UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

EKUIKUI VANILSON DOS ANJOS ROSA

AGRUPAMENTO NÃO SUPERVISIONADO FUZZY C- MEANS

TRABALHO DE CONCLUSÃO DE CURSO

PATO BRANCO

2017

EKUIKUI VANILSON DOS ANJOS ROSA

AGRUPAMENTO NÃO SUPERVISIONADO FUZZY C- MEANS

Trabalho de Conclusão de Curso como requisito parcial à obtenção do título de Bacharel em Engenharia de Computação, do Departamento Acadêmico de Informática da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Dalcimar Casanova

PATO BRANCO

2017

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Pato Branco
Departamento Acadêmico de Informática
Curso de Engenharia de Computação

TERMO DE APROVAÇÃO

Às 10 horas e 20 minutos do dia 05 de dezembro de 2017, na sala V104, da Universidade Tecnológica Federal do Paraná, Câmpus Pato Branco, reuniu-se a banca examinadora composta pelos professores Dalcimar Casanova (orientador), Luciene de Oliveira Marin e Pablo Gauterio Cavalcanti para avaliar o trabalho de conclusão de curso com o título **Agrupamento não supervisionado fuzzy C- Means**, do aluno **Ekuikui Vanilson dos Anjos Rosa**, matrícula 01379305, do curso de Engenharia de Computação. Após a apresentação o candidato foi arguido pela banca examinadora. Em seguida foi realizada a deliberação pela banca examinadora que considerou o trabalho aprovado.

_____
Dalcimar Casanova
Orientador (UTFPR)

_____
Luciene de Oliveira Marin
(UTFPR)

_____
Pablo Gauterio Cavalcanti
(UTFPR)

_____
Profa. Beatriz Terezinha Borsoi
Coordenador de TCC

_____
Prof. Pablo Gauterio Cavalcanti
Coordenador do Curso de
Engenharia de Computação

A Folha de Aprovação assinada encontra-se na Coordenação do Curso.

RESUMO

ROSA, Ekuikui Vanilson dos Anjos. Agrupamento não Supervisionado Fuzzy C-means. 2017. 76f. Monografia (Trabalho de conclusão de Curso) – Curso de Engenharia de Computação. Universidade Tecnológica Federal do Paraná, Campus Pato Branco. Pato Branco, 2017.

A tarefa de agrupamento visa identificar e aproximar os dados similares. Um agrupamento (ou cluster) é uma coleção de dados similares entre si, porém diferentes dos outros registros nos demais grupos. Nesse contexto muitos métodos de agrupamentos já foram propostos e ainda são fonte de diversas pesquisas científicas que objetivam melhor separação dos dados. Dentre os métodos mais conhecidos estão o k-means e o fuzzy c-means e ambos possuem implementações em diversos pacotes comerciais e livres, todavia os usuários comuns tendem a utilizar tais métodos indiscriminadamente, sem conhecer suas implicações e diferenças. Esse trabalho visa contribuir no sentido de dar clareza e exemplos práticos do comportamento do método fuzzy c-means, o qual é baseado na teoria dos conjuntos nebulosos.

Palavra-chave: Agrupamento não supervisionado. Reconhecimento de padrões. Fuzzy c-means.

ABSTRACT

ROSA, Ekuikui Vanilson dos Anjos. Non-supervised clustering fuzzy c-means. 2017. 76f. Monografia (Trabalho de conclusão de Curso) – Computer Engineering major. Federal University of Technology – Paraná, Campus Pato Branco. Pato Branco, 2017.

The clustering task aims to identify and approximate similar data. A cluster is a collection of data similar to each other, but different from the other records in the other clusters. In this context, many of these clusters have already been proposed and are still the origin of several scientific studies aimed at obtaining the best data separation. Among the best known methods are k-means and fuzzy c-means. Both have deployments in various commercial and free packages. However, ordinary users tend to use such methods indiscriminately, without knowing their implications and differences. This work aims to contribute precisely in this sense, to give clarity and practical examples of the behavior of the method fuzzy c-means, which is based on the theory of fuzzy sets.

Keywords: Non-supervised clustering. Pattern Recognition. Fuzzy c-means.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

KDD *Knowledge discovery in databases* ou descoberta de conhecimento em base de dados

FCM Fuzzy c-means

PCA Análise de componentes principais

LVQ *Learning vector quantization* ou aprendizado de quantização vetorial

SOM *self-organizing maps* ou mapas auto-organizados

SUMÁRIO

1. INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

Agrupamento é uma técnica para encontrar aglomerados semelhantes em conjunto de dados. Esses aglomerados semelhantes são, usualmente, chamados de grupos. O agrupamento de dados está presente em diversas áreas em aplicações como processamento de imagens, mineração de dados e processamento de padrões.

A ideia central é formar grupos, onde os dados pertencentes ao mesmo grupo são semelhantes, de forma que dados de grupos diferentes são dissimilares. Classificar dados semelhantes em grupos é importante para as atividades de aprendizado do homem; por exemplo, uma criança aprende de forma natural a distinguir entre cães e gatos, entre homem e mulher. Essa aptidão adquirida faz parte do que chamamos “inteligência” e por isso a pesquisa de agrupamento é frequentemente considerada uma parte da inteligência artificial e do reconhecimento de padrões (LOUREIRO, 2005).

Figura 1 - Agrupamento de objetos
Fonte: Autor desconhecido

Figura 2 - Exemplo de um agrupamento entre objetos
Fonte: Autor desconhecido

A Figura 1 demonstra um exemplo de um conjunto de aglomerados de objetos. Um agrupamento natural possível entre esses objetos da Figura 1 são: funcionários da escola e família (Figura 2), mas também outra pessoa pode fazer o agrupamento entre homens e mulheres. Portanto o agrupamento é uma técnica computacional subjetiva.

As pesquisas na área de agrupamento de dados não servem apenas para identificar estruturas já existentes nos dados, mas também para descobrir conhecimentos novos que, por conseguinte, podem trazer melhora nos processos ou atividades.

Dentro desse contexto temos os métodos de *clustering*, que são técnicas de agrupamento não supervisionadas para fazer agrupamento automático desses dados segundo seu grau de semelhança. Nesta classe de métodos um dos mais conhecidos é o k-means.

Agrupamento k-means é uma heurística de agrupamento não hierárquico que busca minimizar a distância dos elementos a um conjunto de k centros. Este agrupamento permite separar objetos em grupos levando em conta as características destes objetos.

Entretanto o k-means não é adequado para determinados problemas. Há casos onde não basta apenas agrupar esses elementos e ter a plena certeza de que os mesmos estão no grupo certo. Tem casos em que o elemento pode pertencer a um ou mais grupos e, nesse sentido k-means não garante tal abstração.

Para resolver esse problema é necessário desenvolver um algoritmo de clustering fuzzy, que será o responsável por verificar se de fato o objeto pertence a um determinado grupo ou mais.

Métodos fuzzy são baseados em graus de pertinência de cada elemento a um determinado conjunto, ao contrário da lógica booleana nesta lógica temos meio termo. Isto é, admite valores intermediários entre falso (“0”) e o verdadeiro (“1”) (TRONCO, 2015). Dentro dos algoritmos de agrupamento fuzzy temos o algoritmo c-means.

Fuzzy c-means é um algoritmo de agrupamento que permite que um mesmo dado pertença a um ou mais grupos com diferentes graus de pertinência. O fuzzy c-means é praticamente uma implementação fuzzy do algoritmo k-means, compartilhando de muitas de suas vantagens e desvantagens (ROCHA, 2009).

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo deste trabalho é realizar um estudo em detalhe dos métodos de agrupamento k-means e fuzzy c-means, tendo como foco aprender seus conceitos fundamentais.

É fundamental distinguir que, em todo problema de agrupamento, faz-se necessário a tomada de decisões sobre alguns aspectos tais como, que objetos usar para realizar o agrupamento, que medida de similaridade ou dissimilaridade usar, como os grupos devem ser formados e qual método produz melhor otimização de grupo.

1.2.2 Objetivos Específicos

a) Discutir o método de agrupamento k-means não supervisionado.
b) Compreender os conceitos fundamentais de lógica fuzzy.

c) Construir o método de agrupamento fuzzy c-means a partir dos conceitos do k-means e lógica fuzzy.
d) Selecionar um conjunto de dados para teste da técnica.
e) Aplicar a solução construída ao conjunto selecionado.
f) Analisar os resultados do algoritmo aplicado e identificar sua validade.

1.3 JUSTIFICATIVA

Durante as últimas décadas, temos visto um crescimento explosivo da capacidade do homem em gerar e coletar dados. Computadores com grande capacidade de armazenamento têm contribuído para a proliferação de grandes bancos de dados em diversas áreas. Instrumentos científicos, por exemplo, podem produzir terabytes e petabytes de dados a uma taxa de gigabytes por hora. Outros exemplos de avanço¸ nesta área podem ser vistos, frequentemente, no uso de código de barras em produtos comercias e na automação de muitos negócios e transações governamentais. Contudo, também existe a necessidade de uma nova geração de teorias computacionais e ferramentas para assistir o homem na tarefa de extração de informação e conhecimento de grandes volumes de dados. Estas teorias e ferramentas são o foco de uma recente área de pesquisa que trata da descoberta de conhecimento em banco de dados, conhecida como *Knowledge discovery in databases* (KDD) (LOUREIRO, 2005).

Nesse contexto os métodos de agrupamento têm papel central e são muito utilizados nos últimos anos em diversas aplicações.

1.4 ESTRUTURA

Este trabalho está estruturado da seguinte forma: O capitulo 2 apresenta o referencial teórico, onde tem-se uma breve descrição sobre obtenção de conhecimento em bases de dados, (Seção 2.1). Na sequência, é apresentado as etapas do processo KDD (seção 2.2), em seguida são apresentadas as

técnicas de mineração de dados (seção 2.3). Na sequência, é apresentado as medidas de Similaridade e Dissimilaridade (seção 2.3.1), bem como o agrupamento k-means e seu funcionamento (seção 2.3.2). Na seção 2.4 é apresentado a técnica de redução de dimensionalidade, na sequência é apresentado a teoria dos conjuntos nebulosos (seção 2.5), bem como o algoritmo de agrupamento Fuzzy c-means (seção 2.5.6). No Capítulo são apresentadas as bases de dados Íris (seção 3.1), sequência a base de câncer de mama (seção 3.2), e o ambiente computacional (seção 3.3). Finalizando, no Capitulo 4 é apresentado um exemplo número para o algoritmo fuzzy c-means, na sequência são apresentados os resultados e comparação entre os dois métodos, e o Capítulo 5 mostra a conclusão deste trabalho.

.

2. REFERENCIAL TEÓRICO

2.1 KNOWLEDGE DISCOVERY IN DATABASES

Atualmente tem-se a necessidade de observar o homem na função de obtenção de informação e conhecimento em grande quantidade de dados. Estes mecanismos e teorias são o centro de um campo de indagação que trata da exploração de conhecimento em banco de dados, conhecida como *Knowledge Discovery in Databases* (KDD - Descoberta de conhecimento em base de dados).

A exploração de conhecimento em base de dados tem gerado diversos nomes, tais como, *database exploration*, *information discovery, knowledge extraction, pattern processing*, e *data mining* ou mineração de dados como mais conhecido (FAYYAD et al., 1996). Indepentemente do nome, a fundamentação de KDD é obtenção de informação de maneira não trivial, primeiramente desconhecida, e muito útil dos dados (FRAWLEY et al., 1992).

É relevante realçar que KDD descreve toda técnica de descoberta de sabedoria envolvendo a forma: como os dados são acessados e associados, como algoritmos são selecionados e eficazes no seu desempenho, como os resultados são analisados e vistos, e de que forma a relação homem-máquina é formada e mantida de maneira favorável (FAYYAD et al., 1996).

A exploração de conhecimento em banco de dados é um campo amplo que associa sistemas científicos diferentes, com métodos estatísticos, reconhecimento de padrões, alta performance computacional, visualização de dados, aprendizado de máquina, e inteligência artificial, estes campos podem ser vistos na Figura 3. O propósito de agregar é a obtenção de conhecimento de alto nível, a partir de um conjunto de dados de baixo nível, no âmbito de um conjunto maior de banco de dados (FAYYAD et al., 1996).

Figura 3 - Campos de descoberta de conhecimento
Fonte: Autor desconhecido

2.2 PROCESSO KDD

Como referido, KDD é o procedimento de reconhecimento de padrões apropriado, não trivial, recente, utilizáveis e de fácil compreensão a partir de um conjunto de dados. Este procedimento apresenta duas características, iterativa e interativa, formado por uma sequência de tarefas, ilustradas na Figura 4. O processo é dito iterativo quando sua composição é feita por um conjunto de etapas de continuidades, onde pode se dar o caso de voltar a etapas anteriores. O processo é interativo pois, em todas suas fases de construção são tomadas decisões dependentes do domínio de conhecimento por parte do usuário e da aplicação (LOUREIRO, 2005).

Segundo Adriaans e Zantingen (1996), as atividades do KDD podem ser agrupadas em três etapas:

a) Pré-Processamento - controla a aplicação, faz a seleção dos dados onde será executado o processo, faz limpeza dos dados, retira ruídos e separa um conjunto de amostras.

b) Mineração de dados – Faz a seleção dos algoritmos que serão utilizados na aplicação. Esses algoritmos serão responsáveis por buscar os padrões e aplicar os mesmos.

c) Pós - **Processamento** tem como função esclarecer os padrões encontrados, aceitar eles ou não, fazendo uma seleção das descobertas interessantes e usar as mesmas para várias aplicações.

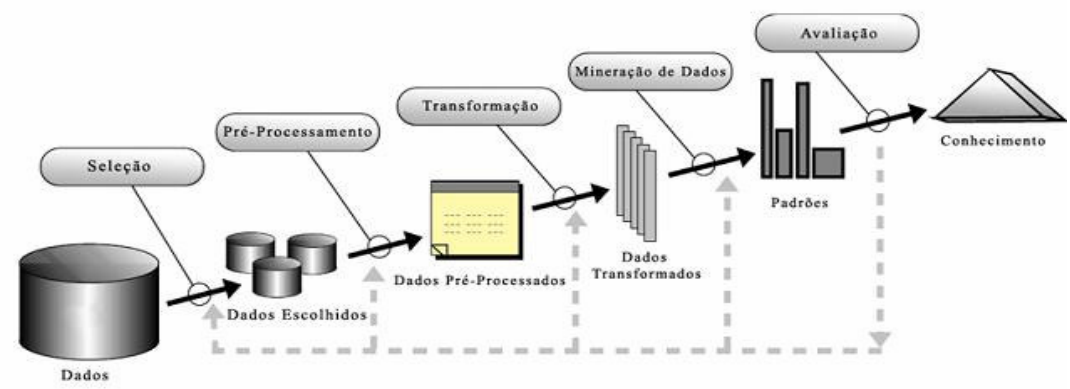

Figura 4 - Tarefas no processo de KDD
Fonte: traduzido de Loureiro (2005, p. 12).

2.3 TÉCNICAS DE MINERAÇÃO DE DADOS

Algumas das técnicas de *data mining* foram criadas com intuito de obter informações e dados, *data mining* é a combinação de diferentes técnicas de sucesso comprovado tais como, inteligência artificial estatística e banco de dados (LOUREIRO, 2005).

As principais divisões dentre as técnicas de mineração de dados são:

- Ferramentas de consulta e técnicas de estatísticas.
- Técnicas de Visualização.
- Classificação.
- Associação.
- Sumarização.
- Análise de Sequência.
- Regressão.
- Clustering.

Ferramentas de consulta e técnicas de estatísticas - Normalmente são utilizadas como o primeiro passo na fase de mineração, como uma análise superficial de um conjunto de dados que se deseja minerar, com uso de ferramentas de consultas (exemplo SQL). A partir de um banco de dados

relacional é possível obter informações bastante importantes de como os dados são distribuídos, através de funções built-in da linguagem SQL. É preciso compreender as particularidades e estruturas de aglomerados de dados que serão minerados antes de aplicarmos os algoritmos de reconhecimento de padrões (CÔRTES et al, 2002).

Com estatísticas do tipo desvio padrão, média aritmética, medidas de dispersão e medidas de separatrizes, constituem-se os passos iniciais na mineração de dados, permitindo obter um conjunto amplo de porção de informações (ADRIAANS et al, 1996).

