ELSON CLAUDIO CORREA MORAES

MÉTODO NÃO SUPERVISIONADO BASEADO EM CURVAS PRINCIPAIS PARA RECONHECIMENTO DE PADRÕES

LAVRAS - MG

2016

ELSON CLAUDIO CORREA MORAES

MÉTODO NÃO SUPERVISIONADO BASEADO EM CURVAS PRINCIPAIS PARA RECONHECIMENTO DE PADRÕES

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, área de concentração em Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

Orientador

Dr. Danton Diego Ferreira

LAVRAS - MG

2015

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo (a) próprio(a) autor(a).

Moraes, Elson Claudio Correa.
Método não supervisionado baseado em curvas principais para reconhecimento de padrões / Elson Claudio Correa Moraes. – Lavras: UFLA, 2016.
132 p.

Dissertação (mestrado acadêmico) – Universidade Federal de Lavras, 2015.
Orientador (a): Danton Diego Ferreira.
Bibliografia.

1. Curvas principais. 2. k-segmentos. 3. Agrupamento. I. Universidade Federal de Lavras. II. Título.

ELSON CLAUDIO CORREA MORAES

MÉTODO NÃO SUPERVISIONADO BASEADO EM CURVAS PRINCIPAIS PARA RECONHECIMENTO DE PADRÕES

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, área de concentração em Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

APROVADA em 21 de dezembro de 2015.

Dr. Bruno Henrique Groenner Barbosa	UFLA
Dr. Ricardo Rodrigues Magalhães	UFLA
Dr. Giovani Bernardes Vitor	UTC

Dr. Danton Diego Ferreira
Orientador

LAVRAS - MG
2015

AGRADECIMENTOS

Ao Basquete UFLA, em especial ao Fabricio Rezende, pelo incentivo de trabalhar na área acadêmica.

Ao meu orientador, Danton Diego Ferreira, pela ajuda, incentivo e constante prontidão diante de qualquer desafio encontrado durante o mestrado.

À Universidade Federal de Lavras - UFLA, ao Departamento de Engenharia - DEG e ao Programa de Pós-Graduação em Engenharia de Sistemas e Automação PPGESISA\UFLA, pela oportunidade de realização do Mestrado.

À Capes, CNPq e FAPEMIG, pela concessão da bolsa de estudos.

RESUMO

No presente trabalho é proposto um novo método de agrupamento e classificação de padrões baseado em Curvas Principais. As Curvas Principais consistem numa generalização não linear da Análise de Componentes Principais e são definidas como curvas suaves, unidimensionais, que modelam um conjunto multidimensional de dados, fornecendo um resumo unidimensional destes. O algoritmo de extração de curvas principais que o método proposto se baseou é o k-segmentos não suave. O método divide a curva principal originalmente obtida pelo algoritmo k-segmentos não suave em duas ou mais curvas, de acordo com o número de agrupamentos definido pelo usuário. Em seguida é calculada a distância dos dados às curvas geradas pelo método e, posteriormente, é feita a classificação dos dados de acordo com o critério da menor distância dos dados às novas curvas. Utilizou-se como métrica para o cálculo da distância o quadrado da distância Euclidiana. O método foi aplicado a cinco bases de dados, duas bidimensionais e três multidimensionais. Os resultados foram comparados com os métodos *k-means* e S*elf Organized Maps*, em que o método proposto superou os demais métodos nas duas bases bidimensionais, com 100% de acerto, e obteve o segundo melhor resultado para as outras bases de dados. O método proposto é mais indicado para agrupamentos com distribuições alongadas e circulares no espaço de parâmetros. Apesar do desempenho alcançado, o método proposto apresentou forte sensibilidade aos parâmetros de entrada como comprimento do segmento e número de segmentos. O problema da sensibilidade aos parâmetros do método será investigado em trabalhos futuros.

Palavras-chave: Curvas principais. k-segmentos. Agrupamento. Reconhecimento de padrões.

ABSTRACT

In this work a new method of data clustering and pattern classification based on principal curves is presented. Principal curves consist of a nonlinear generalization of Principal Component Analysis and are smooth curves, one-dimensional, which model a multidimensional dataset, providing a one-dimensional summary of it. In the proposed method, the principal curves are extracted by the k-segments algorithm. The method divides the principal curves originally obtained by the k-segments algorithm into two or more curves, according to the number of clusters previously defined by the user. Then, the distances from the data to the curves generated by the method are calculated and thereafter it is made sorting the data according to the criterion of the smallest distance from data to the new curves. The square of the Euclidian distance is used. The method was applied to five databases, two two-dimensional and three multidimensional. The results were compared with the methods k-means and Self Organized Maps, where the proposed method outperformed the other methods in two bases (two-dimensional ones) and obtained the second best result in the other databases. The method shown to be suitable for elongated and circular clusters. Despite its high performance, the method shown to be very sensitive to the input parameters (the segment length and the number of segments). The author intend to exploit the problem of the sensitivity of the method in future works.

Keywords: Principal curves. k-segments. Clustering. Pattern recognition.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O tema reconhecimento de padrões é de grande interesse na comunidade científica devido ao fato de ser usado em diversos campos como psicologia, medicina, marketing, finanças, entre outros.

Conforme a sociedade evolui de uma era industrial para uma fase pós-industrial, a recuperação de informações se torna cada vez mais importante, fazendo com que o reconhecimento de padrões seja parte fundamental deste processo, sendo usado na maioria dos sistemas de inteligência computacional.

Métodos de reconhecimento de padrões têm sido usados em inúmeras áreas, dentre elas: análise, segmentação e pré-processamento de imagens, reconhecimento de faces, identificação de impressões digitais, reconhecimento de caracteres, análise de manuscritos, diagnóstico médico, reconhecimento e entendimento de voz, detecção de odores e agrupamento de dados.

