JÚNIOR (2006).

2.2.2.2 Métodos particionais

Os métodos particionais, seguem uma linha adversa dos métodos hierárquicos, pois associam um conjunto de indivíduos a K grupos sem criar uma estrutura hierárquica. Como pode ser visto na figura (5) abaixo, onde o conjunto de dados está sendo particionado em 3 grupos, e ainda na mesma figura são mostradas diferentes etapas de execução do algoritmo de clusterização. É perceptível o modo em que os elementos vão mudando de um cluster para o outro, e como a forma dos mesmos vai se alterando (CAVALCANTI JÚNIOR, 2006).

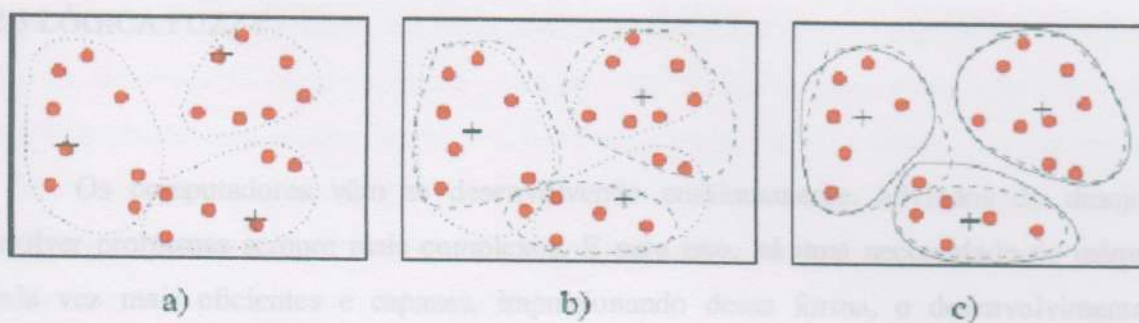


Figura 5: Clusterização de um conjunto de dados em 3 grupos baseadas no algoritmo K-means.
Fonte: CAVALCANTI JÚNIOR (2006).

2.2.2.2.1 K-means (KM)

Conforme Cavalcanti Júnior (2006), o algoritmo K-means, descrito por MacQueen em 1967, é o algoritmo mais conhecido baseado em erro quadrado.

O objetivo do algoritmo de agrupamento por K-médias (K-means clustering) é encontrar a melhor divisão de $\mathcal{P}$ dados em $\mathcal{K}$ grupos, de tal forma que a distância total entre um grupo e seu respectivo centro, somada por todos os grupos, seja a menor possível (PIMENTEL; FRANÇA; OMAR, 2003).

Como em muitos outros algoritmos, a solução que o K-means proporciona, geralmente depende do ponto de partida, mas na maioria das vezes ele consegue encontrar um mínimo local. O K-means mesmo sendo sensível a ruídos e outliers, pode ser considerado um método prático e computacionalmente eficiente. Mas, no entanto, não é aplicável para agrupamentos não-convexos (VALE, 2005).

Segundo Monteiro e Silva et al (apud PIMENTEL; FRANÇA; OMAR, 2003), O método K-means usa os valores dos primeiros $\mathcal{N}$ casos, em um arquivo de dados, presumindo uma média temporária dos $\mathcal{K}$ cluster, onde $\mathcal{K}$ é o número de clusters mencionado pelo usuário. Desta forma, o centro do cluster inicial é formado para cada caso em volta dos dados mais próximos, depois são comparados com os pontos mais distantes e com os outros clusters que foram formados. A partir daí, dentro de um processo atualizado continuamente, e de um processo iterativo, são descobertos os centros dos últimos clusters.

2.3 LÓGICA FUZZY

Os computadores vêm se desenvolvendo continuamente, advindos do desejo de resolver problemas sempre mais complexos. E para isso, há uma necessidade de máquinas cada vez mais eficientes e capazes, impulsionando dessa forma, o desenvolvimento da Inteligência Artificial (LUGER; STUBBLEFIELD; apud ORTEGA, 2001).

Já se observam máquinas que são capazes de imitar parte do raciocínio humano, devido à indução da habilidade para lidar com situações incertas e imprecisas, permitida pelos avanços da Teoria da Lógica Fuzzy (KLIR; YUAN apud ORTEGA, 2001).

Foi um polonês chamado Jan Lukasiewicz (1878-1956), em 1920, que desenvolveu as primeiras noções dos conceitos "vagos" posteriormente combinados a lógica fuzzy. Ele introduziu conjuntos com graus de pertinência 0, $1/2$, e 1, que mais tarde foram expandidos para um número infinito de valores entre 0 e 1 (VIGLIONI, 2007).

Ná década de 60, Lotfi A. Zadeh, professor da Universidade de Berkeley, USA, ao trabalhar com inteligência artificial, observou que a forma abrupta da teoria usual de conjuntos, não era adaptável com a implementação de sistemas especializados, a vista que estes são softwares interativos que podem executar algumas tomadas de decisão, necessitando de mais do que rígidas tomadas binárias, a luz que estas, não contemplam toda a possibilidade de escolhas, desenvolvidas cotidianamente pela mente humana (BILOBROVEC, 2005).

Em 1965, o professor Lotfi A. Zadeh introduziu a lógica fuzzy no contexto científico, através da publicação do artigo *Fuzzy Sets* no *Journal Information and Contro* (JANÉ, 2004).

Também conhecida como lógica difusa, ou lógica nebulosa, a lógica fuzzy tem como objetivo, modelar aproximadamente o modo de raciocínio humano, visando o progresso de sistemas computacionais capazes de tomar decisões racionais em um ambiente imbuído de incerteza e imprecisão. Assim, a lógica fuzzy permite a manipulação de informações imprecisas, como os conceitos de muito, pequeno, alto, bom, quente, frio, etc. proporcionando respostas mais aproximadas o possível para questões com conhecimento pouco exato, incompleto ou não totalmente confiável (ARAÚJO, 2006).

Cox definiu a lógica fuzzy, em 1995, como um modo capaz de combinar a imprecisão dos eventos naturais com o poder computacional das máquinas, para produzir sistemas de respostas inteligentes, robustos e flexíveis (JANÉ, 2004).

Segundo Sandri; Correa (1999), a imprecisão e a incerteza são os principais aspectos da imperfeição da informação. Quanto mais se aumenta a incerteza mais se diminui a

imprecisão e vice-versa. Como exemplo prático, supõe-se que há uma seguridade em informar que um determinado filme começará entre 3h e 4h, se houver uma obrigatoriedade em dar uma informação mais precisa, tende-se a aumentar a incerteza, informando, por exemplo, que o filme começará às 3h30min com uma probabilidade diferente de 1. Informações imprecisas podem ao mesmo tempo ser vagas, como quando na continuidade do exemplo, informa-se que o filme irá começar “por volta” das 3h30min.

Jané (2004), elucida que “Os algoritmos baseados na lógica *fuzzy* apresentam um conceito distinto dos métodos determinísticos e probabilísticos, considerando na verdade a natureza possibilística das variáveis envolvidas”.

Conforme Von Altrock (1996 apud JANÉ 2004), as aplicações da lógica *fuzzy* permeiam diversos campos, tais como: avaliação de crédito, controle de fluxo de caixa, análise de risco, controle de estoques, avaliação de marketing, avaliação de fornecedores, controle de qualidade, e otimização de inventários.

Várias aplicações foram alcançadas recentemente com *fuzzy*, são exemplos reais disso: controle de elevadores, tráfego automotivo, controle automático de foco em câmeras fotográficas digitais e sistemas de acionamento robotizado (YEN; LANGARI; ZADEH, apud JANÉ 2004).

Atualmente, várias empresas famosas possuem laboratórios de pesquisa em lógica *fuzzy* para o desenvolvimento de seus produtos. São exemplos dessas empresas: Siemens, Daimler-Benz, Klockner-Moeller, SGS-Thomson, General Motors, Motorola, Hewlett-Packard, entre várias outras (ORTEGA, 2001).

2.3.1 Conjuntos Fuzzy

Conforme Simões (2003) Os métodos de classificação usuais (*Hard Cluster Analysis*) partem do princípio dos conjuntos clássicos (*Crisp Sets*), que tem como característica a inequivocidade de sua função de pertinência. Fundamentalmente, a teoria dos conjuntos apresenta uma noção rígida fundamental de “pertencer” ou “não pertencer” a um determinado conjunto. Logo, definir um conjunto clássico significa tomar decisões binárias, relacionando a pertinência de um determinado indivíduo com uma dada classe, que irá aceitá-lo (“=1”) ou rejeitá-lo (“=0”).

Segundo Ortega (2001), os conjuntos fuzzy não possuem fronteiras bem definidas. Eles foram criados justamente pra suprir a necessidade dos conjuntos clássicos, onde as transições de uma classe para outra aconteciam de forma abrupta. A definição, propriedades e operações são advindas da generalização da teoria dos conjuntos clássicos.

Os conjuntos fuzzy fornecem o alicerce para a representação fuzzy, a teoria fuzzy define uma função de pertinência ($\mathcal{M}$), que indica quão o elemento u (pertencente ao universo $\mathcal{U}$) pertence a um conjunto χ , numa escala entre 0 e 1: $\mathcal{M}_u(\chi) : \mathcal{U} \rightarrow [0,1]$. Junto com essa função, definem-se todos os operadores atuantes no conjunto, e todo o alicerce lógico necessário para a o desenvolvimento da sua utilização prática (L. A. ZADEH, apud ARAÚJO 2006).