Técnicas de Visualização - Conjunto de recursos utilizados que ajudam na compreensão dos padrões obtidos de um conjunto de dados. (LOUREIRO, 2005).

Classificação – Consiste em identificar em que classe um objeto pertence, verificar as caraterísticas do mesmo e atribuí-lo a um conjunto de classes predefinidas (HAND, 1981).

Associação – É um processo de relações importantes entre objetos em um conjunto de dados (HAN et al, 2002).

Sumarização -Tem como objetivo disponibilizar fácil entendimento dos métodos com definições muito compactas dentro de um conjunto de dados em estudo (CÔRTES et al, 2002).

Análise de sequência – É um processo de regras de associação utilizado para reconhecer as sequências importantes, no qual serão utilizadas para adiantar eventos decorrentes. Portanto o processo de exploração compreende regras e relações atendendo o tamanho temporal (LOUREIRO, 2005).

Regressão – Utilizada quando o registro é reconhecido a partir de um valor numérico. Tem como finalidade estimar o custo de uma variável definida em observação, pelo meio de ligação sequencial ou não com um agregado de outras variáveis que podem ter a possibilidade de caracterizar a variável em observação (DRAPER et al, 1981).

Neste trabalho a técnica empregada é a de **Clustering**, que consiste em dividir um conjunto de dados em diferentes grupos (JAIN, 1988). O foco dessa técnica é obter grupos que são dissimilares uns dos outros, nas quais seus

membros são similares entre si. Os grupos podem ser utilizados a partir dos resultados obtidos para classificar dados novos. Os algoritmos mais utilizados para realizar esse agrupamento são o k-means e fuzzy c-means.

2.3.1 Medidas de Similaridade e Dissimilaridade

Normalmente os algoritmos que executam tarefas de análise de dados empregam alguma medida de similaridade entre vetores durante o seu processo de execução. Estas medidas de similaridade servem para orientar o processo de construção do espaço de decisão que definirá qual é a região de alcance de uma classe de dados no caso da tarefa de classificação, ou seja, quais dos dados pertencem a quais grupos no caso da tarefa de agrupamento (ROCHA et.al, 2012).

A similaridade entre dois vetores diz respeito a uma medida que compara a igualdade dos dois vetores, pode ser também utilizada como uma maneira de medir a diferença entre eles (a dissimilaridade). Normalmente, a dissimilaridade é calculada a partir de uma medida de distância entre dois pontos de um grupo.

No processo de análise de dados podem ser aplicadas diferentes medidas de distância, como Distância de Chebyschev, Distância City-Block, Distância de Hamming, Distância euclidiana, Métrica de Tanimoto, Métrica de Mahalanobis.

Neste trabalho, a medida de dissimilaridade aplicada visa medir o quão dissimilar um ponto de dados é em relação a outro de uma classe ou grupo. A Distância Euclidiana é a métrica escolhida para ser a base do cálculo de dissimilaridade entre pontos de um grupo de dados, a Figura 5 mostra o cálculo da distância Euclidiana entre dois pontos, que é calculada a partir da Equação 1.

$$d(i,j) = \sqrt{((|x_{\downarrow}i1 - x_{\downarrow}j1|^{\uparrow}2 + |x_{\downarrow}i2 - x_{\downarrow}j2|^{\uparrow}2 + \cdots + |x_{\downarrow}ip - x_{\downarrow}jp|^{\uparrow}2))} \quad (1)$$

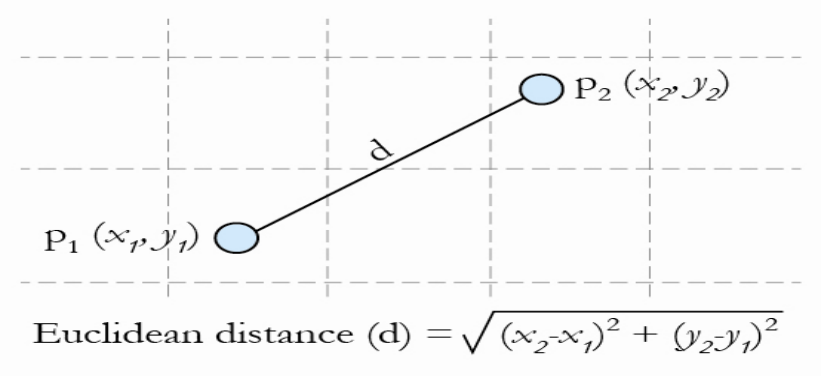

Figura 5 - Distância euclidiana aplicada entre dois pontos
Fonte: Autor desconhecido.

2.3.2 Agrupamento K-Means

Em (1967) McQueen apresentou o algoritmo k-means, o qual é bastante utilizado em várias aplicações, empregando seus centroides de cada grupo como representação de seus pontos. Uma vez que seu desempenho depende da seleção inicial de seus centroides, isso faz com que se torne muito mais eficiente em relação aos algoritmos hierárquicos tradicionais (BRADLEY et al., 1998).

Os centroides finais dos grupos na maioria dos métodos de agrupamento, não apresentam uma solução excelente global, mas sim apresentam apenas uma solução ótima local. Assim sendo, agrupamentos que são totalmente distintos são provenientes às vezes, das escolhas de seus centroides inicias. Segundo alguns pesquisadores, existem vários procedimentos na escolha dos centroides: definir os primeiros k pontos (McQueen, 1967); definir aleatoriamente os k pontos (BALL et al., 1965).

O algoritmo k-means é particional, exigindo a escolha do número de agrupamentos k como entradas. Para encontrar um ótimo local este algoritmo utiliza a técnica de realocação interativa. Por ser um algoritmo simples e com várias implementações, ele é frequentemente utilizado, apesar de sua limitação na definição inicial do número de grupos k (SIMOVIC, 2007).

O algoritmo k-means independentemente de ser bastante utilizado em uma grande gama de aplicações, tem suas desvantagens, dentre elas destacam-se:

- Antes do algoritmo k-means ser inicializado o número de grupos a priori tem que ser conhecido.
- O conjunto de objetos são obrigados a pertencerem a um grupo.
- A convergência do algoritmo k-means é limitada a um domínio local.

O k-means tem por finalidade minimizar o erro quadrático, estabelecido pela Equação (2), onde μ_w é o vetor centroide do grupo G_j e $d\ (x_i\ \mu_w)$ é a distância euclidiana entre x_i e μ_w, isto é, o critério seguido pelo k-means é minimizar a distância entre cada ponto e o centroide do agrupamento no qual o ponto pertence (ANDRADE FILHO, 2013).

$$E = \sum_{w=1}^{k} \sum_{x_i \in G_w} d(\mathbf{x_i}, \mu_w)^2 \qquad (2)$$

A implementação da função de agrupamento de dados k-means, que possui como parâmetros o conjunto de dados *X* e o número de grupos *k,* é mostrado no pseudocódigo abaixo. O processo utilizado é de minimizar o erro quadrático entre os exemplos e o centróide do grupo na qual o exemplo pertence. Assim sendo, como resultado tem-se uma partição composta por grupos de formato hiperesférico com o mesmo tamanho, ou seja, por grupos bem separados (ANDRADE FILHO, 2013).

K-means – Algoritmo1

Entrada: Conjunto de dados X
Entrada: Número de grupos *k*
Saída: Partição do conjunto X em *k* grupos
inicio
  Definir aletoriamente *k* centróides de grupos;
  repita

    para cada *exemplo* $x_i \in X$ *e cada grupo* G_w, $w = 1,\ldots k$ **fazer**

      Calcular a distância entre x_i e o centroide do grupo G_w;
    fim
    para cada *exemplo* x_i **fazer**
      Associar x_i ao grupo com centroide mais próximo;
    fim
    para cada grupo G_w **fazer**
      Recalcular o centroide;
    fim
  até *que nenhuma alteração nas associações de exemplos a grupos seja realizada*;
  retorna *Partição formada pelo* G_k *grupos*
fim

Para os critérios de parada do k-means frequentemente é utilizado: a execução de um determinado número de iterações, e estabilização dos centroides, isto é, quando nenhum elemento muda de posição. Esses critérios de parada podem ser utilizados em conjunto ou individualmente. É preciso enfatizar que o resultado desse algoritmo depende de como os centroides são definidos pelo programador incialmente (ANDRADE FILHO, 2013).

Este algoritmo é extremamente veloz, geralmente convergindo em poucas iterações para uma configuração estável, na qual nenhum elemento está designado para um cluster cujo centro não lhe seja o mais próximo. Um exemplo da execução do algoritmo de k-means pode ser visto na Figura 6 (LINDEN, 2009).

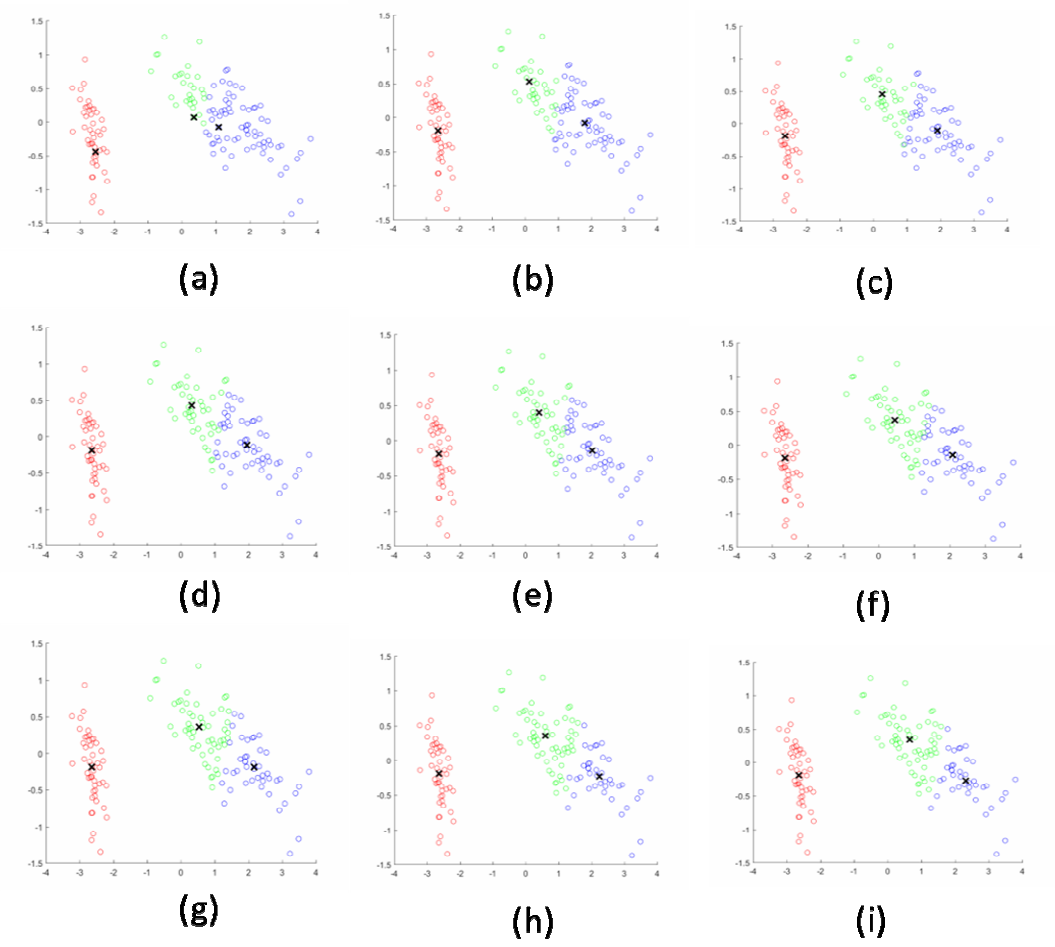

Figura 6 - Exemplo de execução do algoritmo de k-means

Na Figura 6 em (a) cada elemento foi indicado para um dos três grupos aleatoriamente e os centros (**x**) de cada grupo foram calculados. Em (b) os elementos foram designados agora para os grupos cujos centros lhe estão mais próximos e seus centros foram recalculados. Em (c) alguns dos elementos de cor verde e azul mudaram de grupo e seus centros foram recalculados. Em (d) os centros foram recalculados pois alguns elementos de cor azul e verde mudaram de grupo. Em (e) novamente elementos de cor azul e verde mudaram de grupo e seus centros foram recalculados. Em (f) novamente alguns elementos de cor azul e verde mudaram de grupos e seus centros foram recalculados. Em (g) os centros foram novamente recalculados pois alguns elementos de cor azul e verde mudaram de grupo. Em (h) é a etapa final em que elementos de cor azul e verde mudaram de grupo e novamente os centros foram recalculados. Em (i) os centroides foram recalculados e os grupos já estão em sua forma final pois nenhum elemento tanto de cor vermelha azul e verde mudou de grupo. Caso não estivessem, repetir-se-ia as iterações até que estejam na sua forma final.

Este algoritmo iterativo tende a convergir para um mínimo da função de energia definida em cima equação (2).

Um eventual problema é que esta condição enfatiza a questão da homogeneidade e ignora a importante questão da boa separação dos clusters. Isto pode causar uma má separação dos conjuntos no caso de uma má inicialização dos centros, inicialização esta que é feita de forma arbitrária (aleatória) no início da execução. Na Figura 7 pode-se ver o efeito de uma má inicialização na execução de um algoritmo de k-means (LINDEN, 2009). Percebe-se que em determinados casos, quando se possui uma inicialização ruim, o resultado é divergente do esperado.

Outro ponto que pode afetar a qualidade dos resultados é a escolha do número de conjuntos feita pelo usuário. Um número pequeno demais de conjuntos pode causar a junção de dois clusters naturais, enquanto que um número grande demais pode fazer com que um cluster natural seja quebrado artificialmente em dois (LINDEN, 2009).

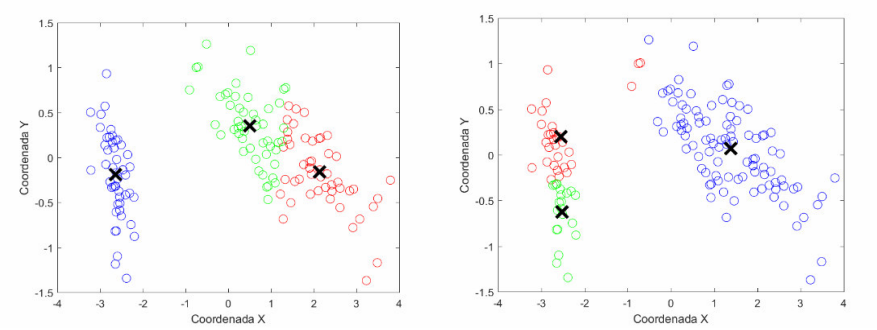

Figura 7 - Exemplo do efeito de uma má inicialização do algoritmo k-means

2.4 ANÁLISE DE COMPONENTES PRINCÍPAIS (PCA)

Criado em 1901 por Karl Pearson, PCA (do inglês *Principal Component Analysis*) será aplicada na análise de agrupamento de dados implementado a partir do algoritmo k-means. Esta técnica tem por objetivo básico, a análise dos dados utilizados tendo em vista sua redução, excluir sobreposições desses dados e seleção das formas de melhor representação de dados com base nas combinações lineares das variáveis originais.

Pesquisar as relações de aglomeração com **p** variáveis correlacionadas podem ser bastante eficiente para se modificar o conjunto de variáveis originais em um conjunto novo de variáveis não correlacionadas denominadas *componentes principais*, havendo caraterísticas especiais com base em variâncias. As variáveis novas são denominadas de combinações lineares das variáveis originais com derivação em ordem decrescente (MARQUES, 1994). A Figura 8 demonstra a uma aplicação da técnica PCA.