Existem várias técnicas de reconhecimento de padrões como as redes neurais artificiais (RNA), árvores de decisão, PSO (*Particle Swarm Optimization*), KNN (*k-Nearest Neighbor*), *k-means*, entre outras. Cada técnica possui vantagens e desvantagens, sendo estas decisivas na escolha da técnica para um dado problema.

Em muitos dos problemas de reconhecimento de padrões não se sabe, a partir da base de dados de treinamento, a qual classe os padrões pertencem e nem o número de classes existentes. Neste tipo de problema, procura-se encontrar similaridades na base de dados e o uso de métodos de reconhecimento de padrões não supervisionados é necessário. Adicionalmente, muitos dos padrões são representados em alta dimensão, o que torna o problema de reconhecimento de padrões mais complexo e, muitas vezes, solucionável ao custo de algum pré-processamento de dados.

O contexto supracitado está presente em muitos dos problemas reais e as soluções atuais não são definitivamente efetivas e requerem melhorias. Os métodos atuais de reconhecimento de padrões não supervisionados são limitados a algumas aplicações que envolvem dados com distribuições específicas. Além disso, é uma realidade conhecida que as bases de dados podem ser de diferentes naturezas e complexidades, e, para um determinado tipo de distribuição, algoritmos de reconhecimento de padrões não supervisionado podem obter resultados diferentes.

Na presente dissertação de mestrado o objetivo principal foi explorar a capacidade de representação de dados da técnica de curvas principais, na solução de problemas de reconhecimento de padrões que requerem aprendizagem (treinamento) não supervisionada.

A técnica de curvas principais vem sendo utilizada com bastante sucesso em diversas aplicações no contexto de reconhecimento de padrões. Curva principal (CP) é uma suavização, curvilínea de dados *d*-dimensionais e foram introduzidas por Hastie e Stuetzle (1989). Trata-se de uma técnica com grande capacidade de representação de dados de alta dimensão em uma única dimensão e apresenta-se como uma opção interessante ao reconhecimento de padrões, visto que a mesma pode extrair padrões compactos de bases de dados com elevada dimensão, não requerendo, portanto, exaustivo pré-processamento. Ademais, a CP é capaz de fornecer uma descrição não linear de um conjunto de dados *d*-dimensionais em uma única dimensão.

O método proposto neste trabalho utiliza o algoritmo *K-segmentos não suave* para extrair a CP e se destaca em relação aos demais métodos baseados em curvas principais por sua simplicidade.

O método foi avaliado utilizando bases bidimensionais e multidimensionais, com um número variado de eventos (dados). Os algoritmos

K-means e o *Self-Organizing Maps* (SOM) foram também aplicados às bases de dados para fins de comparação.

Toda teoria, metodologia, resultados, análises e conclusões são apresentadas neste documento, organizado da seguinte forma. A Seção 2 apresenta uma revisão bibliográfica detalhada acerca das aplicações e características inerentes a reconhecimento de padrões; características ligadas ao treinamento supervisionado e não supervisionado; propriedades dos métodos de classificação não supervisionada e medidas de similaridades; definições matemáticas de *cluster* e os tipos de *clusters*; definições de *cluster* contextualizado sobre classificação não supervisionada; métodos de validação de *clusters*; propriedades de curvas principais e suas aplicações, características do algoritmo de extração de curvas principais usado neste trabalho (k-segmentos não suave) e dos métodos utilizados para fins de comparação (*K-means* e *self organized maps*).

Na Seção 3 é apresentado o passo a passo do método proposto e uma matriz de medidas absoluta que avalia o espalhamento dos dados, podendo avaliar inclusive em bases multidimensionais, em que a análise visual não é possível, podendo inferir os agrupamentos obtidos em termos de homogeneidade e heterogeneidade.

Na Seção 4 está descrita a base de dados utilizada na fase de testes do método proposto.

Na Seção 5 são apresentados os resultados para as bases bidimensionais e multidimensionais que, posteriormente, são comparados com os métodos *K-means* e *self organized maps*. São realizadas análises qualitativas e quantitativas para bases bidimensionais e quantitativas para bases multidimensionais. Também é apresentado um estudo em relação a capacidade de generalização do método proposto frente aos métodos *K-means* e *self organized maps*. São feitas análises do ponto de vista de validações de *clusters* por meio do método

silhouettes e da matriz de medidas aqui proposta. Posteriormente é feita uma apresentação e análise dos parâmetros do método proposto em função da porcentagem de acerto e do resultado da matriz absoluta para três tipos de *clusters*.

Por fim, na Seção 6, é apresentada a conclusão obtida ao final da elaboração deste trabalho, demonstrando suas contribuições, e finalizando com os possíveis trabalhos futuros.

2 REFERENCIAL TEÓRICO

Esta seção aborda uma revisão de alguns importantes conceitos acerca de reconhecimento de padrões. São apresentados conceitos básicos referentes à classificação supervisionada e não supervisionada, medidas de similaridades, validação de agrupamentos e alguns métodos de reconhecimentos de padrões.

2.1 Reconhecimento de padrões

Reconhecimento de padrões é uma área da ciência cujo objetivo é a classificação de objetos dentro de um número de categorias ou classes. As características destes objetos variam de acordo com cada aplicação, que podem ser imagens, caracteres, sinais em forma de onda (como voz, rádio, luz) ou qualquer tipo de medida onde exista a necessidade de ser classificada (THEODORIDIS; KOUTROUMBAS, 2009).

Com o avanço dos recursos computacionais, o interesse na área tem crescido, pois, do ponto de vista computacional, projetar e utilizar métodos de análise e classificação matematicamente complexos já é uma realidade. Kasabov (1996) aponta alguns exemplos de aplicações de reconhecimento de padrões:

a) classificação de doenças;
b) análise, segmentação e pré-processamento de imagens;
c) diagnóstico médico;
d) reconhecimento de faces;
e) mineração de dados;
f) identificação de impressões digitais;
g) reconhecimento de caracteres;
h) marketing;
i) estudos de terremoto.