Segundo Sandri, Correa (1999) A função de pertinência $\mathcal{M}_u(\chi)$ indica o grau de compatibilidade entre o conceito expresso por u , onde:

- $\mathcal{M}_u(\chi) = 1$ indica que χ é completamente compatível com u ;
- $\mathcal{M}_u(\chi) = 0$ indica que χ é completamente incompatível com u ;
- $0 < \mathcal{M}_u(\chi) < 1$ indica que χ é parcialmente compatível com u , com grau $\mathcal{M}_u(\chi)$;

Para maior esclarecimento da teoria dos conjuntos difusos, Araújo (2006) exemplificou a definição de “Jovem”. Para um determinado grupo de pessoas, alguém com 25 anos é jovem, enquanto que para outras pessoas 35 anos “ainda” é jovem. Mesmo sendo complicado estabelecer um limite para o conceito de jovem, pode-se chegar a alguns consensos através do fato de que alguém de 1 ano de idade é definitivamente (100%) jovem e alguém de 100 anos não é nada (0%) jovem. Sendo assim, uma pessoa com 35 anos pode ter um grau de adesão ao conceito jovem entre 0 e 100%. Como os conjuntos fuzzy estabelecem funções de pertinência para que se possa pertencer parcialmente a um ou mais conjuntos, o termo jovem pode ser descrito pelo conjunto fuzzy, assim como mostra o quadro 1 abaixo.

Idade	Grau de Pertinência
25	1
30	0,8
35	0,6
40	0,4
45	0,2
50	0

Quadro 1: Função de pertinência para os conjuntos do exemplo para “jovem”.
Fonte: ARAÚJO (2006).

Por intermédio de uma função fuzzy, pode-se definir conceitos como “pouco jovem” ou “muito jovem”. Isso se deve através da aplicação das operações matemáticas, onde “pouco” e “muito” são definidos por funções matemáticas no conjunto fuzzy “jovem” (ARAÚJO 2006).

2.3.1.1 Operações dos conjuntos fuzzy

Segundo Miranda; Vilela; Kronbauer (2003), é bastante conhecido o emprego dos operadores na lógica booleana, vê-se no exemplo abaixo a aplicação de uma operação “And”.

$$1 \text{ And } 1 \rightarrow 1$$

$$1 \text{ And } 0 \rightarrow 0$$

$$0 \text{ And } 1 \rightarrow 0$$

$$0 \text{ And } 0 \rightarrow 0$$

Na lógica difusa os valores não estão definidos. Quando se aplicam regras da lógica binária a duas variáveis difusas, tem que se considerar uma lógica de valores máximos e mínimos, através do emprego das operações definidas na lógica fuzzy. Pois como a lógica difusa utiliza conjuntos difusos, também haverá operações como união, interseção e complemento, que serão apresentadas separadamente a seguir (MIRANDA, VILELA e KRONBAUER, 2003).

2.3.1.1.1 União

Miranda; Vilela; Kronbauer (2003) elucidam que “Sejam A e B dois conjuntos difusos no universo U com funções de pertinência $\mu_A(x)$ e $\mu_B(x)$ respectivamente, a união dos conjuntos A e B se define por”:

$$A \cup B = \{(\max\{\mu_A(x), \mu_B(x)\}/x \mid x \in U\}$$

A função de pertinência de x ao conjunto união de A e B , onde a função $\max\{\}$ seleciona o valor máximo dos graus de pertinência é (MIRANDA; VILELA; KRONBAUER, 2003):

$$\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}$$

A figura 6 exibe a união entre dois conjuntos difusos:

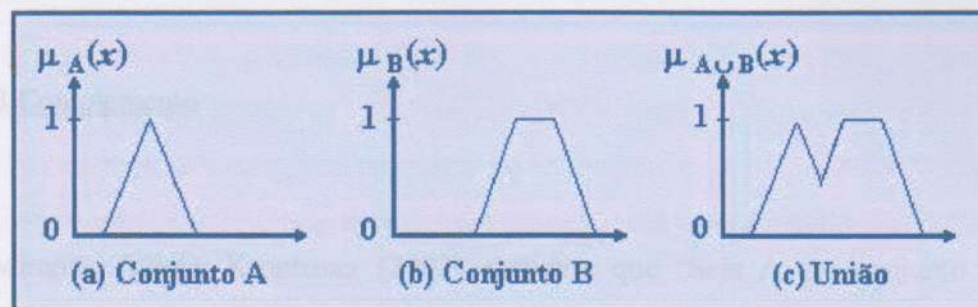


Figura 6: Função de pertinência e operação de união dos conjuntos difusos A e B
Fonte: Miranda; Vilela; Kronbauer (2003).

2.3.1.1.2 Interseção

Miranda; Vilela; Kronbauer (2003) elucidam que “Sejam A e B dois conjuntos difusos no universo U com funções de pertinência $\mu_A(x)$ e $\mu_B(x)$ respectivamente, a interseção dos conjuntos A e B se define por”:

$$A \cap B = \{(\min\{\mu_A(x), \mu_B(x)\})/x \mid x \in U\}$$

A função de pertinência de x ao conjunto interseção de A e B , onde a função $\min\{\}$ seleciona o valor mínimo dos graus de pertinência é (MIRANDA, VILELA e KRONBAUER, 2003):

$$\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}$$

A figura 7 exibe a interseção entre dois conjuntos difusos:

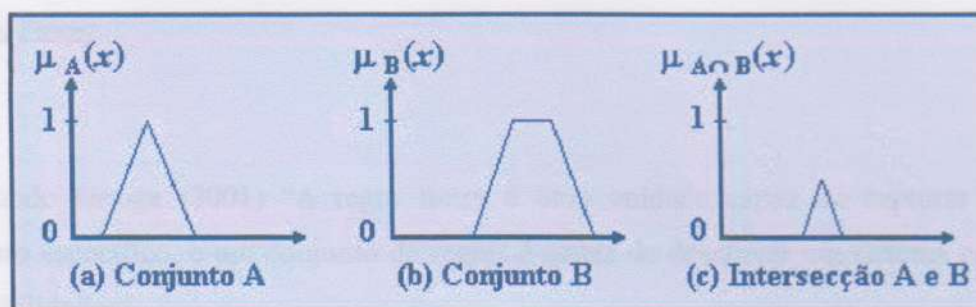


Figura 7: Função de pertinência e operação de intersecção dos conjuntos difusos A e B
Fonte: Miranda; Vilela; Kronbauer (2003).

2.3.1.1.3 Complemento

Miranda; Vilela; Kronbauer (2003) elucidam que “Seja A um conjunto difuso no universo U com uma função de pertinência $\mu_A(x)$, o complemento do conjunto A é o conjunto $\bar{A}$ definido por”:

$$\bar{A} = \{(1 - \mu_A(x)/x) \mid x \in U\}$$

A função de pertinência de x ao conjunto intersecção de $\bar{A}$ é (MIRANDA; VILELA; KRONBAUER, 2003):

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)$$

A figura 8 exhibe os gráficos da função de pertinência de dois conjuntos difusos e seus conjuntos complementares:

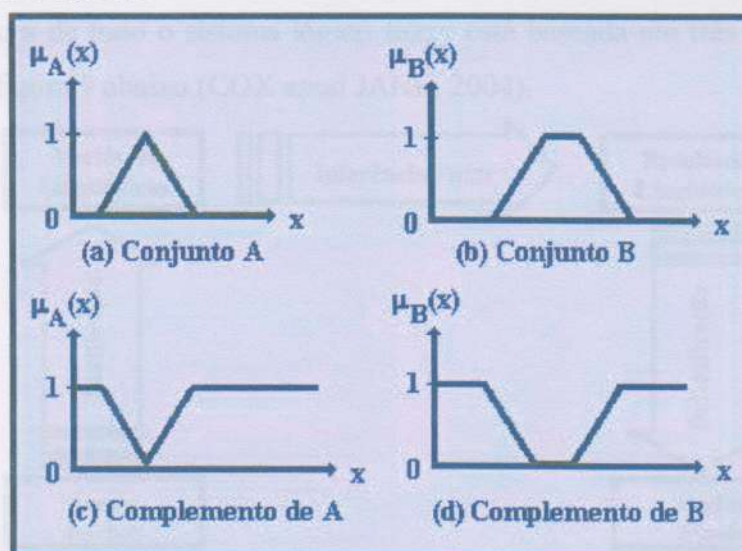


Figura 8: Função de pertinência e complemento dos conjuntos difusos A e B.
Fonte: Miranda; Vilela; Kronbauer (2003).

2.3.2 Regras Fuzzy

Segundo Ortega (2001) “A regra fuzzy é uma unidade capaz de capturar algum conhecimento específico, e um conjunto de regras é capaz de descrever um sistema em suas várias possibilidades”.

As regras fuzzy são estruturas que podem ser entendidas de diversas maneiras, e são bastante utilizadas nas abordagens da teoria fuzzy. Conceitualmente, as regras fuzzy relatam situações específicas, que podem ser enviadas para análise de especialistas. Geralmente sua inferência nos leva em algum resultado desejado. Cada regra fuzzy, assim como as regras dos conjuntos clássicos, é composta por uma parte antecedente (a parte SE) e uma parte conseqüente (a parte ENTÃO), que resulta em uma estrutura do tipo (ORTEGA, 2001):

SE {antecedentes} ENTÃO {conseqüentes}.