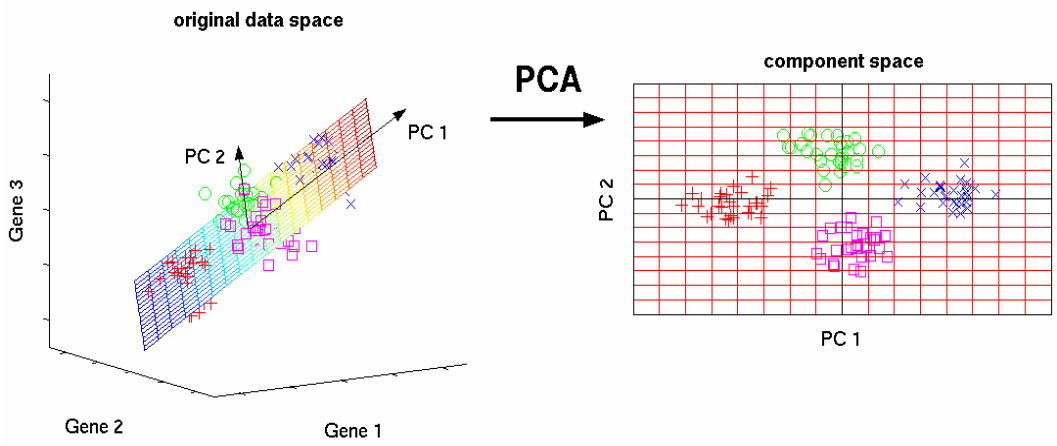

Figura 8 - exemplo da aplicação do PCA

Fonte: Autor desconhecido

As Figuras (9) e (10) abaixo demonstram a implementação do algoritmo k-means com e sem a utilização da técnica de redução de dimensionamento PCA. Na Figura 9 a visualização dos dados é realizada utilizando apenas as primeiras características dos dados (originalmente são 4 características). Importante salientar que é impossível plotar dados em 4 dimensões, logo esse tipo de técnica é comumente empregada na visualização de dados multivariados. Já na Figura 10 os dados estão transformados utilizando método de PCA, ou seja, as 2 dimensões mostradas são uma combinação linear de todas as outras dimensões, ou seja, se está representado toda a informação dos dados de forma resumida, em duas dimensões.

O objetivo da aplicação do PCA nesse caso é apenas para diminuir a dimensionalidade dos dados e assim obter uma melhor visualização dos mesmos (em duas dimensões).

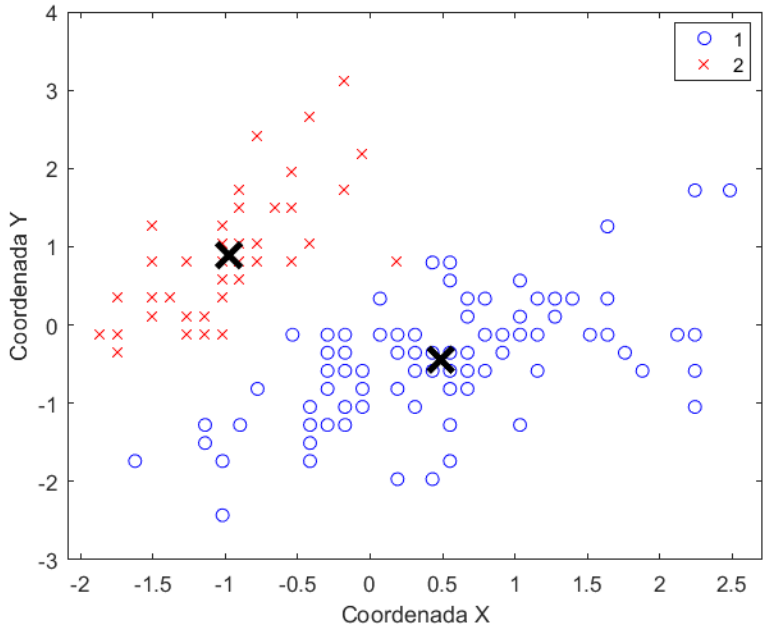

Figura 9 - Representação do algoritmo k-means sem a utilização do PCA

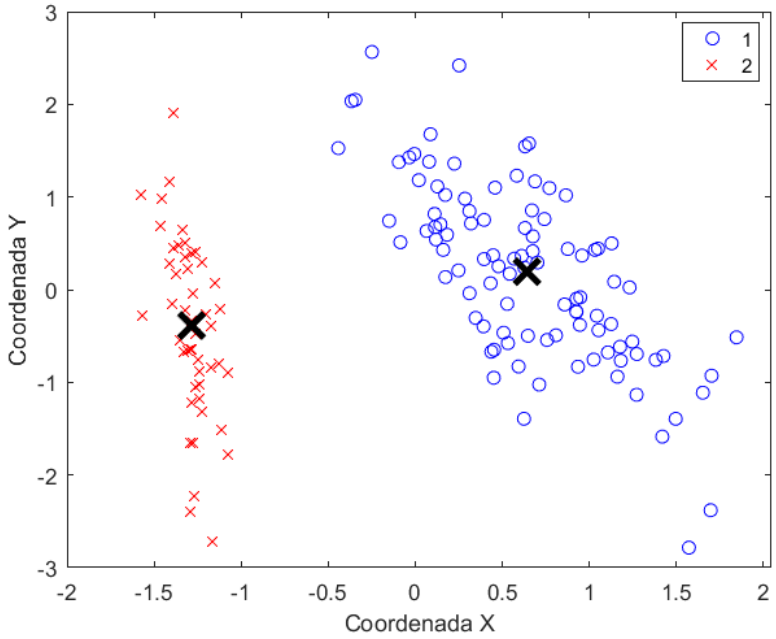

Figura 10 - Representação do algoritmo k-means com a utilização do PCA

2.5 TEORIA DOS CONJUNTOS NEBULOSOS (FUZZY)

Em 1965 por Lofti Asker Zadeh a teoria dos conjuntos nebulosos, foi inserida como um novo modo de descrever princípios indefinidos no compartilhamento de informações entre seres humanos. Princípios do tipo “homens bonitos”, “pessoas baixas”, “curto intervalo de tempo”, “diminuir a velocidade antes”, especificam ideias incertas no qual não podem ser abordados como conjuntos tradicionais. Precisamente, o princípio de conjuntos nebulosos foi criado como forma geral da teoria tradicional dos conjuntos, atribuindo ferramentas pequenas essenciais ao expandir a descrição de operações como aceitabilidade de um componente. Sucessivamente, Zadeh aprofundou sua teoria propondo uma forma próxima, porém capaz de explicar o comportamento de processos muito relevantes, bem como os processos referidos a conhecimentos humano. Nela, é destacado que os métodos quantitativos tradicionais não são propícios para tratar de meios com complexidades semelhantes aos meios humanos. Estes meios são especificados pelo fato de não serem compreendidos em pormenores, por outro lado muitas das vezes podem ser formados de maneira empírica (BONVENTI JUNIOR, 2005).

A partir da modelação de fatores eficientes e criação de classificadores com base na teoria dos conjuntos nebulosos é possível que normas de heurísticas sejam capazes de apreender as técnicas de determinação associadas à especificação ou execução. Com o uso de princípios eventuais, os seres humanos podem empregar técnicas de operações acima de ideias vazias, exemplo: “usarei uma camisa fina, se hoje fizer mais calor”.

Incerteza: Vago (vazio) ou probabilidade. O fato de um modelo demonstrar um pensamento vazio (nebuloso) faz com que o mesmo se diferencie de um modelo probabilístico.

Figura 11 - Exemplo de incerteza
Fonte: autor desconhecido

No exemplo clássico do sorteio das bolas da Figura 12, diz-se que um conjunto de 6 bolas amarelas e vermelhas, existe a probabilidade de 1/3 de sair bola amarela. Entretanto quando a bola é sorteada, a probabilidade se dispersa: a bola é (ou não) amarela. A *possibilidade* de uma bola ser de cor amarela, apontada particularmente, é uma regra *booleana.* O conceito de probabilidade dá-se pela forma de lidar com a incerteza oriunda da ausência de conhecimento ou informação. A variação da probabilidade ocorre por meio de incerteza absoluta (0) e certeza absoluta (1) no qual, um evento está agregado a uma definição (exemplo, um acontecimento) (BONVENTI JUNIOR, 2005).

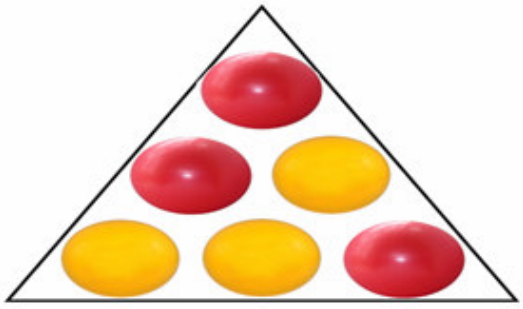

Figura 12 - Exemplo probabilidade

Por outro lado, a vaguidade (vago) calcula a quantidade que um valor numérico se encaixa ao princípio ideal. O vago não se modifica com o passar do tempo, por ser uma propriedade específica de um objeto ou acontecimento. A possibilidade é dada pela abordagem de uma ocorrência para ser analisado

em comparação a um conceito. O princípio da teoria dos *conjuntos nebulosos* – *"fuzzy sets"*- refere-se como uma espécie possibilista de vagueza dos elementos em relação aos conjuntos (BONVENTI JUNIOR, 2005).

Um exemplo elucidativo que foi concedido em 1992 por Bezdek foi: Considere que você está em um deserto com duas garrafas marcadas uma como X e outra como Y. No rótulo da garrafa X lê-se: "a probabilidade dessa garrafa conter líquido potável é 0,91". Na garrafa Y, "o grau de pertinência do conteúdo dessa garrafa em relação ao conjunto de líquido potável é 0,91". Das duas garrafas qual delas você escolheria para beber? O valor de pertinência significa que o conteúdo de Y tem um significado de 0,91 de similaridade com um líquido potável, por exemplo, pode ser, aguardente ou cerveja. O significado da probabilidade, no meio de garrafas observadas, 91 em 100 dispõe um líquido potável e as outras 9 em 100 podem conter um líquido mortífero. Prosseguindo com o conceito de observação, assuma que você pesquisou os conteúdos de X e Y, e descobre que os mesmos contêm ácido clorídrico e água de arroz, respectivamente. Depois da observação, o valor de pertinência de Y continua, enquanto que a probabilidade de afirmação de X vai para zero (BONVENTI JUNIOR, 2005).

2.5.1 Conjuntos Nebulosos

Os elementos contidos em conjuntos clássicos atendem propriedades *precisas*. Um exemplo é o conjunto dos números reais entre 9 e 10 é descrito por um conjunto

H= {x ∈ R| 9 ≤ x ≤10}, ou em termos de uma função de pertinência u_H (x):

$$u_H(x) = \begin{cases} 1, & se\ 9 \leq x \leq 10 \\ 0, & em\ outros\ casos \end{cases} \qquad (3)$$

Todo número real x está ou não em **H**. Pelo fato de que u_H (x) mapeia números reais em uma imagem de dois pontos {0, 1}, os conjuntos convencionais representam uma lógica *bivalorada:* 0 ou 1, sim ou não, aceso ou apagado.

Os conjuntos nebulosos de maneira distinta contêm elementos que dispõem atributos indefinidos de graduação variável, um exemplo disso seria, “aproximadamente 9,5”. A maneira comum de compreensão seria que o número 9 é mais próximo de 9,5 do que o número 11, com isto recomenda-se uma avaliação progressiva da pertinência par ao conjunto de números próximos ao 9,5. O grau de pertinência desse número em relação ao conjunto “aproximadamente 9,5”, é 1, e diminui simetricamente para números equidistantes acima ou abaixo. A função de pertinência referida neste caso como uma representação convexa, com máximo em 9 ,5 (BONVENTI JUNIOR, 2005).

$$u_H(x) \begin{cases} 1 - |x - 9{,}5|; & se\ 8{,}5 \leq \mathbf{x} \leq \mathbf{10{,}5} \\ \mathbf{0}; & em\ outros\ casos \end{cases} \qquad (4)$$

Pelo fato do conceito de nebulosidade estar agregado ao conjunto “aproximadamente 9,5”, não existe uma função de pertinência única. Fica em aberto a questão de decidir *quais* valores tornam esta pertinência nula: se a função deve decair abruptamente ou assintoticamente a partir de uma distância do ponto 9,5. É dado ao agente modelador de tal conjunto a autorização de decidir o formato da função, baseando-se nas propriedades essenciais a este conjunto nas circunstâncias em que está sendo desenvolvida (BONVENTI JUNIOR, 2005).

2.5.2 Relações de Operações na Teoria Nebulosa

Em (1965) foi definido por Zadeh as operações básicas a respeito de conjuntos nebulosos por meio de relações funcionais entre as respectivas funções de pertinência da seguinte forma: dados dois conjuntos nebulosos **A** e **B** estabelecidos pelas suas funções de pertinência $u_{\mathbf{A}}$ (x) e $u._{\mathbf{B}}$ (x), define-se as seguintes operações e relações:

Igualdade: $A = B \Leftrightarrow u_A(x) = u_B(x)$

Continência: $A \subset B \Leftrightarrow u_A(x) \leq u_B(x)$

Complemento: $A \Leftrightarrow u_{\neg A}(x) = 1 - u_A(x)$

Intersecção: $A \cap B \Leftrightarrow u_{A \cap B}(x) = \min(u_A(x), u_B(x))$

União: $A \cup B \Leftrightarrow u_{A \cup B}(x) = max(u_A(x) + u_B(x))$

Percebe-se que as operações de união e intersecção são ambas *associativas* e *comutativas,* ou seja, (A *B*) C = A (B *C*) e AB = BA,

Foram definidas também as seguintes *operações algébricas* por Zadeh (1965):

Produto algébrico: $AB\ u_{AB}(x) = u_A(x)\ u_B(x)$

Soma algébrica: $A+B\ u_{A+B}(x) = u_A(x) + u_B(x)$

É preciso ressaltar que tais operações são realizadas sobre o domínio {x $\in$ R}. Essas funções de pertinência recebem um suporte de um domínio

denominado *universo de discurso.* Em seguida foram propostas outras formas de combinar conjuntos nebulosos a respeito do universo de discurso, concedendo suporte a outras necessidades de modelagem (BLOCH, 1996).

2.5.3 Determinação das Funções de Pertinência

A seleção da função de pertinência e a regulagem dos seus parâmetros depende da natureza do problema. Foi exposta uma lista das formas de u (x) mais usadas na literatura, constantemente de acordo com o conhecimento referente ao conjunto nebuloso (Pedrycz, et al., 1998). O fato de existir várias formas de se usar funções de pertinência, as mesmas podem falhar ao expor a semântica de um modelo em particular (COX, 1994). Funções do tipo *triangular* e *trapezoidal* são exemplos de funções bastante utilizadas em engenharia de controle, devido sua facilidade de serem implementadas em *"Hardware".* Apesar disso, funções de pertinência derivadas de otimização iterativa, no caso dos algoritmos de *agrupamentos nebulosos*, apresentam um formato funcional não-analítico preciso que os conjuntos nebulosos refletem os padrões de pertinência relativos ao conceito a ser estipulado, nem todo momento funções analíticas se adaptam a esta necessidade. Diversos métodos de criação podem gerar funções de pertinência com estas características. Conforme Pedrycz e Gomide (1998), seis classes de métodos experimentais ajudam na determinação de funções de pertinência (BONVENTI JUNIOR, 2005):

1. *Método horizontal de estimação* da função de pertinência;
2. *Método vertical de estimação* da função de pertinência;
3. *Método da comparação par-a-par* de estimação da função de pertinência;
4. *Método do problema baseado em especificação* da função de pertinência;
5. *Otimização paramétrica* para estimação da pertinência;
6. *Agrupamentos nebulosos* para estimação da pertinência.

Neste trabalho, em que o espaço de atributos é o espaço de particionamento de objetos, a função analítica não representa bem a descrição

da função de pertinência. Portanto, o objetivo aqui é estimar a pertinência por agrupamentos nebulosos, no qual pode-se recalcular a cada avaliação dos objetos, obtendo-se um modelo mais preciso de particionamento.

2.5.4 Agrupamentos Nebulosos

Os métodos de agrupamento têm em vista praticamente o particionamento de conjuntos de elementos em grupos similares, em comparação a uma distância estabelecida. Dados com características similares entre si, têm que estar associados a um mesmo grupo, e os dados com dissimilaridades têm que estar associados a grupos diferentes. É um processo não-supervisionado de classificação de padrões, no qual sua organização é estabelecida por regras de dissimilaridade. Com insuficiência na informação prévia com relação a divisão dos elementos no meio de características as técnicas de agrupamento essencialmente são adequadas para estudar as ligações no meio dos elementos e sua organização (JAIN, et al 1999).