Para exemplificar um problema de reconhecimento de padrões, observe as Figuras 1(a) e 1(b) que mostram as imagens de uma lesão benigna e uma lesão maligna, respectivamente. Neste exemplo, o problema real de classificação de lesão é modelado por meio de processamento de imagem, em que o padrão é visto como uma matriz de *pixels*.

O primeiro passo é identificar os eventos mensurando-os, de forma que se possa realizar uma distinção entre as duas classes de problemas. A Figura 2 demonstra uma representação gráfica do valor médio da intensidade da coloração em nível de cinza. Observa-se que eventos da classe A (o) tendem a se espalhar em uma área diferente dos eventos da classe B (+). Assim pode-se traçar uma linha reta que se apresenta como um bom indicador para separar as duas classes de problemas.

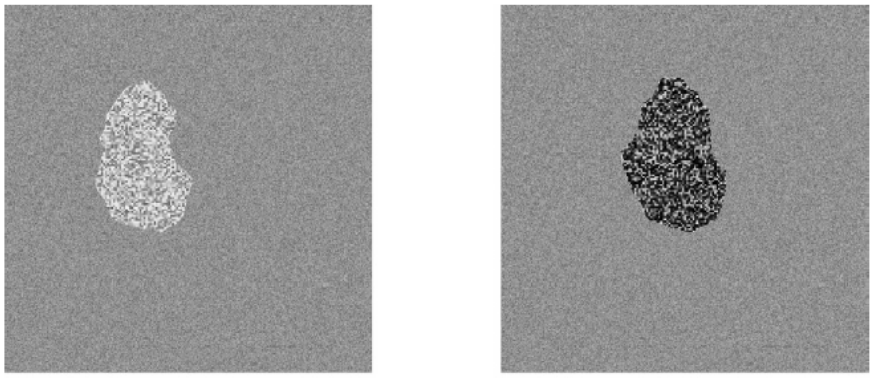


Figura 1 Exemplos de regiões de imagem correspondendo às classes A (a) e B (b)
Fonte: Adaptado de Theodoridis e Koutroumbas (2009)

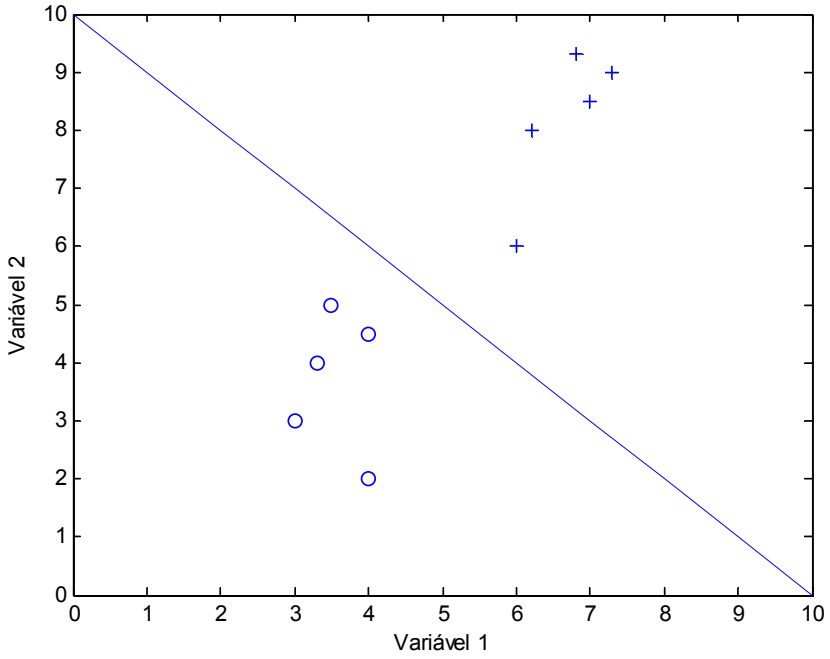


Figura 2 Fronteira de decisão entre duas classes

O procedimento para abstrair as características de uma base de dados real é por meio da construção de um vetor de características:

$$\mathbf{x} = [x_1, x_2, \ldots, x_l]^T \tag{1}$$

em que x_i representa cada variável (característica) extraída do objeto, i=1, ..., l, l é o número total de características da base de dados, T significa transposição. Dessa forma, cada característica do vetor identifica unicamente o padrão singular.

A linha reta na Figura 2 é conhecida como linha de decisão, cujo papel é dividir o espaço em regiões que correspondem à classe A ou à classe B. Se um vetor de características **x** desconhecido estiver na região referente à classe A, o

mesmo é classificado como pertencente à classe A, caso contrário ele é classificado como pertencente à classe B (THEODORIDIS; KOUTROUMBAS, 2009).

Os problemas de reconhecimento de padrões podem ser divididos em problemas linearmente e não linearmente separáveis. Problemas lineares são aqueles em que é possível separar duas ou mais classes usando uma linha reta ou um hiperplano como fronteira de decisão. Em contrapartida, problemas não linearmente separáveis são aqueles cujas classes não podem ser separadas por uma linha reta ou um hiperplano como fronteira de decisão. Pode-se observar na Figura 3 a diferença entre um problema "*or*", o qual é caracterizado por ser linearmente separável, e um problema "*xor*", caracterizado por ser não linearmente separável, onde uma linha ou um hiperplano não é suficiente para solucionar o problema de separação de classes.