Neste entendimento, os antecedentes relatam condições, enquanto as parte conseqüentes relatam uma conclusão ou uma ação que pode ser vista quando as premissas forem verificadas (ORTEGA, 2001).

2.3.3 A Estrutura do Sistema Fuzzy

A estrutura de todo o sistema lógico fuzzy está baseada em três operações que estão explicitadas na figura 9 abaixo (COX apud JANÉ, 2004).

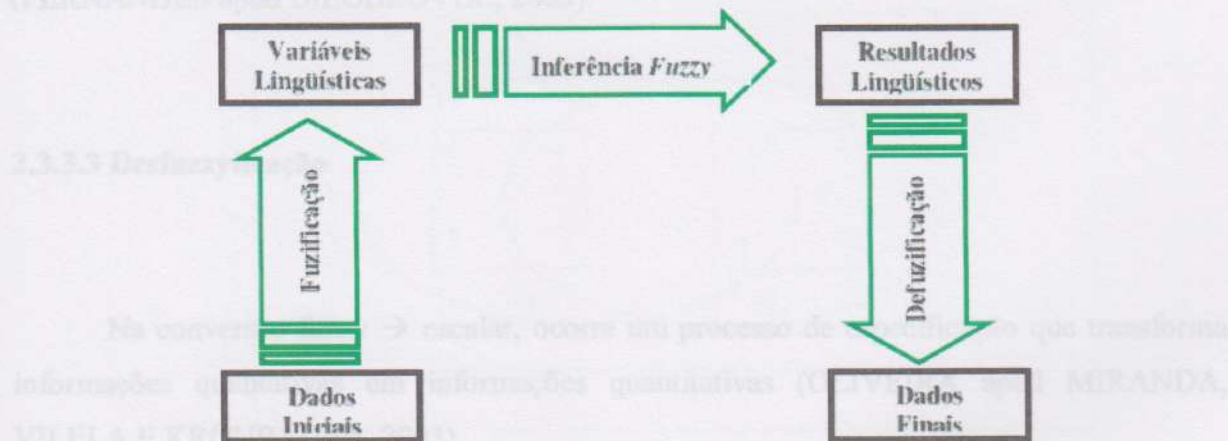


Figura 9: Sistema Lógico Fuzzy.
Fonte: COX apud JANÉ (2004).

2.3.3.1 Fuzzyficação

A fuzzyficação pode ser definida como o processo de associar ou calcular um valor que represente um grau de pertinência de entrada em um ou mais conjuntos fuzzy, onde o grau de pertinência é induzido por uma função de pertinência baseada na experiência ou intuição. É através das funções de pertinência que o controlador sintoniza desejando alcançar respostas desejadas a entradas estabelecidas (CABRAL apud BILOBROVEC, 2005).

Ou seja, é nessa etapa que os valores numéricos são transformados em graus de pertinência que se associam a uma variável lingüística para comportar uma ligação entre os termos lingüísticos como frio, próximo, quente, etc. (MATTOS apud BILOBROVEC, 2005).

2.3.3.2 Inferência

Após ocorrer na fuzzyficação a interpretação das variáveis lingüísticas, a inferência encarrega-se de descrever as situações onde houve reações, sendo assim, ocorre a determinação das regras SE-ENTÃO. A parte SE, chamada de antecedente contém uma ou mais condições, e a parte ENTÃO, chamada de conseqüente, contém uma ou mais ações (BILOBROVEC, 2005).

Cada antecedente possui um grau de pertinência antes indicado pela fuzzyficação. A inferência responsabiliza-se pela avaliação das regras, onde a intensidade é calculada de acordo com os valores dos antecedentes e estão direcionadas para as saídas difusas da regra. (FERNANDES apud BILOBROVEC, 2005).

2.3.3.3 Desfuzzyficação

Na conversão fuzzy $\rightarrow$ escalar, ocorre um processo de especificação que transforma informações qualitativas em informações quantitativas (OLIVEIRA apud MIRANDA, VILELA E KRONBAUER, 2003).

Segundo Sandri e Correa (1999) o procedimento de fuzzyficação corresponde a

identificação do domínio das variáveis de saída, em um correspondente universo de discurso, onde a ação de controle nebulosa inferida, progredi em uma ação de controle não nebulosa. Os métodos de desfuzzyficação mais conhecidos são: Primeiro Máximo (SOM), Método das Médias dos Máximos (MOM) e o Método do Centro da Área (COA).

2.4 COMPLEXO FUZZY CLUSTERING

As técnicas de clusterização tradicionais geram divisões, onde cada indivíduo pertence a um e somente um cluster, concretizando uma abordagem chamada de *clusterização hard*. Nesse processo, portanto, os clusters são disjuntos. Fuzzy clustering é um meio de estender essa abordagem para permitir a associação de um indivíduo com todos os clusters, usando uma função de pertinência (ZADEH apud CAVALCANTI JÚNIOR, 2006).

Cavalcanti Júnior (2006) explica que o conceito de conjuntos fuzzy, oferece a vantagem de expressar a similaridade de um indivíduo que apresenta características de vários grupos, pois, é possível que um algoritmo associe cada indivíduo parcialmente a todos os grupos. Como exemplo, pode-se citar um anfíbio, um indivíduo que possui características de animais aquáticos e terrestres, e tem a necessidade, portanto, de pertencer a mais de um único grupo. Um algoritmo de clusterização fuzzy propõe que cada cluster é um conjunto fuzzy de todos os indivíduos. E que uma partição hard pode ser obtida através de uma partição fuzzy, devido à aplicação de um limite nos valores de pertinência. Na figura 10, encontra-se a ilustração dessa idéia.

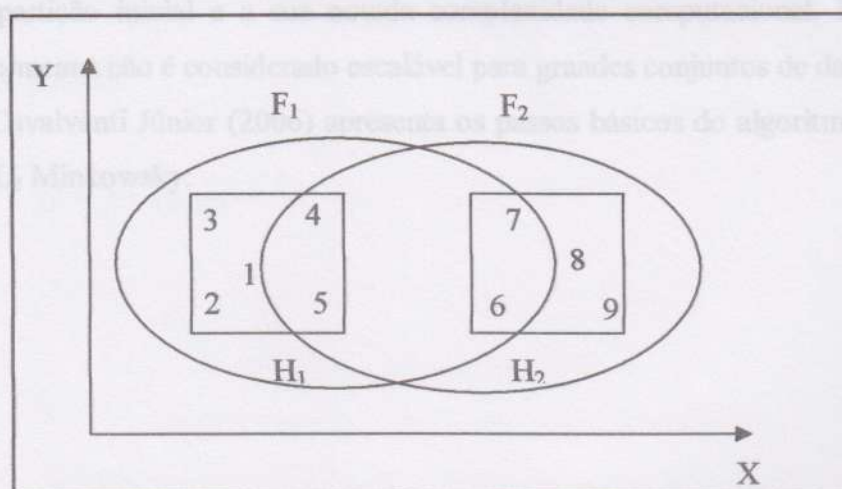


Figura 10: Exemplo de clusterização fuzzy x hard.
Fonte: CAVALCANTI JÚNIOR (2006).

Na figura 10, Cavalcanti Júnior (2006) expõe os retângulos particionando o conjunto de dados em dois hard clusters: $H_1 = \{1,2,3,4,5\}$ e $H_2 = \{6,7,8,9\}$. Um algoritmo de clusterização fuzzy produz os dois clusters fuzzy F_1 e F_2 apresentados nas elipses. Os indivíduos contêm um grau de pertinência definido no intervalo entre 0 e 1 para cada cluster. Os clusters fuzzy F_1 e F_2 podem ser descritos resumidamente como:

$$F_1 = \{(1,0.9), (2,0.8), (3,0.7), (4,0.6), (5,0.55), (6,0.2), (7,0.2), (8,0.0), (9,0.0)\}$$

$$F_2 = \{(1,0.0), (2,0.0), (3,0.0), (4,0.1), (5,0.15), (6,0.4), (7,0.35), (8,1.0), (9,0.9)\}$$

2.4.1 Fuzzy C-Means (FCM)

O fuzzy c-means é o algoritmo de clusterização fuzzy popularmente mais conhecido. Ele trabalha melhor que o algoritmo hard K-Means no propósito de evitar um mínimo local, mas o FCM ainda pode convergir para um mínimo local do critério fundamentado na soma de erros ao quadrado. O principal problema da clusterização fuzzy é o desenvolvimento das funções de pertinência. Escolhas diferentes se baseiam em decomposição de similariedade e centróides de clusters. Em 1981, Bezdek propôs uma generalização por meio de uma família de funções objetivas, chamadas critério, cujo objetivo é fornecer uma partição de conjunto de indivíduos em “C” clusters (XU et al. apud CAVALCANTI JÚNIOR, 2006).

Segundo Cavalcanti Júnior (2006), assim como o K-means, o Fuzzy c-means apresenta problemas para trabalhar com ruídos e anomalias nos dados, além da problemática na definição da partição inicial e a sua notada complexidade computacional. Portanto, o algoritmo Fuzzy c-means não é considerado escalável para grandes conjuntos de dados.