Neste ponto é importante distinguir-se a relação entre classe e grupo. Em um procedimento de agrupamento nebuloso dois ou mais grupos formados, podem ou não, se referir a mesma classe. O padrão da classe pode ser descontínuo no espaço de atributos, por isso muitas das vezes isto ocorre em função de como os mesmos foram selecionados, e em função do domínio em consideração. Por esse fato a importância dos operadores de agregação para oferecer consistência no classificador. Nas redes neurais existe um exemplo muito utilizado em suas discussões tipo *"perceptron"* é a classificação dada pela-tabela verdade *ou-exclusivo*, caso for resolvida por agrupamentos, gera quatro grupos (triviais) para duas classes. Assim sendo, este setor irá tratar como grupos a reunião de valores de atributos que estejam próximos entre si (BONVENTI JUNIOR, 2005). De forma exata Everitt (1980) apresenta algumas definições para "grupo":

1. Conjunto de entidades parecidas, e as entidades que estão em grupos diferentes não são parecidas;

2. Agrupamento no espaço de testes, de forma que a distância entre quaisquer dois pontos em um grupo é menor que aquela entre um ponto de um grupo e um ponto de outro grupo;
3. Região conectada em um espaço multidimensional, com uma alta densidade de pontos, separada de outras regiões de alta densidade por regiões de baixa densidade.

Em casos de técnicas de agrupamento binário as fronteiras entre os grupos têm uma boa definição, e cada elemento pertence ou não a um dado grupo (KLIR, 1995). O termo *binário* é empregado no sentido de que um elemento *pertence ou não* a um dos grupos. Constantemente a distribuição de dados se dá de forma que esta operação se torna demasiada arbitrária, e as fronteiras não ficam exatamente definidas (BONVENTI JUNIOR, 2005).

2.5.5 Algoritmos de Agrupamento Nebuloso

Na parte do processo de construção do classificador nebuloso, é fundamental que cada instância ou dado esteja relacionado a algumas classes com seus respectivos valores de pertinência.

Existem duas categorias de algoritmos para a execução de agrupamentos nebulosos, no qual podem ser considerados:

1. *Aprendizado competitivo*, originado de um modelo biológico sendo uma das espécies de redes neurais artificiais desenvolvidos na atualidade, empregando o modelo *"the winner takes all"*. Dentro deste grupo estão inseridos os *mapas auto-arganizáveis* de kohonen (SOM – *"self-organizing maps"*) e os chamados *"learning vector quantization"* LVQ, com suas respectivas versões para classificação nebulosa.
2. Fundados com base em *otimização de função custo*, caso dos algoritmos de *decomposição de mistura*, o algoritmo *"k-means"* e sua versão nebulosa, *"fuzzy c-means"* (FCM) os algoritmos *possibilísticos*. Nos algoritmos, *decomposição de mistura* a função de custo é produzida sobre vetores de atributos aleatórios e a atribuição dos grupos cumpre argumentos probabilísticos (classificação *bayesiana*). Nos agrupamentos

possibilísticos a função custo considera *quão típico* um vetor é em relação a um grupo, ignorando a restrição probabilística $\sum_{i=1}^{c} Pi(xk) = 1$, ao passo que, no tratamento nebuloso, são construídas as funções de pertinência no espaço de atributos tendo em conta a distribuição dos dados (BONVENTI JUNIOR, 2005).

O algoritmo k-means e sua versão nebulosa fuzzy c-means (FCM), serão os algoritmos bases que serão utilizados neste trabalho, para a etapa de aprendizado não-supervisionado do classificador proposto.

2.5.6 Fuzzy c-means

O algoritmo *"fuzzy c-means"* (FCM) é um algoritmo de cluster desenvolvido por Dunn (1973), e mais tarde melhorado por Bezdek (1981). É útil quando o número necessário de clusters é pré-determinado. Assim, o algoritmo tenta colocar cada um dos pontos de dados em um dos clusters. O que torna o FCM diferente é que ele não decide a adesão absoluta de um ponto de dados a um determinado cluster. Em vez disso, calcula a probabilidade (isto é, o grau de pertinência) que um ponto de dados pertencerá a esse cluster. Assim, dependendo da precisão do agrupamento que é exigido na prática, as medidas de tolerância apropriadas podem ser implementadas. Uma vez que a associação absoluta não é calculada, o FCM pode ser extremamente rápido porque o número de iterações necessárias para alcançar um exercício de agrupamento específico corresponde à precisão necessária. O algoritmo FCM é o algoritmo de agrupamento nebuloso mais aplicado com derivação a partir do algoritmo *"k-means"* (DUDA et al, 2000). Para o cálculo das distâncias entre os objetos, é empregada a medida de similaridade chamada distância euclidiana entre dois pontos no espaço de atributos escolhido. Sendo $X = \{x_1, x_2, ..., x_N\}$ um conjunto de N elementos explícitos por um vetor de atributos x_i, no qual é dividido em classes ou c grupos. Assume-se um particionamento aleatório inicial. Neste caso o algoritmo *"fuzzy c-means"* baseia-se em otimizações iterativas. Este modelo de otimização implica na definição de uma função que obedeça às duas premissas para uma ótima

partição. Tais premissas são: similaridade dos dados do grupo e separabilidade entre os grupos. Isto é, as distâncias d_{ij} em relação ao centro do grupo c_j e os dados x_i de cada grupo são minimizadas. Frequentemente é utilizada a distância euclidiana, se bem que, outras métricas também possam ser utilizadas. (KLIR et al, 1995).

O algoritmo FCM busca minimizar esta distância em relação entre todos os pontos do grupo, onde a função a ser minimizada é:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} \delta_{ij} \left\| x_i - c_j \right\|^2 \qquad (5)$$

Denota-se, N como o número de pontos de dados, C é o número de clusters necessários, c_j é o vetor central para o cluster j e $\delta = \delta_{ij}$ é a matriz de grau de pertinência para o ponto de dados x_i no cluster j. A norma, $d_{ij}=\|x_i-c_j\|$ mede a semelhança (ou proximidade) do ponto de dados x_i com o vetor central c_j do cluster j. Observe que, em cada iteração, o algoritmo mantém um vetor central para cada um dos clusters. Esses pontos de dados são calculados como a média ponderada dos pontos de dados, onde os pesos são dados pelos graus de pertinência.

Para um dado ponto de dados x_i, o grau de sua pertinência ao cluster j é calculado da seguinte forma:

$$\delta_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \qquad (6)$$

Onde m é o fator de nebulosidade e o vetor central c_j é calculado da seguinte forma:

$$c_j = \frac{\sum_{i=1}^{N} \delta_{ij}^{m} x_i}{\sum_{i=1}^{N} \delta_{ij}^{m}} \quad (7)$$

Na equação (7) acima, δ_{ij} é o valor do grau de pertinência calculado na iteração anterior. Nota-se que, no início do algoritmo, o grau de pertinência para o ponto de dados i para agrupar j é inicializado com um valor aleatório θ_{ij}, $0 \leq \theta_{ij} \leq 1$, de modo que $\sum_{j}^{C} \delta_{ij} = 1$,(soma das pertinências de um objeto qualquer é 1).

Nas equações (6) e (7), o fator de nebulosidade m, onde $1 < m < \infty$, mede a tolerância do agrupamento necessário. Esse valor determina o quanto os clusters podem se sobrepor uns com os outros. Quanto maior o valor de m, maior a sobreposição entre os clusters. Em outras palavras, quanto maior o coeficiente de nebulosidade que o algoritmo usa, um número maior de pontos de dados cairá dentro de uma banda fuzzy onde o grau de associação não é nem 0 nem 1, mas em algum lugar intermediário.

A precisão necessária do grau de pertinência determina o número de iterações concluídas pelo algoritmo FCM. Essa medida de precisão é calculada usando o grau de pertinência de uma iteração para a próxima, levando o maior desses valores em todos os pontos de dados considerando todos os clusters. Se representarmos a medida de precisão entre a iteração k e k+ 1 com ε, calculamos seu valor da seguinte maneira:

$$\varepsilon = \Delta_i^N \, \Delta_j^C \left| \delta_{ij}^{(k+1)} - \delta_{ij}^{k} \right| \quad (8)$$

Onde k = iT que é o número de iterações do algoritmo, δ_{ij}^{k} e $\delta_{j}^{(k+1)}$ são respectivamente o grau de pertinência na iteração k e k + 1 e o operador Δ, quando fornecido um vetor de valores, retorna o maior valor nesse vetor.

Entretanto é fundamental ressaltar que os resultados do agrupamento nebuloso dependem do número de grupos *c* escolhidos aleatoriamente e do fator de nebulosidade *m*.

Este algoritmo FCM possui um custo computacional muito alto: *O (iT.N.c),* onde iT é o número de iterações do algoritmo, *N* é o número de pontos no espaço de atributos e *c* é o número de grupos predeterminados. O algoritmo FCM iterativo é descrito a seguir.

FCM- Algoritmo 2

1. Fornecer o número de grupos desejados: $2 \leq c \leq c_{max}$;
2. escolher o nível de nebulosidade: $1 < m < \infty$;
3. criar uma partição inicial aleatória (ou uniforme, ou baseada em alguma heurística), com valores dos centros entre x_{max} e x_{min};
4. escolher um limite $\in$ para convergência
5. Repita:

- Calcular os centros c_j usando a equação da equação 7, com a partição inicial dada;
- atualizar as partições calculando novos valores de pertinência como segue:
- para *i* = 1 até c // Todos grupos
- para *j*= 1 até *N* // Todos os pontos

— calcular d_{ij} euclidiana ;

— se $d_{ij} < d_{min}$ então $\delta_{ij} = 1$; senão atualizar δ_{ij} conforme eq. 6;

6. Até que $\left|\delta_{ij}^{(k+1)} - \delta_{ij}^{k}\right| < \epsilon$

Para os algoritmos que minimizam a função custo emprega-se métodos de cálculo diferencial (BEZDEK, 1980) e precisa-se da informação do número de grupos pretendidos como parâmetro inicial. O maior problema está no fato de que podem convergir em um mínimo local, ocasionando resultados com agrupamentos não otimizados. Apesar disso não apresentam as desvantagens dos algoritmos de aprendizado competitivo, onde a ordem de apresentação dos dados não interessa. O algoritmo *"fuzzy c-means"* oferece o centroide ("valor médio") de cada classe, do mesmo modo aos algoritmos SOM e LVQ, oferece também matriz de pertinências $\delta = \{\delta_{ij}\}$, ($2 \leq i \leq c$, $2 \leq j \leq N$), com *N* dados de c classes (BONVENTI JUNIOR, 2005).

2.5.7 Medidas de Desempenho do Processo de Agrupamento

É importante definir no início o número de grupos para a partição do espaço de atributos. Sobre um processo (semi) automático de classificação, os sistemas nebulosos de reconhecimento de padrões precisam ser competentes para determinar qual o melhor número de classes a ser empregado, com base na qualidade do agrupamento derivado. No sentido de analisar isto, foram criadas várias medidas de qualidade, denominadas de *funções validação*. Hoppner (1999) e colaboradores relacionam as mais importantes e, dentre elas, a mais destacada conhecida como função de validação *S*, dada pela razão entre compactação e separação (XIE, et al, 1991), isto é, a razão entre a distância média dos objetos em relação aos seus centros correspondentes e a mínima distância entre os centros. A função *S* é minimizada quando os grupos são bem definidos, correspondendo a grupos mais compactos e mais separados. Assim sendo uma função que seja capaz de apontar qual a melhor separação, em função de quantos grupos existem, é mais apropriada neste contexto (BONVENTI JUNIOR, 2005).

Define-se a compactação como a função objetivo J já minimizada e dividida pelo número de elementos do espaço de atributos:

$$Comp = \frac{J}{N} = \frac{1}{N}\sum_{i=1}^{c}\sum_{i=1}^{N}\delta_{ij}^{m}\, d_{ij}^{2} \tag{9}$$

a menor distância ente os centros dos grupos é a separação:

$$Sep = \min_{ij} \left\| x_i - c_j \right\| \forall i \neq j \tag{10}$$

Resultando que

$$S = \frac{Comp}{Sep} = \frac{J}{N d_{\min}^{2}} \tag{11}$$

A maneira mais direta de empregar a função *S* para avaliar a qualidade do agrupamento obtido é com a aplicação do algoritmo *"fuzzy c-means"* iterativamente, para $c=2$ até c_{max} e definir o valor c^* que minimize o valor de *S*:

$$c^* = \underset{c}{\arg\min}\left(\min_{\delta} S(\delta, c)\right) \tag{12}$$

É utilizada a função de validação *S* para escolher a melhor partição realizada pelo algoritmo FCM.

É fundamental destacar que, neste modelo exposto, as posições dos centros dos grupos obtidos não são utilizadas, porém apenas os valores da matriz de pertinência δ que é aplicada na classificação dos dados. Para o cálculo dos centros foi utilizada a eq.7 não de forma inútil, mas devido ao fato de que o método utilizado para minimizar a função *J* atualiza alternadamente a matriz de pertinência δ e os centros $\mathbf{x}_{i.}$ (BONVENTI JUNIOR, 2005).

3. BASE DE DADOS E AMBIENTE COMPUTACIONAL

3.1 BASE DE DADOS IRIS

Problemas do tipo que abrangem classificação de padrões requerem um profundo aprendizado do problema. É preciso ter atenção na estrutura e porção dos dados, tempo de desenvolvimento, e dificuldade do problema, com interesse de escolher técnicas que exigem menor esforço computacional com resultados benéficos, gerando soluções ótimas para o problema em estudo.

Foi no repertório *UCI Machine Learning Repository* (BALKE et al., 1998) que se obteve a base de dados Íris. Ìris é um clássico problema de aprendizado de máquina, na qual deseja-se identificar que tipo de flor se tem com base em suas medições distintas.

A representação dos dados de sua planta é dada por 4 características, sendo o comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. A planta Íris, conhecida ainda como “Flor de Lis”, é composta por uma coleção de 150 amostras, e 4 caraterísticas como citado acima. Estas amostras classificam-se em três categorias: Íris Setosa, Íris Virgínica e Íris Versicolor.

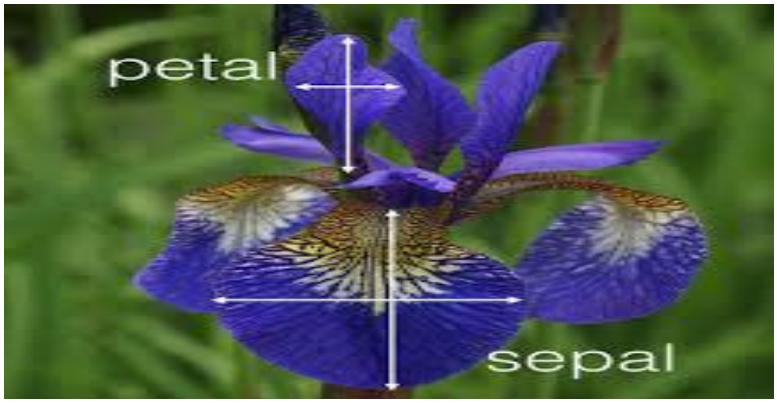

Figura 13 - Características da base de dados Íris
Fonte: Autor desconhecido

3.2 BASE DE DADOS DIAGNÓSTICA DE CÂNCER DE MAMA

Pretende-se a partir do banco de dados disponíveis no repositório da UCI (University of California – Irvine, UCI Machine Learning Repository) com informações a respeito de Câncer de Mama realizar o agrupamento de dados com a técnica de agrupamento fuzzy c-means, e verificar o diagnóstico a partir da análise dos dados, isto é, quantos dados correspondem em M = maligno e B = benigno.

Esta base de dados de câncer de mama foi apresentada na UCI a partir da Universidade de Wisconsin Hospitais, Madison, sendo que as informações foram concedidas pelo Dr. William H. Wolberg (WOLBERG et al 1990).

A base possui 569 instâncias (conjunto de dados), 32 atributos. Os valores dos atributos são calculados a partir da imagem digitalizada de uma amostra aspirada através de uma cânula ou agulha fina da massa de uma mama. Onde cada instância tem uma das 2 classes possíveis: benignos ou malignos, desta forma os valores da classe são binários. Estes atributos são elementos dos relatórios das anomalias nas amostras retiradas de um nódulo do seio para verificar se são malignos (possui câncer) ou benignos (não possui câncer).