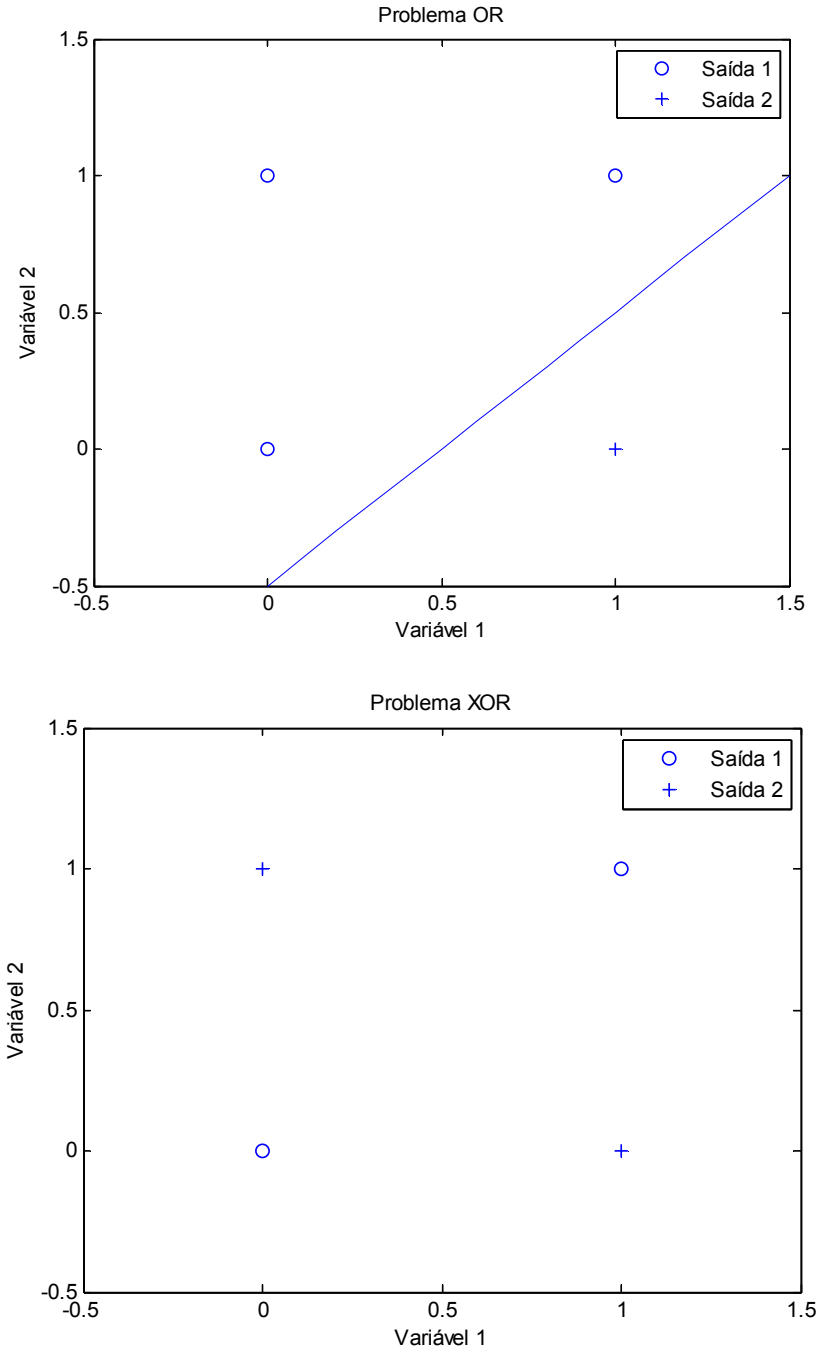


Figura 3 Problemas lineares e não lineares

Em geral, o problema de reconhecimento de padrões pode ser dividido nas etapas sequenciadas pelo diagrama em blocos da Figura 4, independente da ferramenta utilizada.

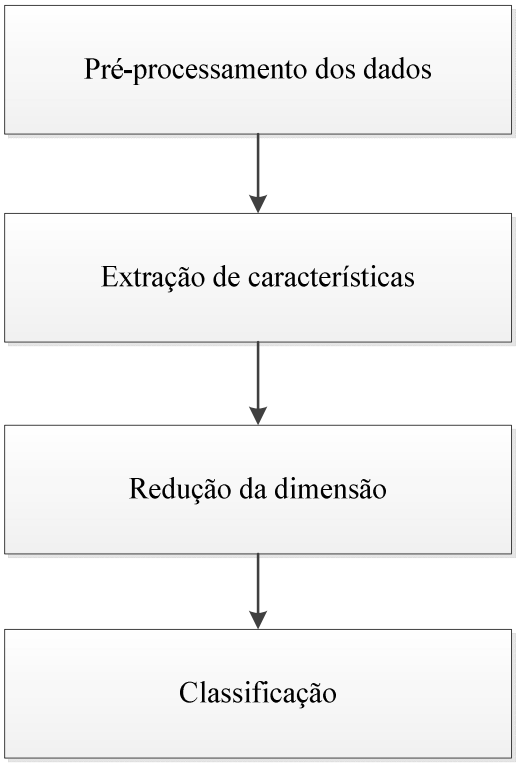


Figura 4 Estágios de um sistema de reconhecimento de padrões

O pré-processamento dos dados consiste em obter os dados e organizá-los em uma base de dados isenta de dados corrompidos e com características relevantes, obtendo assim uma base de dados em que o próximo estágio de extração de características possa ser aplicado sem que ocorram ambiguidades, garantindo assim um resultado mais efetivo (BATCHELOR, 2012).

O estágio de extração de características tem por objetivo extrair características relevantes para reconhecer um padrão. Durante esse estágio, é

importante determinar as características mais relevantes de um determinado problema, sendo a escolha das mesmas, determinada de acordo com a natureza dos dados.

Desta forma, o processo de escolha de quais características são mais relevantes é de grande importância, já que as mesmas devem ser extraídas de modo a servir como elementos básicos de um padrão (DEVIJVER; KITTLER, 1982).