A seguir Cavalcanti Júnior (2006) apresenta os passos básicos do algoritmo Fuzzy c-means, distância L_2 Minkowsky:

1: **Recebe valores de parâmetros**

Determine: c , $2 \leq c < n$; m , $1 < m < \infty$; limite de máximo de iterações; e um escalar ε , $0 < \varepsilon$.

2: **Define matriz de pertinência inicial**

Dado um método de inicialização gere uma matriz de pertinência tal que a pertinência u_{ik} ($k = 1, \dots, n$ e $i = 1, \dots, c$) do indivíduo k ao cluster C_i tal que $u_{ik} > 0$ e $\sum_{i=1}^c u_{ik} = 1$.

3: **Para $t = 1$ até T faça**

4: **Passo de representação:**

{ A pertinência u_{ik} do indivíduo k ao cluster C_i está fixa }

5: **Calcule os protótipos g_i do cluster C_i ($i = 1, \dots, c$) usando a equação**

$$g_{it} = \frac{\sum_{k=1}^n (u_{ik})^m x_{kt}}{\sum_{k=1}^n (u_{ik})^m}$$

6: **Passo de alocação:**

{ Os protótipos g_i dos clusters C_i ($i = 1, \dots, c$) estão fixos }

Atualize o grau de pertinência u_{ik} do indivíduo k ao cluster C_i para $i = 1, \dots, c$ usando a equação

$$u_{ik} = \left[\sum_{h=1}^c \left\{ \frac{\sum_{l=1}^p (x_{kl} - g_{il})^2}{\sum_{l=1}^p (x_{kl} - g_{hl})^2} \right\}^{\frac{1}{m-1}} \right]^{-1} \quad \text{Para } i = 1, \dots, c; k = 1, \dots, n$$

7: **Critério de parada**

8: **Se $|\mathcal{J}^{(t+1)} - \mathcal{J}^{(t)}| \leq \varepsilon$ então**

9: **Pare**

10: **Fim se**

11: **Fim para**

3 METODOLOGIA

O presente trabalho empregou uma pesquisa aplicada, no intuito de gerar conhecimento em relação à clusterização de dados com lógica fuzzy e, a luz desse entendimento, atuou-se numa aplicação prática do tema abrangido.

Constituindo-se de uma abordagem qualitativa, fez-se uma análise da atuante questão, procurando entender os fenômenos associados à mesma com a utilização do método indutivo.

Objetivando um tipo de pesquisa com caráter exploratório, buscou-se identificar por meio de investigações, as qualidades propostas por determinadas formas de agrupamento de dados, que resultassem em conhecimento valioso para futuras tomadas de decisão.

Os principais procedimentos técnicos do presente trabalho envolveram:

- Pesquisas bibliográficas advindas de teses, dissertações, monografias e artigos que permeiam o campo dos procedimentos dos agrupamentos de dados;
- Um estudo de caso observando a capacidade do algoritmo Fuzzy c-means em gerar boas partições em um conjunto de dados;
- Um levantamento para adquirir um conjunto de dados significativos;
- E ainda, uma pesquisa experimental, aplicando o algoritmo Fuzzy c-means, no software MATLAB, para comprovação do levantamento bibliográfico em relação ao processo de clusterizar um conjunto de dados com a lógica fuzzy.

4.1.1 Clusterização com 5 Clusters dos Salários do Mês de Setembro

Inicialmente, houve a importação dos dados para o MATLAB, onde eles foram clusterizados com o algoritmo Fuzzy C-means em 5 clusters, centrando os focos dos agrupamentos dos dados, em distintas longitudes e latitudes.

O quadro abaixo apresenta os dados de entradas referentes aos salários dos funcionários do mês de setembro, as entradas da coluna "A" são referentes à identificação do Funcionário, e as entradas da coluna "B" são referentes ao Salário Líquido dos Funcionários da empresa:

4 ANÁLISE DE RESULTADOS

O algoritmo de clusterização Fuzzy C-means foi utilizado sobre duas amostras de dados bidimensionais com distribuições distintas e prévio conhecimento do número ideal de grupos pertinentes às amostras. Os conjuntos de dados coletados representam os salários líquidos de 119 funcionários de uma empresa, nos meses de setembro e outubro do ano de 2008.

Para a obtenção dos resultados a aplicação foi feita em duas grandes etapas.

Na primeira etapa, os dois conjuntos representativos dos dados (setembro e outubro) são utilizados como entrada e a seguir são clusterizados para alcance de 5 (cinco) clusters, e a partir disso são realizados cálculos com a fórmula da distância euclidiana para adquirir maiores informações sobre os dados e os centróides.

Na segunda etapa, os dois conjuntos representativos dos dados (setembro e outubro) são utilizados como entrada e a seguir são clusterizados para alcance de 3 (três) clusters. Logo após, são adotados os mesmos procedimentos aplicados na primeira etapa.

A seguir, serão apresentadas as duas etapas especificamente.

4.1 CLUSTERIZAÇÃO COM 5 CLUSTERS

4.1.1 Clusterização com 5 Clusters dos Salários do Mês de Setembro

Inicialmente, houve a importação dos dados para o MATLAB, onde eles foram clusterizados com o algoritmo Fuzzy C-means em 5 clusters, centrando os focos dos agrupamentos dos dados, em distintas longitudes e latitudes

O quadro abaixo apresenta os dados de entradas referentes aos salários dos funcionários do mês de setembro, as entradas da coluna "A" são referentes à identificação do Funcionário, e as entradas da coluna "B" são referentes ao Salário Líquido dos Funcionários da empresa:

Entradas Setembro		39	1703	80	261
A	B	40	4460	81	381
1	935	41	851	82	246
2	1554	42	827	83	387
3	976	43	1757	84	810
4	741	44	867	85	466
5	1552	45	1205	86	2516
6	254	46	1075	87	246
7	411	47	1011	88	259
8	620	48	1035	89	779
9	210	49	776	90	1001
10	671	50	725	91	250
11	3969	51	1371	92	1115
12	411	52	300	93	1974
13	2547	53	1289	94	386
14	1062	54	739	95	247
15	1051	55	487	96	700
16	516	56	698	97	552
17	831	57	990	98	1333
18	3606	58	648	99	701
19	1179	59	608	100	651
20	3787	60	321	101	498
21	2334	61	651	102	700
22	2855	62	656	103	356
23	688	63	348	104	720
24	2312	64	161	105	246
25	1780	65	321	106	1047
26	1525	66	587	107	853
27	836	67	635	108	1614
28	1159	68	798	109	1498
29	2578	69	860	110	350
30	534	70	2058	111	411
31	620	71	691	112	356
32	1733	72	350	113	280
33	580	73	699	114	240
34	937	74	606	115	110
35	380	75	702	116	1001
36	915	76	83	117	356
37	704	77	287	118	552
38	2680	78	708	119	381
		79	275		

Quadro 2: Dados de entrada do mês de setembro.
Fonte: Dados da pesquisa (2008).

A seguir serão exibidos o quadro e a figura da representação dos clusters encontrados nesses agrupamentos.

No quadro 3, para facilitar a identificação dos clusters, a coluna “GRUPO” nomeia os clusters de 1 a 5, a coluna “C” representa a latitude das coordenadas dos pontos de amostra, e a coluna “D” a longitude das coordenadas dos pontos de amostra.

Na figura 11, as estrelas azuis representam os dados e os quadrados verdes representam os centróides.

5 Clusters Setembro		
GRUPO	C	D
1	21,3	3945,6
2	37,7	2510,1
3	49,1	1544,9
4	57,1	792,6
5	77,2	330,1

Quadro 3: Localização dos 5 centróides dos dados do mês de setembro.
Fonte: Dados da pesquisa (2008).