Neste trabalho para o agrupamento de dados serão apenas utilizados 10 atributos de valor real demonstrados a seguir:

- **Raio** é a média das distâncias do centro para os pontos no perímetro.
- **Textura** é o desvio padrão de valores de escala de cinza.
- **Perímetro** responsável pela medida de contorno da superfície.
- **Área** responsável pela medida da superfície.
- **Suavidade** gravidade das porções côncavas do contorno da superfície
- **Compacidade** dado pelo (perímetro^2 / área -1, 0).
- **Concavidade** gravidade das porções côncavas do contorno.
- **Pontos côncavos** número de porções côncavas do contorno.
- **Simetria responsável** pela divisão das partes da superfície.
- **Dimensão fractal** ("aproximação do litoral" – 1).

3.3 AMBIENTE COMPUTACIONAL

O ambiente computacional utilizado para a implementação do k-means e o fuzzy c-means é a plataforma Matlab versão R2016a.

4. RESULTADOS E DISCUSSÃO

4.1 EXEMPLO NUMÉRICO FUZZY C -MEANS

Exemplo de aplicação do algoritmo “fuzzy c-means”. Na tabela 1, é feita a avaliação da velocidade e da resistência de 11 jogadores. Para um valor próximo de 1 significa que um jogador é bastante rápido e resistente conforme o caso, um valor próximo de 0 indica que o jogador é lento e não resistente. Neste exemplo o conjunto de dados é separado em 2 grupos (clusters) para observar se os jogadores possuem características especiais para o algoritmo fuzzy c-means. A quantidade de grupos é 2, o parâmetro m = 2 e dois centros são escolhidos inicialmente como c1= (0.2, 0.5) e c2= (0.8, 0.5). Por ser um exemplo numérico, apenas a primeira iteração será demonstrada.

Tabela 1 - Avaliação da velocidade e resistência de 11 jogadores

Jogador	Velocidade	Resistência
1	0.58	0.33
2	0.90	0.11
3	0.68	0.17
4	0.11	0.44
5	0.47	0.81
6	0.24	0.83
7	0.09	0.18
8	0.82	0.11
9	0.65	0.50
10	0.09	0.63
11	0.98	0.24

As funções de pertinência iniciais dos dois grupos são calculadas utilizando a equação 6.

$$\delta_{ij}(x_i) = \frac{\mathbf{1}}{\sum_{k=1}^{2}\left(\frac{\left\|x_i - c_j\right\|^2}{\left\|x_i - c_k\right\|^2}\right)^{2}}$$

Calculando a distância da primeira amostra x1= (0.58 0.33) para o primeiro centro c1= (0.2, 0.5).

$$\left\|x_1 - c_1\right\|^2 = 0.38^2 + 0.17^2 = 0.1733$$

Calculando a distância da primeira amostra x1 = (0.58 0.33) para o segundo centro c2= (0.8, 0.5).

$$||x_1 - c_2||^2 = 0.22^2 + 0.17^2 = 0.0773$$

Calculando a pertinência da primeira amostra para o primeiro centro.

$$\delta_{11}(x1) = \frac{1}{\frac{0.1733}{0.1733} + \frac{0.1733}{0.0773}} = 0.3085$$

Calculando a pertinência da primeira amostra para o segundo centro.

$$\delta_{12}(x1) = \frac{1}{\frac{0.0773}{0.0773} + \frac{0.0773}{0.1733}} = 0.\ 6915$$

Calculando a distância da segunda amostra x2= (0.90 0.11) para o primeiro centro c1= (0.2, 0.5).

$$||x_2 - c_1||^2 = 0.7^2 + 0.39^2 = 0.6421$$

Calculando a distância da segunda amostra x2 = (0.90 0.33) para o segundo centro c2= (0.8, 0.5).

$$||x_2 - c_2||^2 = 0.1^2 + 0.39^2 = 0.1621$$

Calculando a pertinência da segunda amostra para o primeiro centro

$$\delta_{21}(x2) = \frac{1}{\frac{0.6421}{0.6421} + \frac{0.6421}{0.1621}} = 0.2016$$

Calculando a pertinência da segunda amostra para o segundo centro.

$$\delta_{22}(x2) = \frac{1}{\frac{0.1621}{0.1621} + \frac{0.1621}{0.6421}} = 0.7984$$

De forma similar são obtidos valores das demais funções de pertinência. Os resultados são apresentados na tabela 2.

Tabela 2 - Grau de pertinência de cada elemento em cada grupo

Dados	Grau de Pertinência	
	Cluster1	Cluster2
1	0.3085	0.6915
2	0.2016	0.7984
3	0.2665	0.7335
4	0.9762	0.0238
5	0.5481	0.4519
6	0.7927	0.2073
7	0.8412	0.1588
8	0.2213	0.7787
9	0.1000	0.9000
10	0.9473	0.0527
11	0.1289	0.8711

O grau de pertinência de valor 1 indica a máxima pertinência, enquanto que um valor de 0 significa que o dado não pertence no grupo. Sendo assim os dados com maior pertinência ao grupo1 são o quarto, quinto, sexto sétimo e décimo, enquanto que os restantes dos dados têm maior pertinência ao grupo 2.

É necessário fazer a atualização dos centros depois da obtenção dos graus de pertinência a partir da equação 7.

Calculando o novo valor do primeiro centro

$$c_1 = \frac{\sum_{i=1}^{11}(\delta_{c1})^2 . x_{11}}{\sum_{i=1}^{11}(\delta_{c1})^2} = \frac{0.7443}{3.7692} = 0.1975$$

$$c_1 = \frac{\sum_{i=1}^{11}(\delta_{c1})^2 . x_{12}}{\sum_{i=1}^{11}(\delta_{c1})^2} = \frac{1.9392}{3.7692} = 0.5144$$

Calculando o novo valor do segundo centro

$$c_2 = \frac{\sum_{i=1}^{11}(\delta_{c2})^2 . x_{21}}{\sum_{i=1}^{11}(\delta_{c2})^2} = \frac{3.0931}{4.1046} = 0.7536$$

$$c_2 = \frac{\sum_{i=1}^{11}(\delta_{c2})^2 . x_{22}}{\sum_{i=1}^{11}(\delta_{c2})^2} = \frac{1.1808}{4.1046} = 0.2877$$

Esse processo é então repetido até convergência, conforme o algoritmo 2 da seção 2.5.6.

4.2. BASE DE DADOS ÍRIS E PRÉ-PROCESSAMENTO PCA

Com a técnica compreendida espera-se dar clareza como uma técnica de agrupamento poder ser bastante importante para soluções de vários problemas, gerando resultados eficientes e menos custosos.

A Figura 14 apresenta a base de dados Íris com as suas 150 amostras. Essa representação utiliza apenas 2 dimensões dos dados (largura e comprimento da sépala), visto que originalmente essa base possui 4 dimensões, o que tornaria inviável plotar e visualizar tais dados. Esta apresentação da base foi realizada com o intuito de mostrar apenas os 150 objetos contidos da base através do carregamento da base íris_dataset, onde até então não é aplicada nenhuma técnica de redução de dimensionalidade e nem qualquer agrupamento de dados realizados.

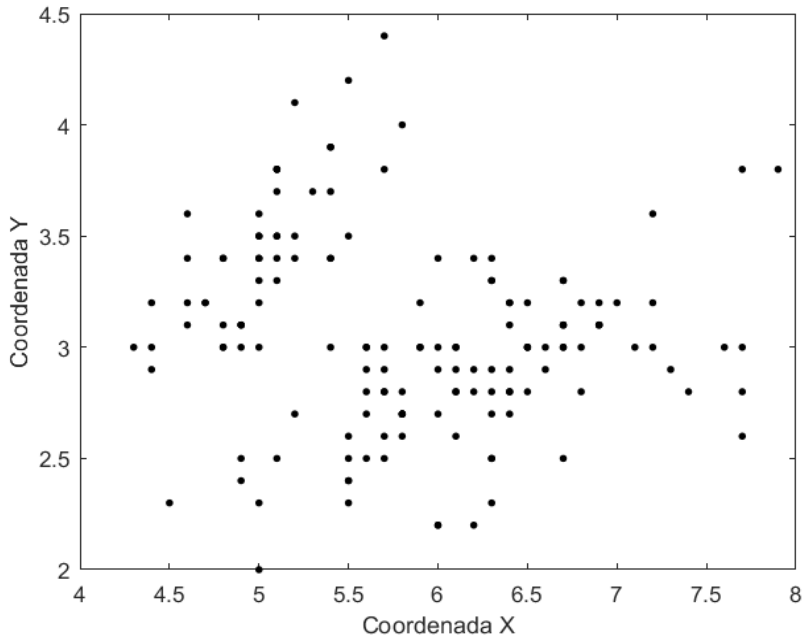

Figura 14 – Representação das 150 amostras da base Íris

Na Figura 15 abaixo é exibida a base de dados Íris com a aplicação da técnica denominada análise de componentes principais (PCA). O PCA foi

empregado sobre a base Íris para a redução de dimensionalidade dos seus dados visando uma fácil visualização dos mesmos. Diferentemente da Figura 14, onde apenas 2 dimensões foram escolhidas ao acaso e plotadas, na Figura 15 as duas dimensões visualizadas representam uma combinação linear das variáveis originais.

PCA pode ser feito por decomposição em autovalores (valores próprios) de uma matriz covariância, geralmente depois de centralizar a matriz de dados para cada atributo. Ou seja, normalizou-se a base de dados utilizando método z-score em cada dimensão do conjunto de dados, calculou-se a matriz covariância, encontrou-se os autovalores e autovetores, ordenou-se os autovetores resultantes com base nos autovalores obtidos e por fim projetou-se o conjunto de dados originais transformados pela rotação dada por essa matriz resultante. O trecho de código responsável por essa transformação pode ser visualizado na no algoritmo 3. A variável X recebida como parâmetro são as variáveis originais. A variável V são os autovalores ordenados de forma crescente, a variável PC são os autovetores ordenados segundo a mesma ordem de V e a variável de retorno signal são os dados transformados.

PCA - Algoritmo 3

```
function [PC V signals] = runPCA(X)
[N,P] = size(X);
% normalizaçao
mn = mean(X,1);
X = X - repmat(mn,N,1);
% cálculo da matriz de covariância
covariance = 1 / (N-1) * (X') * X;
% calculando autovalores e autovetores
[PC, V] = eig(covariance);
% separando a diagonal principal
V = diag(V);
% ordenando V e PC Segundo ordem decrescente de V
[junk, rindices] = sort(-1*V);
V = V(rindices);
PC = PC(:,rindices);
%% projetando (rotacionando) as variáveis originais pelos autovetores PV. Apenas 2 componentes principais mantidas nesse caso.
signals = X * PC(:,2);
```

Portanto, como a dimensão original da base Íris é de 4 dimensões, com a técnica PCA aplicada chegou-se para duas dimensões através de combinações lineares das variáveis originais obtendo assim uma melhor visualização dos dados.

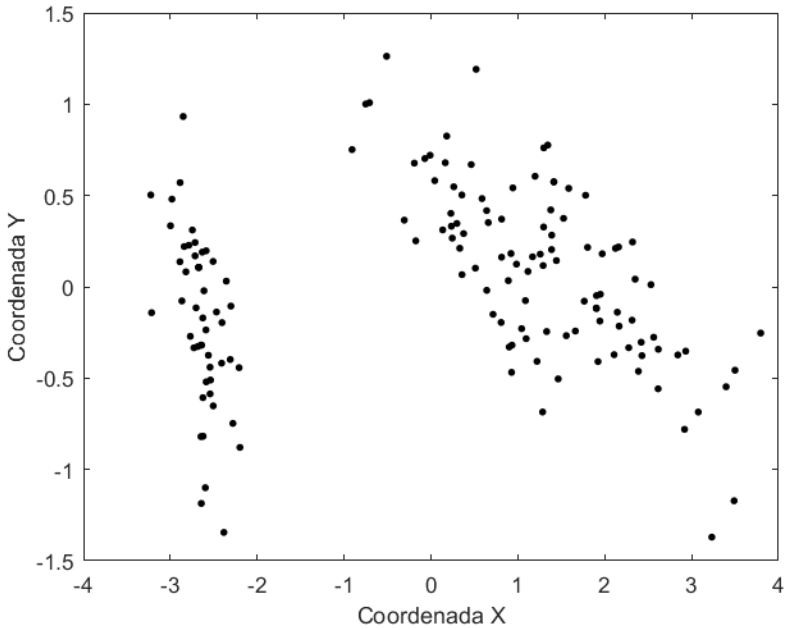

Figura 15 – Base de dados Íris com PCA

4.3 RESULTADO DO MÉTODO K–MEANS

Para ilustrar os resultados fornecidos pelo k-means implementado através do sistema iterativo Matlab, será considerado o conjunto de dados da Base Íris que possui 150 amostras com a técnica de redução de dimensionalidade (PCA) já aplicada.

Abaixo é apresentado os resultados do método k-means, com a definição final dos centros, o número de iterações, soma de suas distâncias bem como a soma total de melhor distância dentro de cada grupo. Como a medida de dissimilaridade é a distância euclidiana, é possível verificar a dispersão dos elementos em relação ao seu centro calculando-se a soma do quadrado das distâncias entre os pontos de um grupo em relação ao respectivo centro. Por exemplo, as observações do grupo de cor verde e vermelho apresentam uma dispersão maior no seu agrupamento, portanto nota-se

claramente que existem elementos que ficaram bem distantes em relação ao seu centro, enquanto que no grupo de cor azul a dispersão dos elementos no agrupamento é menor, o que pode ser observado através da Figura 16.

Para gerar os dados que seguem e o resultado da Figura 16, a função de implementação requer conjunto de dados, o número de centros, no qual para este trabalho foram definidos três centros definidos inicialmente de forma aleatória (k=3), que podem ser vistos na Tabela 2, o tipo de distância no caso foi escolhido a distância Euclidiana e o número de vezes que o k-means será executado que pode ser visto na Tabela 3.

Tabela 3 – Centros dos três grupos

Centros	Coordenada eixo X	Coordenada eixo Y
C1	2.3465	-0.2724
C2	-2.6408	-0.1905
C3	0.6644	0.3303

Tabela 4 - Números de iterações e soma das distâncias

Replicar	Iterações	Soma total das distâncias
1	9	63.8738
2	7	63.8738
3	6	128.275
4	9	63.8738
5	3	130.134
6	3	63.8738
7	5	63.8738
Melhor soma total das distancias		**63.8738**

É necessário rodar o k-means N vezes pois, como os centros iniciais são escolhidos aleatoriamente, diferentes resultados podem ser obtidos. Em duas das sete repetições o k-means encontrou um mínimo local. Uma vez que cada uma dessas sete repetições começa de centroides inicialmente selecionados de forma aleatória, isso faz com que o k-means as vezes encontre mais de um mínimo local. Portanto, a solução que o k-means retorna é aquela com a menor soma total de distâncias em relação aos centróides, em todas as iterações, neste caso a soma total de melhor distância é: **63.8738.**

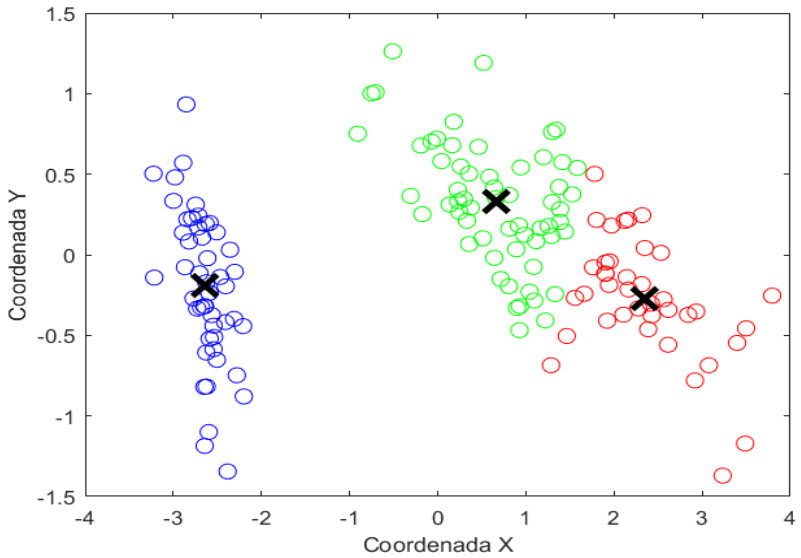

Figura 16 – Agrupamento de dados com o algoritmo k-means

4.4 RESULTADO DO MÉTODO FUZZY C-MEANS

Para fazer o agrupamento dos elementos da Figura 17 é utilizado o algoritmo FCM Esta função exige que se tenha um conjunto de dados de entradas. Novamente neste estudo, foi utilizado os dados provenientes da base de dados Íris pós redução de dimensionalidade com o método PCA.