O termo "*The course of dimensionality*" foi criado por Bellman em 1961 (BELLMAN, 1961), ele afirma que com o aumento da dimensão dos dados o número de problemas que surgem aumentam proporcionalmente. Dessa forma, a redução da dimensionalidade é uma etapa importante para identificar atributos que são relevantes em uma base de dados, removendo características redundantes e irrelevantes (ALBUS et al., 2012).

Métodos de redução de dimensionalidade aplicam diversas técnicas que permitem reduzir espaços de altas dimensões para dimensões menores. São exemplos de algumas dessas técnicas, a Análise de componentes principais (PCA), curvas principais e rede de Kahonen (KOHONEN, 2012).

No estágio de classificação ocorre a rotulação dos dados, estes são classificados ou rotulados de acordo com os resultados dos estágios anteriores. As técnicas de classificação de padrões podem ser supervisionadas e não supervisionadas. Na classificação supervisionada é fornecida a identificação de cada objeto da base de dados e na classificação não supervisionada as categorias envolvidas são conhecidas, mas a base de dados não possui nenhum tipo de rótulo. Um maior detalhamento de ambos os tipos de classificação será abordado na próxima seção.

2.2 Treinamento supervisionado e não supervisionado

De acordo com Antonio et al. (2002), uma vez extraída as características é necessária a classificação do dado (objeto ou padrão), sendo necessária a distinção entre as classes de dados, de forma que dados de mesma classe serão classificados da mesma forma. Caso o classificador exija um amplo conhecimento *a priori* da estrutura estatística dos padrões a serem analisados e o padrão de entrada for identificado como membro de uma classe pré-definida pelos padrões de treinamento, a classificação será dada na forma supervisionada. Por outro lado, se o classificador utilizar determinado modelo estatístico, ajustando-se mediante processos adaptativos e a associação entre padrões se fizer com base em similaridades entre os padrões de treinamento, a classificação será dada na forma não supervisionada (RAMIREZ; SPRECHMANN; SAPIRO, 2010).

Na Figura 1 da seção anterior foi exposta uma imagem médica, em que se assumiu *a priori* um conjunto de dados de treinamento disponível e foi possível projetar um classificador para esta base de dados. Neste exemplo, a base de dados disponível para o treinamento do classificador é previamente rotulada, portanto, é caracterizado como treinamento supervisionado.

Por outro lado, nem sempre existe a disponibilidade de uma base de dados previamente rotulada e, neste caso, o treinamento do classificador passa a ser não supervisionado, uma vez que não se conhecem os rótulos dos padrões.

Neste tipo de problema, o classificador tem como objetivo buscar por similaridades entre os eventos da base de dados, gerando assim agrupamentos de acordo com as similaridades e/ou dissimilaridades previamente definidas. Este tipo de classificação é comumente utilizado em aplicações como sensoriamento remoto, segmentação de imagens, codificações de fala, entre outros.

Uma das principais dificuldades no projeto de classificadores não supervisionados é a etapa de validação do classificador, já que não se tem em mãos a rotulação dos padrões. Dessa forma, muitas aplicações práticas requerem a análise dos dados após o agrupamento, ou por meio de especialista da área de aplicação do problema, ou por meio de medidas intra e entreclasses, bem como testes estatísticos. No entanto, o que se tem como resultado da validação é se os agrupamentos foram bem construídos, mas não há a certeza do acerto da classificação.

2.2.1 Métodos de classificação

Os métodos de classificação não supervisionada podem ser divididos em quatro categorias (THEODORIDIS; KOUTROUMBAS, 2009).

a) Sequenciais: são algoritmos simples e rápidos e têm a capacidade de produzir um único agrupamento. O resultado depende da ordem que a base de dados é apresentada ao método. O método tende a apresentar agrupamentos compactos em forma de hiperesferas.
b) Hierárquicos: nesta categoria há uma hierarquia de relacionamento entre os elementos, existem duas subcategorias, a aglomerativa que trabalha operando conjuntos de elementos isolados e a divisiva, que se inicia com um grande conjunto e vai quebrando-o em partes até chegar a elementos isolados. A principal vantagem dos algoritmos hierárquicos é o fato de que eles oferecem toda estrutura dos dados além dos *clusters*, permitindo assim o acesso a subconjuntos dentro da base de dados. Ademais, eles permitem visualizar diretamente através do dendôgrama, a forma como os dados se ligam e a verdadeira semelhança entre dois diferentes pontos (LIDEN, 2009).

c) Baseados na otimização da função de custo: são algoritmos que são quantificados por uma função de custo. Normalmente, o número de agrupamentos é mantido fixo nos parâmetros do algoritmo. Estes algoritmos utilizam conceitos de cálculo diferencial e tentam minimizar a função de custo, ocorre a convergência do algoritmo quando um mínimo local ou global é encontrado.
d) Outros: esta categoria contém algumas técnicas especiais de agrupamento que não podem ser atribuídas a qualquer uma das categorias acima. Incluem algoritmos como *Branch and bound*, algoritmos genéticos, métodos de relaxamento estocásticos, busca por vales (*Valley-seeking*), algoritmos de aprendizagem competitiva e algoritmos baseados nas técnicas de transformação morfológica.

2.3 Medidas de similaridades

As medidas de similaridade são muito importantes na área de reconhecimento de padrões não supervisionado.

Em um algoritmo de agrupamento de dados, o critério de busca baseia-se em uma função de similaridade, ou seja, uma função que recebe dois objetos e retorna a distância entre eles. Bons resultados consistem em obter agrupamentos com alta homogeneidade interna e alta heterogeneidade externa, isso se traduz na forma que eventos de uma mesma classe devem estar bem próximos um do outro e eventos de outra classe devem, preferencialmente, estar distantes dos eventos de outras classes (LIDEN, 2009).