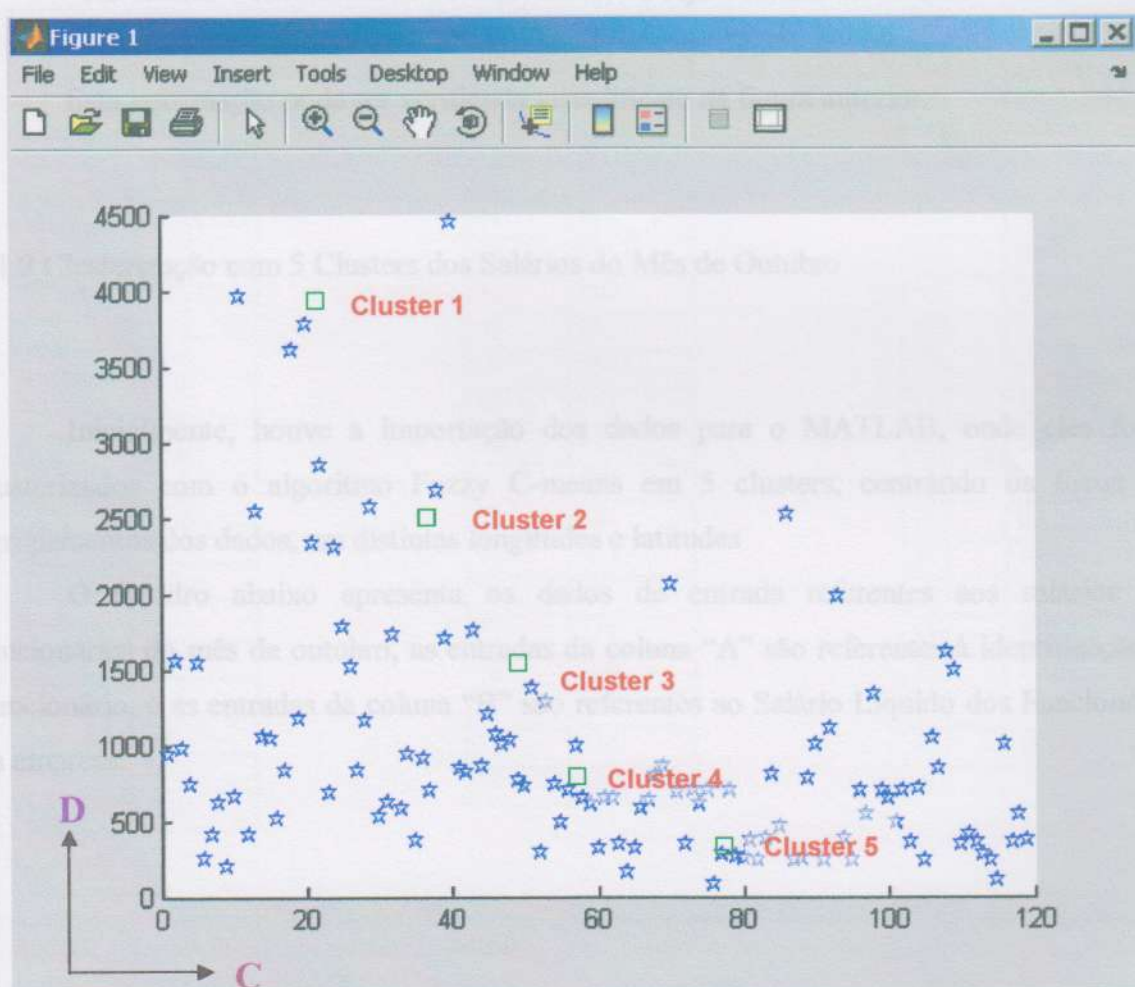


Figura 11: Clusters Setembro (5 Centróides dos dados).
Fonte: Dados da pesquisa (2008).

É interessante notar que o número de iterações realizadas para se concluir a clusterização foi de 65 iterações.

Após a clusterização, com o objetivo de medir o grau de similaridade dos indivíduos para cada centróide, calculou-se a distância euclidiana de cada elemento padrão ao centro de cada cluster, para identificar a qual classe cada indivíduo pertence, de acordo com a seguinte equação:

$$D(i, j) = \left(\sum_{t=1}^p |x_{it} - x_{jt}|^{1/2} \right)^2$$

Com o cálculo da distância euclidiana verificou-se que o cluster 4, é a classe que contém maior número de elementos, apesar de apresentar proximidade com o cluster 5, pois:

- No cluster 1 encontram-se 4 indivíduos, ou seja 3% dos elementos
- No cluster 2 encontram-se 8 indivíduos, ou seja 7% dos elementos
- No cluster 3 encontram-se 15 indivíduos, ou seja 12,5% dos elementos
- No cluster 4 encontram-se 52 indivíduos, ou seja 44% dos elementos
- No cluster 5 encontram-se 40 indivíduos, ou seja 33,5% dos elementos

Esta informação pode ser verificada visualmente na figura anterior.

4.1.2 Clusterização com 5 Clusters dos Salários do Mês de Outubro

Inicialmente, houve a importação dos dados para o MATLAB, onde eles foram clusterizados com o algoritmo Fuzzy C-means em 5 clusters, centrando os focos dos agrupamentos dos dados, em distintas longitudes e latitudes

O Quadro abaixo apresenta os dados de entrada referentes aos salários dos funcionários do mês de outubro, as entradas da coluna "A" são referentes à identificação do Funcionário, e as entradas da coluna "B" são referentes ao Salário Líquido dos Funcionários da empresa:

Entradas Outubro					
A	B				
1	935	39	4460	80	246
2	1528	40	1198	81	367
3	976	41	904	82	430
4	1542	42	3467	83	466
5	1552	43	867	84	2411
6	232	44	2116	85	328
7	411	45	1083	86	246
8	1240	46	1051	87	779
9	210	47	981	88	1661
10	5612	48	776	89	250
11	411	49	737	90	1115
12	2547	50	172	91	1974
13	1062	51	460	92	386
14	1051	52	1270	93	247
15	516	53	868	94	700
16	831	54	516	95	552
17	3606	55	783	96	1333
18	1179	56	1013	97	1029
19	3787	57	652	98	1311
20	2332	58	608	99	735
21	2855	59	293	100	700
22	658	60	311	101	356
23	2259	61	656	102	720
24	1780	62	348	103	271
25	1525	63	336	104	1047
26	836	64	587	105	853
27	1109	65	635	106	914
28	2578	66	798	107	1498
29	1010	67	875	108	350
30	1199	68	2001	109	411
31	800	69	691	110	356
32	580	70	432	111	280
33	937	71	699	112	240
34	364	72	606	113	110
35	915	73	702	114	1001
36	736	74	83	115	356
37	2681	75	273	116	552
38	1703	76	708	117	356
		77	275	118	1104
		78	314	119	381
		79	356		

Quadro 4: Dados de entrada do mês de outubro.
Fonte: Dados da pesquisa (2008).

A seguir serão exibidos o quadro e a figura da representação dos clusters encontrados nesses agrupamentos.

No quadro 5, para facilitar a identificação dos clusters, a coluna “GRUPO” nomeia os clusters de 1 a 5, a coluna “C” representa a latitude das coordenadas dos pontos de amostra, e a coluna “D” a longitude das coordenadas dos pontos de amostra.

Na figura 12, as estrelas azuis representam os dados e os quadrados verdes representam os centróides.

5 Clusters Outubro		
GRUPO	C	D
1	26,3	4400,7
2	34,8	2535,9
4	45,1	1533,9
4	56,6	874
5	76	347,8

Quadro 5: Localização dos 5 centróides dos dados do mês de outubro.

Fonte: Dados da pesquisa (2008).

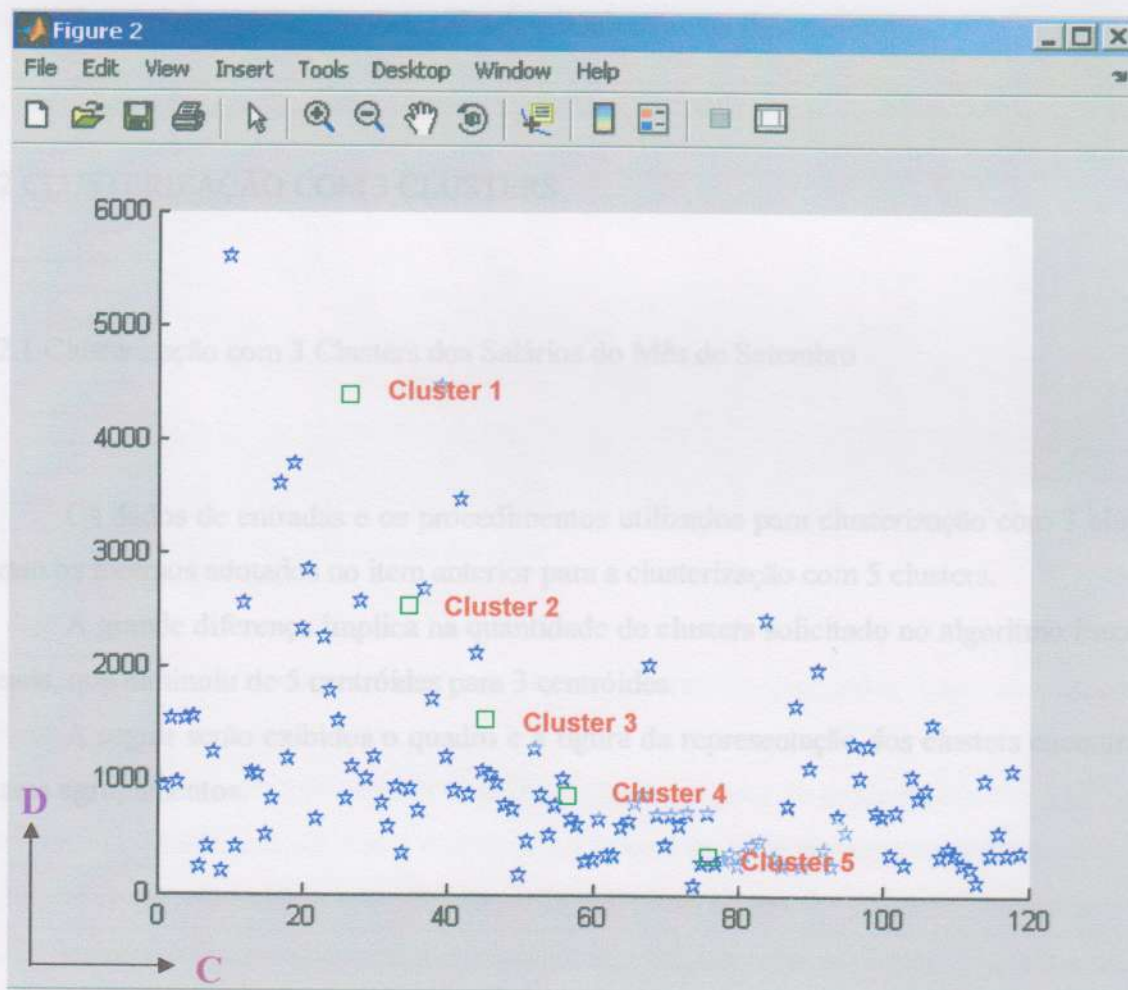


Figura 12: Clusters Outubro (5 Centróides dos dados).