Utilizou-se a distância euclidiana como medida de distância. Novamente foram definidos três centros onde são atribuídos os valores de forma aleatória com c=3 (número de clusters), com base nesses valores, associa-se cada elemento ao valor no qual possui menor distância, formando os 3 grupos. Em seguida, é calculado o centro de cada grupo formado, e os elementos são re-associados ao centro que lhe é mais próximo. Desta forma, os cálculos continuam, iterativamente, até que as diferenças entre os centros do passo anterior e do atual sejam mínimas. No FCM cada elemento pertence a todos os grupos, porém com graus de pertinência diferentes, sendo que o elemento é designado no grupo onde o seu grau de pertinência é maior.

A finalidade das funções de pertinência do FCM é fundamental para lidar com as regiões de incerteza durante o agrupamento dos dados. Nas

observações da Figura 17 os elementos do grupo de cor verde alguns deles estão associados ao grupo de cor vermelha e alguns elementos do grupo de cor vermelha estão associados ao grupo de cor verde. Esta dispersão se dá pelo fato de que um elemento pode pertencer a mais de um grupo consoante o seu maior grau de pertinência.

A Tabela 5 apresenta o valor dos centros finais, a Tabela 6 apresenta o valor da função objetivo, essa função atribui a cada partição um valor de que dever ser otimizado, assim obtém a melhor avaliação após cada uma das interações do método.

Tabela 5 – valor dos centros finais

Centros	Coordenada Eixo X	Coordenada Eixo Y
C1	-2.6216	-0.1744
C2	2.22884	-0.2266
C3	0.6362	0.3108

Tabela 6 – Número de iterações

Contagem de iterações	**Função objetivo**
1	277.527035
2	216.562385
3	183.757655
4	113.569074
5	84.960791
6	65.795676
7	56.158152
8	52.735547
9	51.086888
10	50.238041
11	49.805439
12	49.592261
13	49.490813
14	49.443887
15	49.422626
16	49.413132
17	49.408934

18	49.407090
19	49.406283
20	49.405931
21	49.405778
22	49.405711
23	49.405682
24	49.405670
25	49.405664

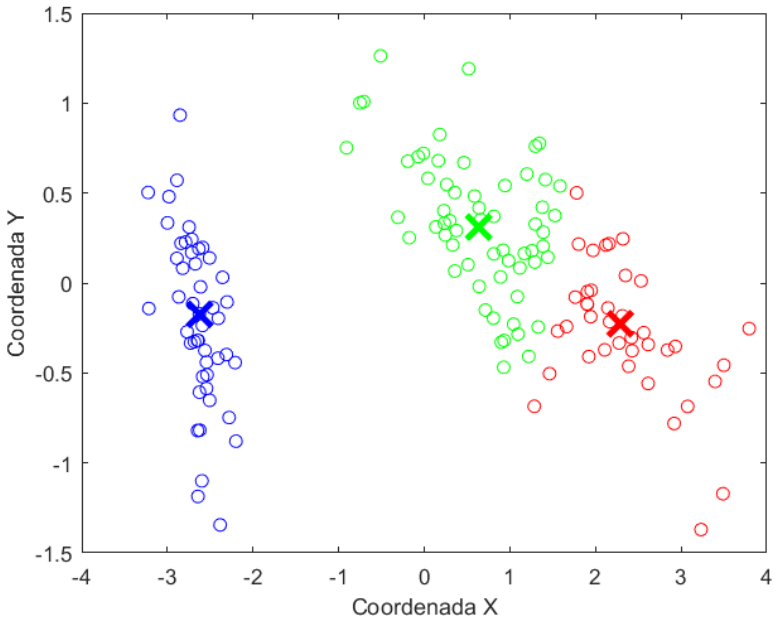

Figura 17 - Agrupamento de dados com o algoritmo fuzzy c-means

4.5 COMPARAÇÃO DOS MÉTODS K-MEANS E FUZZY C-MEANS

Este trabalho descreveu em detalhes duas técnicas de agrupamento que podem ser utilizadas para fazer agrupamento não supervisionado de dados em um conjunto de dados: k-means e fuzzy c-means. Foi mostrado que o funcionamento dos dois algoritmos é semelhante, no entanto, atende-se ao fato de que o k-means consiste em uma aproximação ***hard,*** isto é, o algoritmo do k-means tem que executar várias operações aritméticas enquanto que o fuzzy c-means tem uma aproximação ***soft*** isto é, simplifica essas operações aritméticas e aborda as operações com 0 ou 1. Observa-se, no entanto, que os resultados foram bastante semelhantes visualmente. Todavia se investigados cada ponto em detalhe percebemos que alguns pontos pertencem a certo grupo, mas também pertencem parcialmente a outro grupo. A Tabela 7 apresenta alguns desses pontos e sua pertinência para cada um dos 3 grupos. Por exemplo o ponto 10 tem chance absoluta de pertencer unicamente ao grupo 3 pelo fato de possuir 98% de grau de pertinência para este grupo. Para o ponto 52 é muito difícil garantir de forma absoluta de que o ponto pertença a um dos três grupos, o grau de pertinência apresentado pelo ponto permite dizer que o ponto 52 pertença aos dois primeiros grupos.

Tabela 7 – Grau de pertinência dos pontos para cada um dos três grupos.

Pontos	δ_{11}	δ_{21}	δ_{31}
1	0.0024	0.0011	0.9965
2	0.0111	0.0050	0.9839
3	0.0133	0.0062	0.9805
4	0.0213	0.0095	0.9692
5	0.0031	0.0015	0.9954
6	0.0434	0.0197	0.9369
7	0.0087	0.0040	0.9873
8	0.0000	0.0000	1.0000
9	0.0467	0.0213	0.9320
10	0.0074	0.0033	0.9894
11	0.0216	0.0101	0.9683
12	0.0022	0.0010	0.9968
13	0.0158	0.0071	0.9771
14	0.0509	0.0246	0.9244
15	0.0704	0.0362	0.8934
16	0.1017	0.0523	0.8460
17	0.0331	0.0161	0.9508
18	0.0019	0.0009	0.9972
20	0.0647	0.0297	0.9056
21	0.0107	0.0050	0.9843

22	0.0156	0.0068	0.9776
23	0.0071	0.0032	0.9896
24	0.0227	0.0113	0.9660
25	0.0118	0.0049	0.9832
26	0.0123	0.0051	0.9826
27	0.0110	0.0047	0.9842
28	0.0025	0.0011	0.9964
29	0.0041	0.0018	0.9941
30	0.0019	0.0009	0.9972
31	0.0122	0.0054	0.9824
32	0.0131	0.0057	0.9812
33	0.0105	0.0046	0.9849
34	0.0329	0.0161	0.9510
35	0.0624	0.0315	0.9060
36	0.0074	0.0033	0.9894
37	0.0056	0.0026	0.9918
38	0.0159	0.0075	0.9766
39	0.0074	0.0033	0.9894
40	0.0401	0.0186	0.9413
41	0.0004	0.0002	0.9994
42	0.0026	0.0012	0.9962
43	0.0887	0.0401	0.8711
44	0.0290	0.0136	0.9574
45	0.0049	0.0021	0.9930
46	0.0269	0.0115	0.9616
47	0.0159	0.0071	0.9770
48	0.0110	0.0050	0.9840
49	0.0165	0.0075	0.9760
50	0.0157	0.0073	0.9769
51	0.0009	0.0004	0.9987
52	0.4442	0.5154	0.0404
53	0.7688	0.2017	0.0295
54	0.3492	0.6228	0.0280
55	0.8784	0.0748	0.0469
56	0.7899	0.1899	0.0202
57	0.9954	0.0036	0.0010
58	0.6962	0.2753	0.0284
59	0.5828	0.1303	0.2869
60	0.7525	0.2219	0.0256
61	0.8535	0.0809	0.0656
62	0.6403	0.1415	0.2182
63	0.9765	0.0177	0.0058
64	0.9400	0.0392	0.0208
65	0.9114	0.0778	0.0108
66	0.8249	0.0866	0.0885
67	0.7059	0.2556	0.0385
68	0.9990	0.0007	0.0002
69	0.9479	0.0339	0.0182
70	0.9325	0.0571	0.0104
71	0.8849	0.0660	0.0490
72	0.8253	0.1581	0.0166
73	0.9510	0.0343	0.0147
74	0.7322	0.2474	0.0205
75	0.9475	0.0452	0.0073
76	0.9028	0.0795	0.0177
77	0.7777	0.1923	0.0300

78	0.5226	0.4510	0.0265
79	0.3054	0.6739	0.0207
80	0.9735	0.0222	0.0043
81	0.7787	0.0981	0.1232
82	0.8395	0.0853	0.0753
83	0.8055	0.0948	0.0997
84	0.9233	0.0471	0.0296
85	0.6728	0.3040	0.0232
86	0.9877	0.0093	0.0030
87	0.8664	0.1126	0.0210
89	0.5590	0.4087	0.0324
90	0.9837	0.0135	0.0028
91	0.9492	0.0333	0.0175
92	0.9013	0.0606	0.0381
93	0.9489	0.0364	0.0147
94	0.9249	0.0646	0.0106
95	0.9450	0.0355	0.0195
96	0.5986	0.1306	0.2708
97	0.9620	0.0261	0.0120
98	0.9641	0.0247	0.0112
99	0.9757	0.0170	0.0073
100	0.9526	0.0376	0.0097
101	0.5232	0.1208	0.3560
102	0.9614	0.0260	0.0127
103	0.0304	0.9654	0.0042
103	0.6581	0.3156	0.0264
104	0.0269	0.9688	0.0043
105	0.1279	0.8612	0.0108
105	0.0245	0.9725	0.0030
106	0.1332	0.8362	0.0306
107	0.8192	0.1257	0.0551
108	0.0691	0.9181	0.0128
109	0.0725	0.9191	0.0084
110	0.0971	0.8830	0.0200
111	0.2207	0.7630	0.0163
112	0.2361	0.7475	0.0164
113	0.0058	0.9935	0.0007
114	0.7032	0.2664	0.0304
115	0.5171	0.4558	0.0271
116	0.0808	0.9122	0.0071
117	0.0744	0.9190	0.0066
118	0.1755	0.7772	0.0473
119	0.1728	0.7840	0.0432
120	0.7314	0.2397	0.0289
121	0.0113	0.9871	0.0016
122	0.8062	0.1725	0.0214
123	0.1425	0.8242	0.0333
124	0.6198	0.3582	0.0220
125	0.0037	0.9958	0.0005
126	0.0439	0.9486	0.0074
127	0.7378	0.2426	0.0196
128	0.6902	0.2889	0.0209
129	0.0889	0.9024	0.0087
130	0.0175	0.9799	0.0025
131	0.0572	0.9326	0.0102
132	0.1780	0.7743	0.0477

133	0.0838	0.9077	0.0085
134	0.5441	0.4337	0.0222
135	0.3609	0.6147	0.0245
136	0.1046	0.8732	0.0222
137	0.0113	0.9875	0.0012
138	0.0926	0.8995	0.0079
139	0.8078	0.1752	0.0170
140	0.0200	0.9777	0.0023
141	0.0008	0.9990	0.0001
142	0.0710	0.9216	0.0074
143	0.6581	0.3156	0.0264
144	0.0188	0.9784	0.0028
145	0.0064	0.9927	0.0009
146	0.0575	0.9371	0.0054
147	0.5297	0.4463	0.0241
148	0.1704	0.8170	0.0126
149	0.0826	0.9102	0.0072
150	0.6380	0.3398	0.0222

4.6 RESULTADOS PARA BASE DE DADOS DE CÂNCER

A seguir é demonstrado o resultado do teste feito a partir da base de dados de diagnóstico de câncer de mama. Comparando o resultado do agrupamento observa-se na Figura 18 que alguns pontos do grupo de cor azul ficaram bastante distante de seu cluster, isso se dá por causa do conjunto de dados de entrada que é bastante diversificado, isto é, quanto mais diversificado for o conjunto mais afastados dos seus clusters podem ficar seus pontos.

Observando a Tabela 8 no Anexo A é possível verificar se um elemento é maligno ou benigno através de seu grau de pertinência. Na coluna do valor previsto são atribuídos para os elementos com câncer maligno (M) o valor correspondente 2 e para os elementos com câncer benigno (B) foi atribuído o valor correspondente 1. Essa atribuição é realizada pela comparação do grau de pertinência entre o valor do grupo δ11 ou δ21. Dessa forma se o valor do elemento para o grupo δ21 for maior que o valor do elemento do δ11 esse elemento passa a ser considerado como maligno (M = 2) e caso o valor do elemento para o grupo δ11 for maior em relação ao grupo δ21 esse elemento é considerado benigno (B = 1). A coluna do valor real é a coluna que diz respeito ao valor real atribuído para cada elemento na base de dados para fazer o diagnóstico desses elementos como maligno (M) ou benigno (B) tal valor pode ser denominado como o **que é**. A partir da análise destas duas colunas, com

resultado expresso na coluna “Acertos” pode-se verificar se o agrupamento conseguiu diferenciar pessoas com cânceres malignos ou benignos.

Todavia podemos observar alguns casos interessantes. As amostras 11 e 35, por exemplo, ambas foram agrupadas no grupo de benignas. Todavia seu grau de pertinência demonstra dúvida (valores próximos a apenas 50% de certeza). Essas amostras seriam boas candidatas à uma observação mais cuidadosa, com exames mais aprofundados, por exemplo. Já as amostras 6 e 9 foram classificadas com alto grau de confiança, dispensando futuras investigações.

No total o número de acertos foi 438, o que corresponde a 76% dos casos de teste. Embora melhores números possam ser alcançados com outras técnicas de agrupamento, classificação e redução de dimensionalidade, tal número aparenta ser satisfatório para um primeiro estudo, especialmente se considerarmos o grau de confiança como fator para afirmar com plena certeza que o câncer é maligno ou benigno.

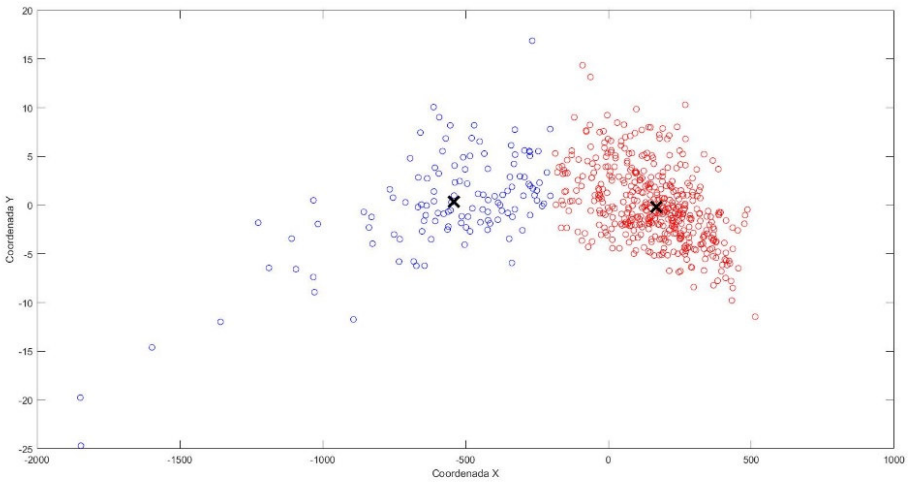

Figura 18 – Diagnóstico câncer de mama com fuzzy c-means

5. CONCLUSÃO

Neste trabalho foi proposto um conceito de conjuntos nebulosos para análise de dados. No cotidiano pessoas pretendem demonstrar e guardar enormes quantidades de dados com a finalidade de analisar e usar esses dados. Para atender essa quantidade de dados existe dois mecanismos fundamentais para lidar com esse volume de dados que são: classificação de dados ou de Clustering (agrupamento), ou seja, pode-se formar grupos ou classificar os mesmos em várias categorias.

Este conceito inclui o algoritmo fuzzy c-means que permite associar um objeto com todos os clusters utilizando uma função de pertinência (Zadeh, 1965). O conceito de conjuntos fuzzy proporciona a vantagem de expressar este tipo de contexto em que um objeto compartilha similaridade com vários grupos (clusters) através da possibilidade do algoritmo associar cada objeto parcialmente a todos os grupos.