Durante o processo de agrupamento, a medida de similaridade deve ser previamente definida. A medida de similaridade escolhida influencia diretamente no resultado do agrupamento. Isso se dá pelo fato de que a medida de similaridade deve corresponder às características usadas para distinguir um

agrupamento do outro, dessa forma não há nenhuma medida de similaridade padrão e melhor para todos os tipos de agrupamento (BATCHELOR, 2012).

A seguir são apresentadas algumas medidas de similaridade que serão utilizadas neste trabalho. O termo *t* se refere ao *t*-ésimo atributo dos objetos X_{it} e X_{jt}.

- Distância Euclidiana

$$d(X_i X_j) = \sqrt{\sum_{p=1}^{t} (X_{ip} - X_{jp})} \tag{2}$$

- Distância *Manhattan* ou *City Block*

$$d(X_i X_j) = \sum_{p=1}^{t} |(X_{ip} - X_{jp})| \tag{3}$$

Existem outras medidas de similaridade e dissimilaridade. Um melhor detalhamento pode ser encontrado em (THEODORIDIS; KOUTROUMBAS, 2009).

2.4 Tipos de *clusters*

Em Everitt (1981), *cluster* (ou agrupamento) é definido como regiões contínuas de um espaço que contenha uma elevada densidade de pontos, separados a partir de outras regiões de baixa densidade de pontos.

Sendo assim, nossa base de dados é definida como:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \quad (4)$$

em que $\mathbf{X}$ é um conjunto de eventos $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ e $c_1, c_2, \dots, c_m$ são *clusters* que devem seguir as seguintes condições:

- $c_1 \neq \emptyset$, $i = 1,\dots, m$
- $\cup_{i=1}^{m} c_i = X$
- $c_i \cap c_j = \emptyset, i \neq j, i, j = 1, \dots, m$

Assim, os eventos contidos no *cluster* c_i possuem certa similaridade entre si e certa dissimilaridade em comparação com os eventos do *cluster* $c_j, j \neq i$. Quantificar em termos de similaridade depende muito da distribuição dos eventos, veja na Figura 5 diferentes distribuições. Sendo assim, pode ser necessária uma investigação a cerca da medida de similaridade a ser utilizada para se obter *clusters* que modelem bem a base de dados.

(a) *Clusters* compactos (b) *Clusters* alongados (c) *Clusters* esféricos

Figura 5 Diferentes distribuições de dados

2.4.1 Definições de *cluster* dentro do contexto da classificação não supervisionada

A classificação não supervisionada busca por similaridades ou características em comum dentro de uma base de dados, permitindo que se forme um determinado agrupamento ou *cluster* de acordo com as semelhanças e diferenças entre os padrões da base de dados.

Agrupamentos podem ser aplicados em diversos campos como em ciências da vida (biologia, zoologia) (SILVA, 2014), ciências da terra (geografia, geologia), ciências sociais (sociologia, arqueologia), entre outros (THEODORIDIS; KOUTROUMBAS, 2009). Um ponto crucial é a escolha do critério de agrupamento, observe o exemplo a seguir.

Imagine os seguintes animais, ovelha, cão, gato (mamíferos), cobra, lagarto (repteis), tilápia, tubarão (peixe). É necessário criar um critério de agrupamento, por exemplo, o critério de existência de pulmões, que levaria a duas divisões, um agrupamento com tilápia e tubarão e outro com os demais animais (Figura 6(a)). Se o critério de agrupamento da temperatura do sangue for considerado, animais de sangue frio (animais ectodérmicos) formariam um grupo composto por tubarão, tilápia, cobra e lagarto e o outro agrupamento seria composto por animais de sangue quente (animais endotérmicos) composto por ovelha, cão, gato (Figura 6(b)).

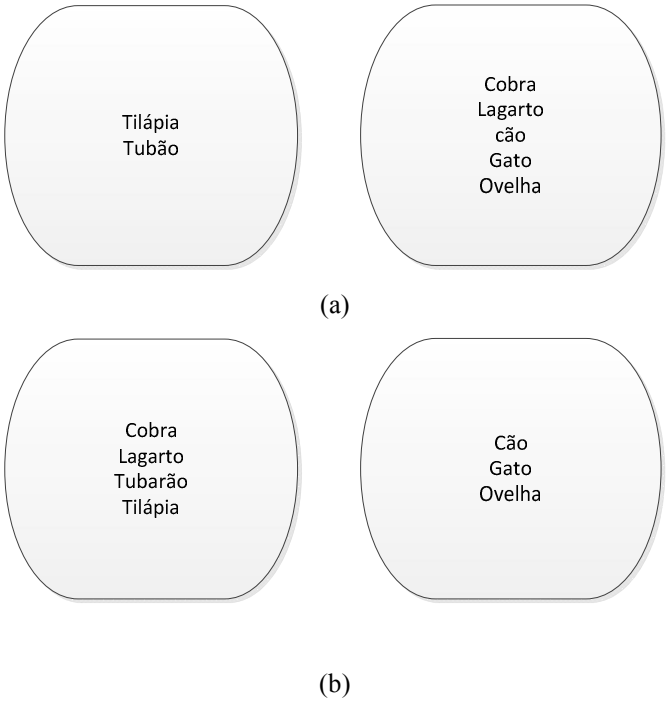


Figura 6 Critério de agrupamento: (a) existência de pulmões e; (b) temperatura do sangue

O exemplo ilustrado na Figura 6 demonstra o quanto é importante a definição do critério de agrupamento, critérios incorretos podem levar a agrupamentos incorretos e até vazios, e esta definição deve ser feita por um especialista da área, uma vez que o projetista do algoritmo que realiza os agrupamentos necessita desta informação para realizar os agrupamentos de forma correta de acordo com as necessidades do especialista.