Fonte: Dados da pesquisa (2008).

É interessante notar que o número de iterações realizadas para se concluir a clusterização foi de 100 iterações.

Após a clusterização, com o objetivo de medir o grau de similaridade dos indivíduos para cada centróide, calculou-se a distância euclidiana de cada elemento padrão ao centro de cada cluster, para identificar a qual classe cada indivíduo pertence, de acordo com a devida equação vista anteriormente.

Com o cálculo da distância euclidiana verificou-se que o cluster 4, continua sendo a classe que contém maior número de elementos, e que esta classe, por sua vez, aumentou ainda mais a proximidade com o cluster 5, pois:

- No cluster 1 encontram-se 4 indivíduos, ou seja 3% dos elementos
- No cluster 2 encontram-se 9 indivíduos, ou seja 7,5% dos elementos
- No cluster 3 encontram-se 14 indivíduos, ou seja 12% dos elementos
- No cluster 4 encontram-se 47 indivíduos, ou seja 39,5% dos elementos
- No cluster 5 encontram-se 45 indivíduos, ou seja 38% dos elementos

Esta informação pode ser verificada visualmente na figura anterior.

4.2 CLUSTERIZAÇÃO COM 3 CLUSTERS

4.2.1 Clusterização com 3 Clusters dos Salários do Mês de Setembro

Os dados de entradas e os procedimentos utilizados para clusterização com 3 clusters foram os mesmos adotados no item anterior para a clusterização com 5 clusters.

A grande diferença implica na quantidade de clusters solicitado no algoritmo Fuzzy c-means, que diminuiu de 5 centróides para 3 centróides.

A seguir serão exibidos o quadro e a figura da representação dos clusters encontrados nesses agrupamentos.

Com o cálculo da distância euclidiana verificou-se que o cluster 3, é a classe que contém maior número de elementos, e que esta por sua vez apresenta uma proximidade com o

3 Clusters Setembro		
GRUPO	C	D
1	26,5	3358,1
2	50,1	1454,1
3	67,7	532

Quadro 6: Localização dos 3 centróides dos dados do mês de setembro.

Fonte: Dados da pesquisa (2008).

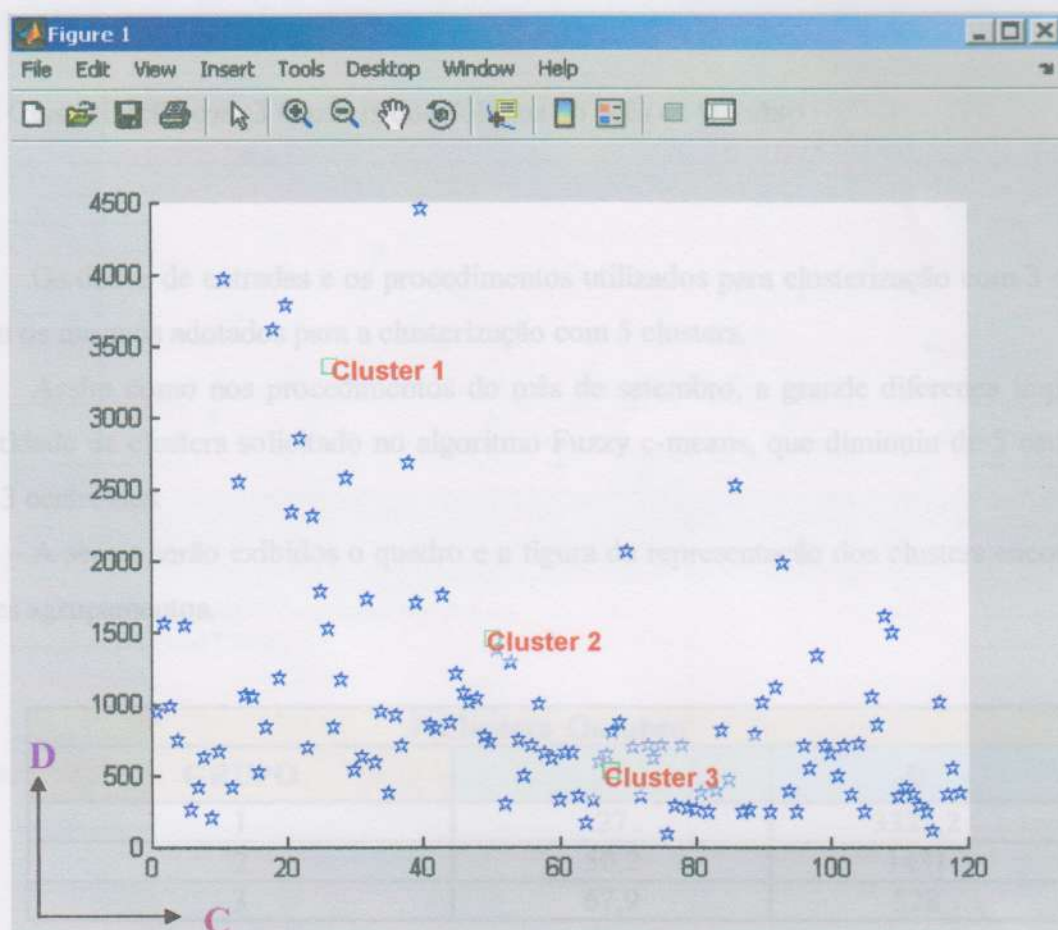


Figura 13 : Clusters Setembro (3 Centróides dos dados).

Fonte: Dados da pesquisa (2008).

É interessante notar que o número de iterações realizadas para se concluir a clusterização foi de 100 iterações.

Após a clusterização, com o objetivo de medir o grau de similaridade dos indivíduos para cada centróide, calculou-se a distância euclidiana de cada elemento padrão ao centro de cada cluster, para identificar a qual classe cada indivíduo pertence.

Com o cálculo da distância euclidiana verificou-se que o cluster 3, é a classe que contém maior número de elementos, e que esta por sua vez apresenta uma proximidade com o

cluster 2, porém menor se compara a proximidade das classes quando clusterizadas com 5 clusters, pois:

- No cluster 1 encontram-se 9 indivíduos, ou seja 8% dos elementos
- No cluster 2 encontram-se 28 indivíduos, ou seja 23% dos elementos
- No cluster 3 encontram-se 82 indivíduos, ou seja 69% dos elementos

Esta informação pode ser verificada visualmente na figura anterior.

4.2.2 Clusterização com 3 Clusters dos Salários do Mês de Outubro

Os dados de entradas e os procedimentos utilizados para clusterização com 3 clusters foram os mesmos adotados para a clusterização com 5 clusters.

Assim como nos procedimentos do mês de setembro, a grande diferença implica na quantidade de clusters solicitado no algoritmo Fuzzy c-means, que diminuiu de 5 centróides para 3 centróides.

A seguir serão exibidos o quadro e a figura da representação dos clusters encontrados nesses agrupamentos.

3 Clusters Outubro		
GRUPO	C	D
1	27	3321,2
2	50,2	1431
3	67,9	528

Quadro 7: Localização dos 3 centróides dos dados do mês de Outubro.

Fonte: Dados da pesquisa (2008).