Ao longo deste trabalho foram analisados algoritmos de agrupamento (clusterering) não supervisionado. Clustering serve para fazer agrupamento de dados segundo seus graus de semelhança, com aplicações em várias áreas como em mineração de dados e reconhecimento de padrões.

Este trabalho reportou a proposição de dois algoritmos o fuzzy c-means e o k-means. Embora os resultados dos agrupamentos sejam bastante semelhantes, é necessário ressaltar que o fuzzy c-means tem como princípio o conceito de conjuntos nebulosos, que tem seus objetos representados com tratamento de atributos definidos, mas com classificação vaga. Uma vez que suas classes são vagas (conjuntos nebulosos), permite que um objeto pertença a mais de um grupo segundo seu grau de semelhança, ao passo que o k-means seus dados são agrupados consoante suas caraterísticas, isto é, um objeto só pode pertencer a um único grupo consoante seu grau de semelhança.

Pode-se compreender, ao realizar os experimentos, que o método apresentado neste trabalho apresenta uma convergência satisfatória para as soluções ótimas das instâncias testadas.

REFERÊNCIAS

ADRIAAANS, P.; Zantinge. D., **Data Mining**, Addison Wesley, Harlew, (1996).

ANDRADE FILHO, José Augusto. **Definição automática da quantidade de atributos selecionados em tarefas de agrupamentos de dados**. 2013. 73 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação – ICMC, Universidade de São Paulo – USP, São Carlos, 2013.

BALL, G.H. and Hall, D.J. **ISODATA, a Novel Method of Data Analysis and Pattern Classification**. Stanford Research Institute, Menlo Park (1965).

BALKE, C. L.; MERZ, C.J, UCI **Repository of Machine Learning Databases**. Irvine. CA. USA: University of California, 1998.

BEZDEK, J. C.: "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York, 1981.

BEZDEK, J. C.; PAL, S. K. **Fuzzy models for pattern recognition**. New York: IEEE
Press, 1992.

BEZDEK, J C. Pattern Recognition **Fuzzy Objective Function Algortithms**, 1 ed. Kluwer Academic Publishers, 1981.

BEZDEK, J. C. **A convergence theorem for the fuzzy isodata clustering algorithms.IEEE** Transactions on Pattern Analysis and Machine Intelligence, v. 2, p. 1–8, 1980.

BLOCH, I**. Information combination operators for data fusion: a comparative review with classification**. IEEE Trans on Systems, Man and Cybernetics - part A: Systems and Humans, v. 26, n. 1, jan 1996.

BONVETI, W. Jr. **Aprendizado nebuloso híbrido e incremental para classificar pixels por cores.** Tese (Doutorado). Universidade de São Paulo, São Paulo SP Brasil, 2005.

BRADLEY, P. s., Fayyad, U. M., **Refining initial points for K-means clustering**, in 15th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA pp. 91-99, 1998.

CÔRTES, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de Dados – Funcionalidades e Técnicas e Abordagens. 2002.** PUC. Disponível em **<**http://www.dbd.puc-rio.br/depto_informatica/02_10_cortes.pdf**>.** Acesso em: 3 jun. 2017.

COX, E. **The fuzzy systems handbook**. Cambridge MA USA: Academic Press, 1994.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classificatio**n. 2nd. ed. New York:

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

DRAPER, N. R., Smith, H., **Applied Regression Analysis**, John Wiley, New York 19981.

DRAPER, N. R.; Smith, H., **Applied Regression Analysis**, John Wiley, New York 1981.

EVERITT, B. **Cluster Analysis**. New York, USA: Academic Press, 1980.

FAYYAD, U., Piatsky-Shapiro, G., Smyth, P., **'From data mining to knowledge discovery in databases'**, AI Magazine 17(3), 37-54, (1996).

FRAWLEY, W., Piatetsky-Shapiro, G., Matheus, C. (1992), **'Knowledge discovery databases** - anoverview' Repartined in AI Magazine pp. 1-30.

GUSTAFSON, D. E.; KESSEL, W. C. **Fuzzy clustering with a fuzzy covariance matrix**.In: Proc IEEE CDC. San Diego CA: [s.n.], 1979. p. 761–766.

HAND, D. J., **Discrimination and Classification**, Wiley, Chichester 1981.

HAN, J. & Kamber, M., **Data mining: concepts an thechniques***,* Academic Press, London,2002.

HOPPNER, F. et al. **Fuzzy cluster analysis: methods for classification, data analysis and image recognition**. Chichester, England: John Wiley & Sons Ltd., 1999.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. **Data clustering**: a review. ACM Computing Surveys, v. 31, n. 3, sep 1999.

JAIN, A. K., Dubes, R. **Algoritms for Clustering** Data Prentice Hall, New JERSEY (1988).

LOUREIRO, Julina de Aguiar. **Técnicas de agrupamento de dados na mineração de dados químicos.** 122f. Tese (Mestrado) – Ciências da Computação, Universidade Federal de Pernambuco, (2005).

KLIR, G. J.; YUAN, B. **Fuzzy sets and fuzzy logic**. New Jersey USA: Prentice Hall, 1995.

MARQUES, J.M.. **O método da análise de componentes principais na detecção e identificação de outliers múltiplos em fototriangulação.** Tese (Pós-Graduação). Universidade Federal do Paraná, Curitiba PR Brasil, 1994.

MCQUEN, J**'Some methods for classification and analysts of multivarte observations'**, Berkeley Symposium on mathematics 1, S.281-298 (1967).

PEARSON, K. «**On Lines and Planes of Closest Fit to Systems of Points in Space**» (PDF). Philosophical Magazine. 2 (6): 559–572, 1901.

PEDRYCZ, W.; GOMIDE, F. **An introduction to fuzzy sets: analysis and design**. Cambridge MA USA: The MIT Press, 1998.

Ricardo LINDEN. **Técnicas de Agrupamento**. Revista de Sistemas de Informação da FSMA. n. 4. pp. 18-36. 2009.

ROCHA, Henrique Santos C., **Fuzzy C-Means na Seleção de Currículos.** Disponível em: <https://pt.slideshare.net/hscr/fuzzy-cmeans-na-seleo-de-currculos-projeto-de-aplicao>. Acesso em: 31 Mar. 2017.

ROCHA, Thiago; PERES, Serajane P.; BÍSCARO, Helton H.; MADEO Renata C. B., BOSCARIOLI, Clodis. Tutorial sobre **Fuzzy-c-Means e Fuzzy Learning Vector Quantization: Abordagens Híbridas para Terefas de Agrupamento e Classificação**. Revista de Informática Teórica e Aplicada v. 19, n.1, p. 120-163, 2012.

SIMOVIC, D. A. **Data mining algorithms I: Clustering**. Handbook of Applied Algorithms- Capitulo 6, p. 177-218, 2007.

TRONCO, Tania Regina. **Algoritmo de Agrupamento Fuzzy C-Means para Aprendizado e Tomada de Decisão em Redes de ópticas de Próxima Geração.** 2015. 137f. Tese (Doutorado) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2015.

University of California – Irvine, UCI Machine Learning Repository. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>. Acesso em 20 de novembro de 2017.

WOLBERG, W. H. and O.L. Mangasarian: "**Multisurface method of pattern separation for medical diagnosis applied to breast cytology**", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, pp 9193-9196, Dez. 1990.

XIE, X. L.; BENI, G. A **validity measure for fuzzy clustering. IEEE Transactions on Pattern** Analysis and Machine Intelligence, v. 13, p. 841–847, 1991.

ZADEH, L. A. **Fuzzy logic. neural networks and soft computing.** Comunications of the ACM. ACM Press. vol 37 no. 3. pp. 77-84. New York. USA. March-1994

ZADEH, L. A. **Fuzzy sets. Information Control**, v. 8, p. 338–353, 1965.

ANEXO A – Tabela com os diagnósticos de câncer de mama

Tabela 8 – Grau de pertinência dos pontos para cada um dos dois grupos.

Ponto	δ_{11}	δ_{21}	Valor previsto	Valor real	Acertos
1	0.1242	0.8758	2	2	1
2	0.0239	0.9761	2	2	1
3	0.0001	0.9999	2	2	1
4	0.9845	0.0155	1	2	0
5	0.0157	0.9843	2	2	1
6	0.9998	0.0002	1	2	0
7	0.0728	0.9272	2	2	1
8	0.9790	0.0210	1	2	0
9	0.9976	0.0024	1	2	0
10	0.9996	0.0004	1	2	0
11	0.6210	0.3790	1	2	0
12	0.6653	0.3347	1	2	0
13	0.0125	0.9875	2	2	1
14	0.6608	0.3392	1	2	0
15	0.9785	0.0215	1	2	0
16	0.9070	0.0930	1	2	0
17	0.8706	0.1294	1	2	0
18	0.6174	0.3826	1	2	0
19	0.0070	0.9930	2	2	1
20	0.9846	0.0154	1	1	1
21	0.9977	0.0023	1	1	1
22	0.9486	0.0514	1	1	1
23	0.8360	0.1640	1	2	0
24	0.0494	0.9506	2	2	1
25	0.3264	0.6736	2	2	1
26	0.3049	0.6951	2	2	1
27	0.9241	0.0759	1	2	0
28	0.0269	0.9731	2	2	1
29	0.7805	0.2195	1	2	0
30	0.2080	0.7920	2	2	1
31	0.0306	0.9694	2	2	1
32	0.9961	0.0039	1	2	0
33	0.3395	0.6605	2	2	1
34	0.0024	0.9976	2	2	1
35	0.5948	0.4052	1	2	0

36	0.4201	0.5799	2	2	1
37	0.9371	0.0629	1	2	0
38	0.9971	0.0029	1	2	0
39	0.8466	0.1534	1	2	0
40	0.9874	0.0126	1	2	0
41	0.9860	0.0140	1	2	0
42	0.9803	0.0197	1	2	0
43	0.0211	0.9789	2	2	1
44	0.9922	0.0078	1	2	0
45	0.9956	0.0044	1	2	0
46	0.0391	0.9609	2	2	1
47	0.9232	0.0768	1	1	1
48	0.9950	0.0050	1	2	0
49	0.9973	0.0027	1	1	1
50	0.9868	0.0132	1	1	1
51	0.9939	0.0061	1	1	1
52	0.9821	0.0179	1	1	1
53	0.9956	0.0044	1	1	1
54	0.0808	0.9192	2	2	1
55	0.8209	0.1791	1	2	0
56	0.9900	0.0100	1	1	1
57	0.0042	0.9958	2	2	1
58	0.9097	0.0903	1	2	0
59	0.9965	0.0035	1	1	1
60	0.9311	0.0689	1	1	1
61	0.9617	0.0383	1	1	1
62	0.9303	0.0697	1	1	1
63	0.9232	0.0768	1	2	0
64	0.9441	0.0559	1	1	1
65	0.9997	0.0003	1	2	0
66	0.8943	0.1057	1	2	0
67	0.9471	0.0529	1	1	1
68	0.9864	0.0136	1	1	1
69	0.9404	0.0596	1	1	1
70	0.9995	0.0005	1	1	1
71	0.0101	0.9899	2	2	1
72	0.9381	0.0619	1	1	1
73	0.2648	0.7352	2	2	1
74	0.9756	0.0244	1	2	0
75	0.9994	0.0006	1	1	1
76	0.5663	0.4337	1	2	0
77	0.9875	0.0125	1	1	1
78	0.1168	0.8832	2	2	1

79	0.0046	0.9954	2	2	1
80	0.9993	0.0007	1	1	1
81	0.9883	0.0117	1	1	1
82	0.9977	0.0023	1	1	1
83	0.1945	0.8055	2	2	1
84	0.0094	0.9906	2	2	1
85	0.9965	0.0035	1	1	1
86	0.0400	0.9600	2	2	1
87	0.9205	0.0795	1	2	0
88	0.0392	0.9608	2	2	1
89	0.9991	0.0009	1	1	1
90	0.9160	0.0840	1	1	1
91	0.9022	0.0978	1	1	1
92	0.7898	0.2102	2	2	1
93	0.9902	0.0098	1	1	1
94	0.9890	0.0110	1	1	1
95	0.8340	0.1660	1	2	0
96	0.0079	0.9921	2	1	0
97	0.9976	0.0024	1	1	1
98	0.9558	0.0442	1	1	1
99	0.9908	0.0092	1	1	1
100	0.9270	0.0730	1	2	0
101	0.9764	0.0236	1	2	0
102	0.9027	0.0973	2	1	0
103	0.9984	0.0016	1	1	1
104	0.9571	0.0429	1	1	1
105	0.9697	0.0303	1	1	1
106	0.9959	0.0041	1	2	0
107	0.9908	0.0092	1	1	1
108	0.9992	0.0008	1	1	1
109	0.0865	0.9135	2	2	1
110	0.9870	0.0130	1	1	1
111	0.9543	0.0457	1	1	1
112	0.9999	0.0001	1	1	1
113	0.9399	0.0601	1	1	1
114	0.9691	0.0309	1	1	1
115	0.9335	0.0665	1	1	1
116	0.9958	0.0042	1	1	1
117	0.9386	0.0614	1	1	1
118	0.8733	0.1267	1	2	0
119	0.6607	0.3393	1	2	0
120	0.1565	0.8435	2	2	1
121	0.9886	0.0114	1	1	1

122	0.0384	0.9616	2	2	1
123	0.1651	0.8349	2	2	1
124	0.9290	0.0710	1	1	1
125	0.9896	0.0104	1	1	1
126	0.9730	0.0270	1	1	1
127	0.9816	0.0184	1	2	0
128	0.0076	0.9924	2	2	1
129	0.8853	0.1147	1	1	1
130	0.0000	10.000	2	2	1
131	0.9981	0.0019	1	1	1
132	0.7442	0.2558	1	2	0
133	0.5877	0.4123	1	2	0
134	0.7140	0.2860	1	1	1
135	0.0400	0.9600	2	2	1
136	0.9993	0.0007	1	2	0
137	0.9931	0.0069	1	1	1
138	0.9878	0.0122	1	1	1
139	0.8803	0.1197	1	2	0
140	0.9840	0.0160	1	1	1
141	0.9536	0.0464	1	1	1
142	0.5792	0.4208	1	2	0
143	0.9874	0.0126	1	1	1
144	0.9987	0.0013	1	1	1
145	0.9756	0.0244	1	1	1
146	0.9948	0.0052	1	1	1
147	0.9947	0.0053	1	2	0
148	0.8622	0.1378	1	1	1
149	0.9297	0.0703	1	1	1
150	0.9752	0.0248	1	1	1
151	0.9978	0.0022	1	1	1
152	0.9239	0.0761	1	1	1
153	0.9577	0.0423	1	1	1
154	0.9832	0.0168	1	1	1
155	0.9940	0.0060	1	1	1
156	0.9986	0.0014	1	1	1
157	0.1903	0.8097	2	2	0
158	0.3911	0.6089	2	1	0
159	0.9972	0.0028	1	1	1
160	0.9789	0.0211	1	1	0
161	0.9924	0.0076	1	1	1
162	0.0032	0.9968	2	2	1
163	0.0007	0.9993	2	2	1
164	0.9990	0.0010	1	1	1

165	0.1437	0.8563	2	2	1
166	0.8613	0.1387	1	1	1
167	0.9762	0.0238	1	1	1
168	0.3745	0.6255	2	2	1
169	0.1517	0.8483	2	2	1
170	0.8682	0.1318	1	1	1
171	0.9989	0.0011	1	1	1
172	0.9850	0.0150	1	2	0
173	0.7710	0.2290	1	2	0
174	0.9807	0.0193	1	1	1
175	0.9738	0.0262	1	1	1
176	0.9321	0.0679	1	1	1
177	0.9585	0.0415	1	1	1
178	0.5228	0.4772	1	2	0
179	0.9967	0.0033	1	1	1
180	0.9990	0.0010	1	1	1
181	0.2638	0.7362	2	2	1
182	0.0197	0.9803	2	2	1
183	0.7021	0.2979	1	2	0
184	0.9884	0.0116	1	1	1
185	0.8248	0.1752	1	2	0
186	0.9635	0.0365	1	1	1
187	0.0718	0.9282	2	2	1
188	0.9924	0.0076	1	1	1
189	0.9941	0.0059	1	1	1
190	0.9989	0.0011	1	1	1
191	0.9578	0.0422	1	2	0
192	0.9992	0.0008	1	1	1
193	0.9536	0.0464	1	1	1
194	0.9997	0.0003	1	2	0
195	0.8896	0.1104	1	2	0
196	0.9982	0.0018	1	1	1
197	0.9727	0.0273	1	2	0
198	0.0921	0.9079	1	2	0
199	0.0050	0.9950	2	2	1
200	0.9268	0.0732	1	2	0
201	0.9987	0.0013	1	1	1
202	0.2153	0.7847	2	2	1
203	0.1435	0.8565	2	2	1
204	0.9669	0.0331	1	2	0
205	0.9999	0.0001	1	1	1
206	0.8134	0.1866	1	2	0
207	0.9561	0.0439	1	1	1