De acordo com Theodoridis e Koutroumbas (2009), os seguintes passos devem ser seguidos para se realizar uma tarefa de agrupamento:

a) seleção de características - a seleção de características deve ser feita de modo a codificar o máximo possível as informações a respeito da base de dados;

b) medidas de similaridades - consiste na medida de quanto semelhante ou diferente é o evento da base de dados;
c) critério de agrupamento - neste passo é indispensável a presença do especialista, uma vez que neste ponto será disposto o critério que será buscado na base de dados;
d) métodos de agrupamento - escolha do algoritmo que realizará o agrupamento na base de dados;
e) validação dos resultados - após a realização dos agrupamentos é necessário verificar sua exatidão, em casos não supervisionados, a presença de um especialista é indispensável;
f) interpretação dos resultados - nesta fase é necessária a presença de um especialista para interpretar os resultados dos agrupamentos, a fim de tirar as conclusões de forma correta.

2.5 Validação de *clusters*

Um dos pontos mais importantes na análise de agrupamentos é a avaliação dos resultados de agrupamento. Por meio desta analise é possível encontrar a divisão de grupos que melhor se adapta a base de dados.

Deve-se destacar que resultados obtidos por métodos de validação de agrupamento não descartam a necessidade da presença de um especialista contextualizado na base de dados, pois os métodos de validação são apenas ferramentas para que o mesmo possa validar o agrupamento feito pelo método, principalmente em métodos não supervisionados.

Para Theodoridis e Koutroumbas (2009), de um modo geral, existem três abordagens para avaliar a qualidade de um agrupamento:

a) critérios externos - busca-se avaliar o resultado de um algoritmo de agrupamento com base em uma estrutura pré-estabelecida imposta a

um conjunto de dados e que reflete, intuitivamente, como os dados devem estar agrupados. Estes critérios são úteis para permitir uma avaliação e comparação objetiva entre diferentes algoritmos de agrupamento aplicados a bases de dados;

b) critérios internos - avaliam os resultados do agrupamento usando apenas a informação inerente à própria base de dados. Podem ser divididas em medidas que avaliam o ajuste ou correspondência entre os dados e o agrupamento obtidos, com o objetivo de medir o quanto um determinado agrupamento de dados corresponde a estrutura de agrupamento natural dos dados e em métodos de determinação do número de grupos existentes na base de dados;
c) critérios relativos - são baseados em critérios internos da classificação anterior, que avaliam os resultados do mesmo algoritmo de agrupamento, mas usando diferentes valores para os parâmetros de entrada nos algoritmos (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001).

2.5.1 *Silhouettes*

Silhoettes é um método de interpretação e validação da consistência dos agrupamentos, mensurando assim a medida de qualidade do agrupamento. Foi desenvolvida por Rousseeuw em 1987 (ROUSSEEUW, 1987).

A definição de *silhouettes* é feita a seguir:

$$Sil(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}} \tag{5}$$

em que $a(i)$ é a distância média dos eventos a partir de outros eventos dentro de um mesmo agrupamento, e $b(i)$ é a menor distância média de seus vizinhos. A equação (5) pode ser descrita como:

$$Sil(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & Se\ a(i) < b(i) \\ 0 & Se\ a(i) = b(i) \\ \dfrac{a(i)}{b(i)} - 1 & Se\ a(i) > b(i) \end{cases} \quad (6)$$

Valores *Sil*(*i*) próximos de 1 significa que as distâncias do evento $\mathbf{x}_i$ ao *cluster* ao qual ele pertence é significantemente menor do que as distâncias entre ele e seus *clusters* vizinhos. Esta é uma indicação de que $\mathbf{x}_i$ está bem agrupado. Valores *Sil*(*i*) próximos de -1 significa que a distâncias do evento $\mathbf{x}_i$ ao *cluster* ao qual ele pertence é significantemente alta em comparação às distâncias entre ele e seus *clusters* vizinhos. Esta é uma indicação de que $\mathbf{x}_i$ não está bem agrupado. Valores *Sil*(*i*) próximos de 0 significa que o evento $\mathbf{x}_i$ está próximo da borda de separação entre dois *clusters*.

Uma forma quantitativa de avaliar os *clusters* usando o método *silhouettes*, é utilizar a comparação entre as médias das *silhouettes*:

$$Sil_m = \frac{1}{m}\sum_{j=1}^{m} Sil_j \quad (7)$$

Valores maiores de Sil_m indicam que os grupos estão melhores definidos e valores menores indicam que os objetos daquele grupo não estão bem agrupados. Este índice é comumente utilizado para se ter uma ideia de quantos agrupamentos existem no banco de dados. O gráfico de Sil_m *versus m* (número de *clusters*) pode dar este indicativo (SILVA et al., 2014; THEODORIDIS; KOUTROUMBAS, 2009).

2.6 Curvas principais

Curvas principais foram definidas por Hastie e Stuetzle (1989) como sendo curvas suaves, unidimensionais, que passam no meio de um conjunto de dados multidimensional, fazendo uma representação compacta do mesmo. Outras definições foram propostas como em Delicado (2001), em que curvas principais são baseadas no conceito de componentes principais. Similar à definição de Delicado, em Jolliffe e Hope (1996) curvas principais foram definidas como uma generalização não linear da técnica de análise de componentes principais (PCA - *principal components analisys*) (JOLLIFFE, 2002).

PCA é uma técnica de processamento estatístico de sinais que aplica uma transformação ortogonal num conjunto de dados de observações, transformando-os em variáveis linearmente descorrelacionadas, chamadas de componentes principais. Componentes principais (JOLLIFFE, 2002) são representados pelos autovetores associados aos autovalores da matriz de covariâncias do conjunto de dados, os quais representam as direções ortogonais e o módulo de cada direção, respectivamente. O método permite uma representação compacta dos dados e permite a reconstrução dos dados em eixos cujas direções correspondem a uma ordenação maximizada da variância dos dados em ordem decrescente.

Conforme definido em Faier (2006), do ponto de vista matemático, as curvas principais são definidas a partir do conceito de autoconsistência, ou seja, os pontos que compõem a curva principal são a média dos dados que nela se projetam.