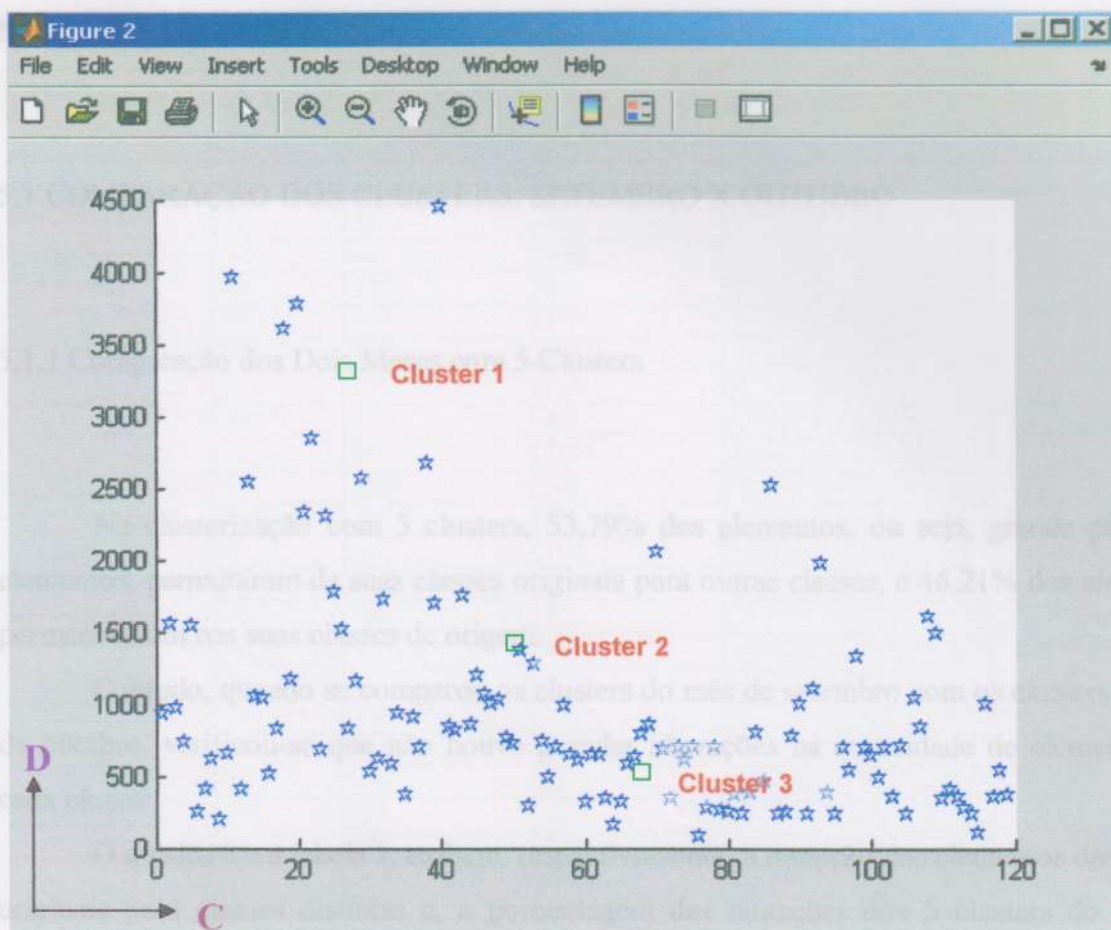


Figura 14 : Clusters Outubro (3 Centr ides dos dados).

Fonte: Dados da pesquisa (2008).

  interessante notar que o n mero de itera  es realizadas para se concluir a clusteriza  o continuou sendo de 100 itera  es.

Ap s a clusteriza  o, com o objetivo de medir o grau de similaridade dos indiv duos para cada centr ide, calculou-se a dist ncia euclidiana de cada elemento padr o ao centro de cada cluster, para identificar a qual classe cada indiv duo pertence.

Com o c lculo da dist ncia euclidiana verificou-se que o cluster 3,   a classe que cont m maior n mero de elementos, e que esta por sua vez apresenta uma proximidade um pouco maior com o cluster 2 do que no m s de setembro, mesmo que essa proximidade ainda continue menor quando comparada a proximidade das classes clusterizadas com 5 clusters, pois:

- No cluster 1 encontram-se 9 indiv duos, ou seja 8% dos elementos
- No cluster 2 encontram-se 29 indiv duos, ou seja 24% dos elementos
- No cluster 3 encontram-se 81 indiv duos, ou seja 68% dos elementos

Esta informa  o pode ser verificada visualmente na figura anterior.

5 CONCLUSÃO

5.1 COMPARAÇÃO DOS CLUSTERS: SETEMBRO X OUTUBRO

5.1.1 Comparação dos Dois Meses com 5 Clusters

Na clusterização com 5 clusters, 53,79% dos elementos, ou seja, grande parte dos elementos, permutaram de suas classes originais para outras classes, e 46,21% dos elementos permaneceram nas suas classes de origem.

Contudo, quando se comparou os clusters do mês de setembro com os clusters do mês de outubro, verificou-se que não houve grandes alterações na quantidade de elementos em cada cluster.

O quadro 8 e a tabela 1, exibem, respectivamente, a mutação dos elementos das classes originais para classes distintas e, a porcentagem das mutações dos 5 clusters do mês de setembro para o mês de outubro.

Mutação dos Elementos com 5 Clusters				
Entradas			Clusters (de 1 a 5)	
A	B (SETEMBRO)	B (OUTUBRO)	Setembro	Outubro
4	741	1542	4	3
8	620	1240	4	3
10	671	5612	4	1
11	3969	411	1	5
12	411	2547	5	2
13	2547	1062	2	4
15	1051	516	4	5
16	516	831	5	4
17	831	3606	4	1
18	3606	1179	1	4
19	1179	3787	3	1
20	3787	2332	1	2
22	2855	658	2	4
23	688	2259	4	2
24	2312	1780	2	3
26	1525	836	3	4
28	1159	2578	4	2
29	2578	1010	2	4
30	534	1199	5	4

32	1733	580	3	5
34	937	364	4	5
35	380	915	5	4
37	704	2681	4	2
38	2680	1703	2	3
39	1703	4460	3	1
40	4460	1198	1	4
42	827	3467	4	2
43	1757	867	3	4
44	867	2116	4	2
45	1205	1083	3	4
50	725	172	4	5
51	1371	460	3	5
52	300	1270	5	3
53	1289	868	3	4
54	739	516	4	5
55	487	783	5	4
58	648	608	4	5
59	608	293	4	5
62	656	348	4	5
65	321	635	5	4
68	798	2001	4	3
70	2058	432	2	5
74	606	83	4	5
75	702	273	4	5
76	83	708	5	4
78	708	314	4	5
84	810	2411	4	2
86	2516	246	2	5
87	246	779	5	4
88	259	1661	5	3
89	779	250	4	5
91	250	1974	5	3
92	1115	386	4	5
93	1974	247	3	5
94	386	700	5	4
96	700	1333	4	3
97	552	1029	5	4
105	246	853	5	4
107	853	1498	4	3
108	1614	350	3	5
109	1498	411	3	5
114	240	1001	5	4
116	1001	552	4	5
118	552	1104	5	4

Quadro 8: Mutação dos elementos com 5 clusters.

Fonte: Dados da pesquisa (2008).

Porcentagem da Mutação dos 5 Clusters					
Grupo	1	2	3	4	5
Setembro	3%	7%	12,5%	44%	33,5%
Outubro	3%	7,5%	12%	39,5%	38%

Tabela 1: Porcentagem da mutação dos elementos com 5 clusters.

Fonte: Dados da pesquisa (2008).

Assim, podem ser averiguados os elementos que permutaram de suas classes originais para classes distantes, como quando um elemento permuta do cluster 1 direto para o cluster 5.

E, a luz desses pressupostos, o dono da empresa pode examinar esses elementos que tiveram alto grau de variação em relação ao recebimento do salário de um mês para o outro. Ou seja, o administrador pode verificar no mês de outubro quais funcionários receberam bem mais, ou bem menos, do que no mês de setembro, e assim investigar a causa dessa variação.

O quadro abaixo exhibe os funcionários que tiveram alto grau de variação em relação a sua remuneração mensal:

Funcionários com alto grau de variação salarial (5 Clusters)				
Entradas			Clusters (de 1 a 5)	
A	B (SETEMBRO)	B (OUTUBRO)	Setembro	Outubro
10	671	5612	4	1
11	3969	411	1	5
12	411	2547	5	2
17	831	3606	4	1
18	3606	1179	1	4
40	4460	1198	1	4
70	2058	432	2	5
86	2516	246	2	5

Quadro 9: Funcionários que apresentaram alto grau de variação salarial quando clusterizados com 5 clusters.
Fonte: Dados da pesquisa (2008).

5.1.1 Comparação dos Dois Meses com 3 Clusters

Na clusterização com 3 clusters, apenas 1% dos elementos permutou de sua classe original para outra classe, ou seja, 99% dos elementos permaneceram nas suas classes de origem.

Quando se comparou os 3 clusters do mês de setembro com os 3 clusters do mês de outubro, verificou-se que houve ainda menos alterações na quantidade de elementos em cada cluster, se comparado a quantidade de elementos de cada cluster quando ocorreu a clusterização com 5 clusters.

O quadro 10 e a tabela 2, exibem, respectivamente, a mutação do único elemento que saiu da sua classe original para outra classe distinta e, a porcentagem das mutações dos 3

clusters do mês de setembro para o mês de outubro.

Muta��o do Elemento com 3 Clusters				
Entradas			Clusters (de 1 a 3)	
A	B (SETEMBRO)	B (OUTUBRO)	Setembro	Outubro
57	990	652	3	2

Quadro 10: Muta  o do elemento com 3 clusters.

Fonte: Dados da pesquisa (2008).

Porcentagem da Muta��o dos 3 Clusters			
Grupo	1	2	3
Setembro	8%	23%	69%
Outubro	8%	24%	68%

Tabela 2: Porcentagem da muta  o dos elementos com 3 clusters.

Fonte: Dados da pesquisa (2008).

Diferentemente do que ocorreu na clusteriza  o com 5 clusters, o motivo que fez o elemento 57 mudar de classe n o foi pela alta varia  o em sua faixa salarial, e sim pela maior proximidade que ele apresentou do centr ide 3 (saindo dessa forma do cluster 2) quando calculada a dist ncia euclidiana dos elementos para os clusters do m s de outubro. Pois, outros elementos apresentaram uma varia  o salarial bem maior que a do elemento 57 e, ainda assim permaneceram em seus clusters de origem.

5.2 COMPARA  O DOS PROCEDIMENTOS: 5 CLUSTERS X 3 CLUSTERS

Na finaliza  o da clusteriza  o dos dados dessa pesquisa foi poss vel verificar varia  es significativas das informa  es de cada indiv duo, como o grau de pertin ncia da varia  o salarial dos funcion rios da empresa.