208	0.3272	0.6728	2	2	1
209	0.9960	0.0040	1	1	1
210	0.7957	0.2043	1	1	1
211	0.0139	0.9861	2	2	1
212	0.9939	0.0061	1	1	1
213	0.2958	0.7042	2	2	1
214	0.2230	0.7770	2	2	1
215	0.9573	0.0427	1	2	0
216	0.9785	0.0215	1	2	0
217	0.9947	0.0053	1	1	1
218	0.9648	0.0352	1	1	1
219	0.0022	0.9978	2	2	1
220	0.0015	0.9985	2	2	1
221	0.9834	0.0166	1	1	1
222	0.9866	0.0134	1	1	1
223	0.9621	0.0379	1	1	1
224	0.7148	0.2852	1	2	0
225	0.9919	0.0081	1	1	1
226	0.9286	0.0714	1	1	1
227	0.9675	0.0325	1	1	1
228	0.8703	0.1297	1	1	1
229	0.9998	0.0002	1	1	1
230	0.9994	0.0006	1	2	0
231	0.3507	0.6493	2	2	1
232	0.9868	0.0132	1	1	1
233	0.9844	0.0156	1	1	1
234	0.0218	0.9782	2	2	1
235	0.9506	0.0494	1	1	1
236	0.9631	0.0369	1	1	1
237	0.1389	0.8611	2	2	1
238	0.0181	0.9819	2	2	1
239	0.9459	0.0541	1	1	1
240	0.2862	0.7138	2	2	1
241	0.9804	0.0196	1	1	1
242	0.9997	0.0003	1	1	1
243	0.9853	0.0147	1	1	1
244	0.9722	0.0278	1	1	1
245	0.0035	0.9965	2	2	1
246	0.9702	0.0298	1	1	1
247	0.9933	0.0067	1	1	1
248	0.9987	0.0013	1	1	1
249	0.9731	0.0269	1	1	1
250	0.9894	0.0106	1	1	1

251	0.0359	0.9641	2	2	1
252	0.9896	0.0104	1	1	
253	0.0002	0.9998	2	2	1
254	0.2679	0.7321	2	2	1
255	0.0014	0.9986	2	2	1
256	0.9638	0.0362	1	2	0
257	0.0004	0.9996	2	2	1
258	0.8192	0.1808	1	2	1
259	0.6831	0.3169	1	2	0
260	0.7525	0.2475	1	2	0
261	0.0134	0.9866	2	2	1
262	0.2567	0.7433	2	2	1
263	0.2235	0.7765	2	2	1
264	0.7217	0.2783	1	2	0
265	0.2678	0.7322	2	2	1
266	0.0546	0.9454	2	2	1
267	0.9729	0.0271	1	1	1
268	0.9868	0.0132	1	1	1
269	0.9987	0.0013	1	1	1
270	0.9725	0.0275	1	1	1
271	0.9378	0.0622	1	1	1
272	0.9849	0.0151	1	1	1
273	0.0801	0.9199	2	2	1
274	0.9542	0.0458	1	1	1
275	0.1281	0.8719	2	2	1
276	0.9953	0.0047	1	1	1
277	0.9870	0.0130	1	1	1
278	0.0224	0.9776	2	2	1
279	0.9819	0.0181	1	1	1
280	0.9738	0.0262	1	1	1
281	0.0076	0.9924	2	2	1
282	0.9938	0.0062	1	1	1
283	0.0057	0.9943	2	2	1
284	0.6002	0.3998	1	2	0
285	0.9982	0.0018	2	1	0
286	10.000	0.0000	1	1	1
287	0.9962	0.0038	1	1	1
288	0.9983	0.0017	1	1	1
289	0.9865	0.0135	1	1	1
290	0.9869	0.0131	1	1	1
291	0.9170	0.0830	1	1	1
292	0.8659	0.1341	1	1	1
293	0.9986	0.0014	1	1	1

294	0.9948	0.0052	1	1	1
295	0.9999	0.0001	1	1	1
296	0.9765	0.0235	1	1	1
297	0.9781	0.0219	1	1	1
298	0.9945	0.0055	1	2	0
299	0.9373	0.0627	1	1	1
300	0.9690	0.0310	1	1	1
301	0.0009	0.9991	2	2	1
302	0.9995	0.0005	1	1	1
303	0.0047	0.9953	2	2	1
304	0.9691	0.0309	1	1	1
305	0.9887	0.0113	1	1	1
306	0.9918	0.0082	1	1	1
307	0.9944	0.0056	1	1	1
308	0.9389	0.0611	1	1	1
309	0.9847	0.0153	1	1	1
310	0.9958	0.0042	1	1	1
311	0.9921	0.0079	1	1	1
312	0.8996	0.1004	1	1	1
313	0.9994	0.0006	1	1	1
314	0.9900	0.0100	1	1	1
315	0.9300	0.0700	1	1	1
316	0.9999	0.0001	1	1	1
317	0.9987	0.0013	1	1	1
318	0.0882	0.9118	2	2	1
319	0.9384	0.0616	1	1	1
320	0.9998	0.0002	1	1	1
321	0.9658	0.0342	1	1	1
322	0.0100	0.9900	2	2	1
323	0.9994	0.0006	1	1	1
324	0.0079	0.9921	2	2	1
325	0.9983	0.0017	1	1	1
326	10.000	0.0000	1	1	1
327	0.9528	0.0472	1	1	1
328	0.9969	0.0031	1	1	1
329	0.5769	0.4231	1	2	0
330	0.5403	0.4597	1	2	0
331	0.6329	0.3671	1	2	0
332	0.9985	0.0015	1	1	1
333	0.9847	0.0153	1	1	1
334	0.9854	0.0146	1	1	1
335	0.9990	0.0010	1	1	1
336	0.2910	0.7090	2	2	1

337	0.9985	0.0015	1	1	1
338	0.0281	0.9719	2	2	1
339	0.9613	0.0387	1	1	1
340	0.1611	0.8389	2	2	1
341	0.9285	0.0715	1	1	1
342	0.9510	0.0490	1	1	1
343	0.9811	0.0189	1	1	1
344	0.0000	10.000	2	2	1
345	0.9924	0.0076	1	1	1
346	0.9649	0.0351	1	1	1
347	0.9968	0.0032	1	1	1
348	0.8941	0.1059	1	1	1
349	0.9885	0.0115	1	1	1
350	0.9937	0.0063	1	1	1
351	0.9926	0.0074	1	1	1
352	0.7205	0.2795	1	2	0
353	0.2228	0.7772	2	2	1
354	0.8134	0.1866	1	2	1
355	0.9839	0.0161	1	1	1
356	10.000	0.0000	1	1	1
357	0.9987	0.0013	1	1	1
358	0.9699	0.0301	1	1	1
359	0.9370	0.0630	1	1	1
360	0.9503	0.0497	1	1	1
361	10.000	0.0000	1	1	1
362	0.9920	0.0080	1	1	1
363	0.9998	0.0002	1	1	1
364	0.5087	0.4913	1	1	1
365	0.9900	0.0100	1	1	1
366	0.0146	0.9854	2	2	1
367	0.0028	0.9972	2	2	1
368	0.9984	0.0016	1	1	1
369	0.0990	0.9010	2	2	1
370	0.0770	0.9230	2	2	1
371	0.5017	0.4983	1	2	0
372	0.8228	0.1772	1	1	1
373	0.0434	0.9566	2	2	1
374	0.0266	0.9734	2	2	1
375	0.9785	0.0215	1	1	1
376	0.6453	0.3547	1	1	1
377	0.9704	0.0296	1	1	1
378	0.9863	0.0137	1	1	1
379	0.9776	0.0224	1	1	1

380	0.9776	0.0224	1	2	0
381	0.9844	0.0156	1	1	1
382	0.9807	0.0193	1	1	1
383	0.9971	0.0029	1	1	1
384	0.9989	0.0011	1	1	1
385	0.9932	0.0068	1	1	1
386	0.8996	0.1004	1	2	0
387	0.9987	0.0013	1	1	1
388	0.9680	0.0320	1	1	1
389	0.9859	0.0141	1	1	1
390	0.0009	0.9991	2	2	1
391	0.9649	0.0351	1	1	1
392	0.9346	0.0654	1	1	1
393	0.7536	0.2464	1	2	0
394	0.0506	0.9494	2	2	1
395	0.9969	0.0031	1	1	1
396	0.9589	0.0411	1	1	1
397	0.9879	0.0121	1	1	1
398	0.9991	0.0009	1	1	1
399	0.9822	0.0178	1	1	1
400	0.9946	0.0054	1	1	1
401	0.1350	0.8650	2	2	1
402	0.9963	0.0037	1	1	1
403	0.9969	0.0031	1	1	1
404	0.9992	0.0008	1	1	1
405	0.9993	0.0007	1	1	1
406	0.9800	0.0200	1	1	1
407	0.6149	0.3851	1	1	1
408	0.9985	0.0015	1	1	1
409	0.1395	0.8605	2	2	1
410	0.9991	0.0009	1	1	1
411	0.9878	0.0122	1	1	1
412	0.9809	0.0191	1	1	1
413	0.9469	0.0531	1	1	1
414	0.8552	0.1448	1	1	1
415	0.8078	0.1922	1	2	0
416	0.9950	0.0050	1	1	1
417	0.9477	0.0523	1	1	1
418	0.6067	0.3933	1	2	0
419	0.9998	0.0002	1	1	1
420	0.9828	0.0172	1	1	1
421	0.9902	0.0098	1	1	1
422	0.9106	0.0894	1	1	1

423	0.9899	0.0101	1	1	1
424	0.9803	0.0197	1	1	1
425	0.9541	0.0459	1	1	1
426	0.9541	0.0459	1	1	1
427	0.9602	0.0398	1	1	1
428	0.9688	0.0312	1	1	1
429	0.9770	0.0230	1	1	1
430	0.9830	0.0170	1	1	1
431	0.9996	0.0004	1	2	0
432	0.8688	0.1312	1	1	1
433	0.9993	0.0007	1	2	0
434	0.0052	0.9948	2	2	1
435	0.0182	0.9818	2	1	0
436	0.8870	0.1130	1	2	0
437	0.9659	0.0341	1	1	1
438	0.9990	0.0010	1	1	1
439	0.9572	0.0428	1	1	1
440	0.9706	0.0294	1	1	1
441	0.9609	0.0391	1	1	1
442	0.9804	0.0196	1	2	0
443	0.2665	0.7335	1	1	1
444	0.9747	0.0253	1	1	1
445	0.9712	0.0288	1	2	0
446	0.1423	0.8577	2	1	0
447	0.9962	0.0038	1	2	0
448	0.1569	0.8431	2	1	0
449	0.8853	0.1147	1	1	1
450	0.9063	0.0937	1	2	0
451	0.0426	0.9574	2	1	0
452	0.9947	0.0053	1	2	0
453	0.0000	10.000	2	1	0
454	0.9963	0.0037	1	1	1
455	0.9252	0.0748	1	1	1
456	0.9999	0.0001	1	1	1
457	0.9880	0.0120	1	1	1
458	0.9913	0.0087	1	1	1
459	0.9942	0.0058	1	1	1
460	0.9976	0.0024	1	1	1
461	0.9545	0.0455	1	2	0
462	0.2618	0.7382	2	2	1
463	0.2960	0.7040	2	1	0
464	0.9230	0.0770	1	1	1
465	0.9908	0.0092	1	1	1

466	0.9944	0.0056	1	1	1
467	0.9928	0.0072	1	1	1
468	0.9944	0.0056	1	1	1
469	0.9529	0.0471	1	2	0
470	0.1586	0.8414	2	1	0
471	0.9900	0.0100	1	1	1
472	0.9539	0.0461	1	1	1
473	0.9973	0.0027	1	1	1
474	0.8666	0.1334	1	1	1
475	0.9989	0.0011	1	1	1
476	0.9767	0.0233	1	1	1
477	0.9992	0.0008	1	1	1
478	0.9510	0.0490	1	1	1
479	0.9660	0.0340	1	1	1
480	0.9891	0.0109	1	2	0
481	0.5705	0.4295	1	1	1
482	0.9981	0.0019	1	1	1
483	0.9636	0.0364	1	1	1
484	0.9919	0.0081	1	1	1
485	0.9824	0.0176	1	1	1
486	0.7480	0.2520	1	1	1
487	0.9998	0.0002	1	1	1
488	0.8979	0.1021	1	2	0
489	0.0017	0.9983	2	1	0
490	0.9925	0.0075	1	2	0
491	0.4538	0.5462	2	1	0
492	0.9991	0.0009	1	1	1
493	0.1393	0.8607	2	2	1
494	0.1155	0.8845	2	1	1
495	0.9997	0.0003	1	1	1
496	0.9940	0.0060	1	1	1
497	0.8760	0.1240	1	1	1
498	10.000	0.0000	1	1	1
499	0.9999	0.0001	1	2	1
500	0.0456	0.9544	2	2	1
501	0.0221	0.9779	2	1	1
502	0.8623	0.1377	1	2	0
503	0.9681	0.0319	1	1	1
504	0.9997	0.0003	1	2	0
505	0.1425	0.8575	2	1	0
506	0.9398	0.0602	1	1	1
507	0.9482	0.0518	1	1	1
508	0.9978	0.0022	1	1	1

509	0.9790	0.0210	1	1	1
510	0.5604	0.4396	1	2	0
511	0.7826	0.2174	1	1	1
512	0.9936	0.0064	1	1	1
513	0.8765	0.1235	1	2	0
514	0.9884	0.0116	1	1	1
515	0.9075	0.0925	1	2	0
516	0.8411	0.1589	1	1	1
517	0.9857	0.0143	1	2	0
518	0.0601	0.9399	2	2	1
519	0.0007	0.9993	2	1	0
520	0.9999	0.0001	1	1	1
521	0.9999	0.0001	1	1	1
522	0.9430	0.0570	1	2	0
523	0.1856	0.8144	2	1	0
524	0.9849	0.0151	1	1	1
525	0.9825	0.0175	1	1	1
526	0.9553	0.0447	1	1	1
527	0.9300	0.0700	1	1	1
528	0.9904	0.0096	1	1	1
529	0.9992	0.0008	1	1	1
530	0.9695	0.0305	1	1	1
531	0.9968	0.0032	1	1	1
532	0.9930	0.0070	1	1	1
533	0.9916	0.0084	1	1	1
534	0.9803	0.0197	1	2	0
535	0.0162	0.9838	2	1	0
536	0.9787	0.0213	1	2	0
537	0.0187	0.9813	2	2	1
538	0.9403	0.0597	1	1	1
539	0.9894	0.0106	1	1	1
540	0.9150	0.0850	1	1	1
541	0.9121	0.0879	1	1	1
542	0.9121	0.0879	1	1	1
543	0.9886	0.0114	2	1	0
544	0.9103	0.0897	1	1	1
545	0.8941	0.1059	1	1	1
546	0.9939	0.0061	1	1	1
547	0.9752	0.0248	1	1	1
548	0.9814	0.0186	1	1	1
549	0.9660	0.0340	1	1	1
550	0.9647	0.0353	1	1	1
551	0.9528	0.0472	1	1	1

552	0.9775	0.0225	1	1	1
553	0.9772	0.0228	1	1	1
554	0.9823	0.0177	1	1	1
555	0.9990	0.0010	1	1	1
556	0.9452	0.0548	1	1	1
557	0.9983	0.0017	1	1	1
558	0.9649	0.0351	1	1	1
559	0.9617	0.0383	1	1	1
560	0.9477	0.0523	1	1	1
561	0.9094	0.0906	1	1	1
562	0.9887	0.0113	1	1	1
563	0.9650	0.0350	1	2	0
564	0.9843	0.0157	1	2	0
565	0.8117	0.1883	1	2	0
566	0.0305	0.9695	2	2	1
567	0.0758	0.9242	2	2	1
568	0.0073	0.9927	2	2	1
569	0.4521	0.5479	2	1	0