Mas no momento em que a quantidade de clusters diminuiu, a dist ncia entre os centr ides das classes entre os indiv duos tamb m decresceu, dificultando a percep  o de varia  es significativas das informa  es de cada indiv duo.

Logo, a clusteriza  o efetuada com 5 clusters foi bem mais produtiva do que a clusteriza  o efetuada com 3 clusters.

5.3 CONCLUSÃO SOBRE O ALGORITMO FUZZY C-MEANS

Este trabalho analisou a proposição do algoritmo Fuzzy c-means em clusterizar elementos com 5 e 3 clusters.

Com base nos resultados obtidos, observou-se que mesmo com a diminuição da percepção de variações significativas quando decrescida a quantidade de clusters, o algoritmo Fuzzy c-means apresentou boa escalabilidade para particionar conjuntos de dados representativos, atingindo os resultados esperados no término desta pesquisa.

Apesar da sua notada complexidade matemática, o algoritmo operou no MATLAB de forma rápida e precisa, mantendo quase sempre o mesmo número de iterações para a finalização da clusterização.

REFERÊNCIAS

ALMEIDA, Evert Elvis Batista; SOUZA, Adauto José Ferreira. **Clusterização de dados superparamagnética**. VII Jornada de Ensino, Pesquisa e Extensão da Universidade Federal Rural de Pernambuco (VII JEPEX), 2007. Disponível em: <<http://www.adtevento.com.br/jepex/cdrom/resumos/R0138-1.pdf>>. Acesso em: 15 maio 2008.

AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de mineração de dados para gerar conhecimento a partir de bases de dados**. Universidade Federal de Pernambuco. Graduação em Ciência da Computação - Centro de Informática, 2006. Disponível em: <<http://www.cin.ufpe.br/~tg/2006-2/tmas.pdf>>. Acesso em: 30 maio 2008.

ARAÚJO, Anderson Viçoso de. **Árvores de decisão fuzzy na mineração de imagens do sistema footscanage**. Dissertação (Mestrado), Setor de Ciências Exatas, Universidade Federal do Paraná, 2006.

BILOBROVEC, Marcelo. **Sistema especialista em lógica fuzzy para o controle, gerenciamento, e manutenção da qualidade em processo de aeração de grãos**. Dissertação (Mestrado) - Universidade Tecnológica Federal do Paraná, Campus Ponta Grossa. Curso de Pós-Graduação em Engenharia de Produção. Ponta Grossa, 2005. Disponível em: <http://www.pg.cefetpr.br/ppgep/dissertacoes/diss_2005/Dissertacao_Final_Marcelo.pdf>. Acesso em: 09 junho 2008.

CAVALCANTI JÚNIOR, Nicomedes Lopes. **Clusterização baseada em algoritmos fuzzy**. Recife, Universidade Federal de Pernambuco, 2006. Dissertação (Mestrado) - UFPE. CIN - Ciências da Computação, 2006. Disponível em: <http://www.btd.ufpe.br/tedeSimplificado/tde_busca/arquivo.php?codArquivo=170>. Acesso em: 15 maio 2008.

FERRO, Mariza; LEE, Hwei Diana; **O Processo de KDD - Knowledge Discovery in Database para aplicações na medicina**. In: Seminc 2001. Cascavel. Anais da Seminc 2001. Disponível em: <http://labi.pti.org.br:8090/labi/documentos/SEMINC03_1.pdf>. Acesso em: 30 maio 2008.

GONÇALVES, Lóren Pinto Ferreira. **Avaliação de ferramentas para mineração de dados como fonte de dados relevantes para tomada de decisão: aplicação na Rede Unidão de Supermercados São Leopoldo - RS**. Universidade Federal do Rio Grande do Sul - Escola de Administração - Programa de Pós Graduação em Administração - Grupo de Estudos em Sistemas de Informação e de Apoio à Decisão, 2001. Disponível em: <http://volpi.ea.ufrgs.br/teses_e_dissertacoes/td/000410.pdf>. Acesso em: 30 maio 2008.

JANÉ, Dario de Almeida. **Uma introdução ao estudo da lógica fuzzy**. Hórus – Revista de Humanidades e Ciências Sociais Aplicadas, Ourinhos/SP, Nº 02, 2004. Disponível em: <http://www.faeso.edu.br/horus/artigos%20anteriores/2004/artigo_dario.pdf>. Acesso em: 06 junho 2008.

KAESTNER, Celso Antônio Alves. **Introdução à descoberta de conhecimento e mineração de dados**. Disciplina do Curso de especialização em Gestão de Tecnologia da Informação, Curitiba, Universidade Tecnológica Federal do Paraná, 2008.

MELLO, Carlos Eduardo R.; SILVA, Geraldo Zimbrão; SOUZA, Jano M. **Extensão do WEKA para métodos de agrupamento com restrição de contigüidade**. IX Brazilian Symposium on GeoInformatics, 2007, Campos do Jordão, Brasil. Disponível em: <<http://www.geoinfo.info/geoinfo2007/papers/S7P3.pdf>>. Acesso em: 27 abril 2008.

MIRANDA, Pedro; VILELA, Mauro Barbosa Jr.; KRONBAUER, Diego. **Sistema de controle difuso de mamdani, aplicações: pêndulo invertido e outras**. Monografia do término de curso em Análise de Sistemas. Universidade Federal de Mato Grosso do Sul, Departamento de Computação e Estatística, Centro de Ciências Exatas e Tecnologia, Mato Grosso do Sul, 2003. Disponível em: <<http://www.dct.ufms.br/~mzanusso/producao/PedroMir.pdf>>. Acesso em: 09 junho 2008.

NAVEGA, Sérgio. **Princípios essenciais do data mining**. São Paulo, 2002.

OLIVEIRA, Tatyana Bitencourt Soares. **Clusterização de dados utilizando técnicas de redes complexas e computação bioinspirada**. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação – ICMC/USP. Universidade de São Paulo, São Paulo, 2008. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-01042008-142253/publico/dissertacao.pdf>>. Acesso em: 15 maio 2008.

ORTEGA, Neli Regina Siqueira. **Aplicação da teoria de conjuntos fuzzy a problemas da biomedicina**. Tese (Doutorado) - Instituto de Física da Universidade de São Paulo. São Paulo, 2001.

PIMENTEL, Edson P.; FRANÇA, Vilma F. de; OMAR, Nizam. **A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização**. XIV Simpósio Brasileiro de Informática na Educação - NCE - IM/UFRJ - Rio de Janeiro, 2003. Disponível em: <<http://www.nce.ufrj.br/sbie2003/publicacoes/paper52.pdf>>. Acesso em: 02 junho 2003.

SANDRI, Sandra; CORREA, Cláudio. **Mini-curso de lógica nebulosa**. V Escola de Redes Neurais, Promoção: Conselho Nacional de Redes Neurais, INPE, São José dos Campos – SP,

1999. Disponível em: < <http://www.ele.ita.br/cnrm/minicursos-5ern/log-neb.pdf>>. Acesso em: 09 junho 2008.

SILVA, Marcelino Pereira dos Santos. **Mineração de dados - conceitos, aplicações e experimentos com Weka**. Escola Regional de Informática Rio de Janeiro - Espírito Santo, 2004. Vitória - ES, Rio das Ostras - RJ. Disponível em: <<http://www.sbc.org.br/bibliotecadigital/?module=Public&action=PublicationObject&subject=153&publicationobjectid=9>>. Acesso em: 30 março 2008.

SILVEIRA, Rosemari de Freitas. **Mineração de dados aplicada à definição de índices em sistemas de raciocínio baseados em casos**. Universidade Federal do Rio Grande do Sul - Instituto de Informática - Curso de Pós-Graduação WEB e Sistemas de Informação, 2003. Disponível em: < <http://www.inf.ufrgs.br/gpesquisa/bdi/publicacoes/files/Websis-RosemariSilveira.pdf>>. Acesso em: 30 maio 2008.

SIMÕES, Rodrigo Ferreira. **Complexos industriais no espaço: uma análise de fuzzy cluster**. Belo Horizonte: UFMG/Cedeplar, 2003. Disponível em: <<http://www.cedeplar.ufmg.br/pesquisas/td/TD%20209.pdf>>. Acesso em: 15 maio 2008.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to data mining**. Addison-Wesley, 2004.

VALE, Marcos Neves do. **Agrupamentos de dados: avaliação de métodos e desenvolvimento de aplicativo para análise de grupos**. Dissertação (Mestrado), PUC Rio de Janeiro, Programa de Pós-Graduação em Engenharia Elétrica, 2005. Disponível em: < http://www.maxwell.lambda.ele.puc-rio.br/cgi-bin/db2www/PRG_0651.D2W/SHOW?Mat=&Sys=&Nr=&Fun=&CdLinPrg=pt&Cont=7975:pt>. Acesso em: 15 maio 2008.

VIGLIONI, Giovanni Melo Carvalho. **Metodologia para previsão de demanda ferroviária utilizando data mining**. Dissertação (Mestrado) - Instituto Militar de Engenharia, Rio de Janeiro, 2007. Disponível em: <<http://transportes.ime.eb.br/DIS/DIS229.pdf>>, Acesso em: 13 junho 2008.