Uma curva unidimensional, num espaço de dimensões $\mathfrak{R}^d$, é um vetor $\mathbf{f}(t)$ de d funções contínuas de uma única variável t, ou seja, $\mathbf{f}(t) =$

$\{f_1(t), \ldots f_d(t)\}$. Essas funções são denominadas funções de coordenada e o parâmetro t proporciona o ordenamento ao longo da curva.

Seja $\mathbf{x}$ um vetor aleatório em $\Re^d$, possuindo densidade de probabilidade h, e momento de segunda ordem finito. Seja $\mathbf{f}$ uma curva suave parametrizada no intervalo fechado $I \subseteq \Re$, que não intercepta a si própria e de comprimento finito dentro de uma esfera de dimensões finitas em $\Re^d$.

O índice de projeção é definido como:

$$t_f(\mathbf{x}) = sup_t\{t: \|\mathbf{x} - \mathbf{f}(T)\| = inf_\mu||\mathbf{x} - \mathbf{f}(\mu)||\} \quad (8)$$

em que $\mathbf{x}$ é um evento arbitrário pertencente a $\mathbf{X}$ e μ é uma variável auxiliar definida em $\Re$. O índice de projeção $t_f(\mathbf{x})$ é o valor de t para o qual a curva principal $\mathbf{f}(t)$ está mais próxima de $\mathbf{x}$. Se houver mais de um valor possível, o maior deles é selecionado. A Figura 7 demonstra o índice de projeção relativo a uma curva principal. São mostrados cinco dados $\mathbf{x}_1$ ... $\mathbf{x}_5$, os quais projetam na curva principal respectivamente os pontos $\mathbf{f}\left(t_f(\mathbf{x}_1)\right), \ldots, \mathbf{f}\left(t_f(\mathbf{x}_5)\right)$.

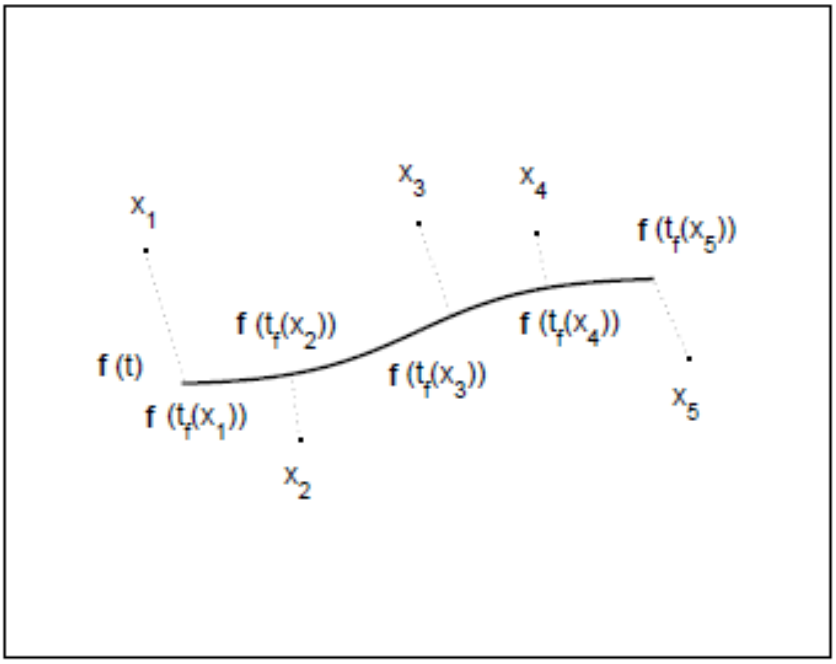


Figura 7 Projeção de dados na curva principal

A característica de autoconsistência decorre da propriedade dos pontos que compõem uma curva principal consistirem na média dos dados que nela projetam conforme em (8). Uma curva é autoconsistente se para todos os valores do parâmetro *t*, seus pontos forem a média das realizações de **X** que na curva se projetam ortogonalmente.

Diversos algoritmos foram propostos para extrair curvas principais. Em Hastie e Stuetzle (1989), curvas principais foram extraídas baseando-se no conceito de autoconsistência, entretanto o algoritmo posposto apresentou problemas como: tendências em trechos de curvatura, convergência e a existência de uma curva principal que não ocorria para qualquer direção.

Em Banfield e Raftery (1992) foi proposta uma correção para a tendência de estimação do algoritmo proposto por Hastie e Stuetzle (1989). Kégl et al. (2000) propuseram um algoritmo baseado em linhas poligonais, mais

eficiente e que apresenta melhores resultados que o algoritmo proposto por Hastie e Stuetzle. No trabalho de Eibek, Tutuz e Evers (2005) foi proposto um algoritmo baseado na definição de Delicado e Huerta (2003), em que se enfatizam a definição de curvas principais a partir de pontos orientados autoconsistentes. Em Verbeek et al. (2001) foi proposto o algoritmo k-segmentos não suave que além de ser robusto, e ter convergência garantida, é menos susceptível a mínimos locais. Este algoritmo é utilizado no presente trabalho para a extração de curvas principais.

2.6.1 Aplicações de curvas principais

Na presente dissertação de mestrado é apresentada uma técnica de reconhecimento de padrões utilizando curvas principais, sendo assim uma revisão acerca dos trabalhos em que curvas principais foram utilizadas para fins de reconhecimento de padrões foi feita e é apresentada nesta seção.

O trabalho de Banfield e Raftery (1992), propõe um método para a identificação de contornos de blocos de gelo com o uso da técnica de curvas principais associada ao algoritmo de propagação baseado no método da erosão (EP - *erosion-propagation*) (MURTAGH, 1985). Inicialmente o algoritmo EP seleciona os agrupamentos de *pixels* de borda dos blocos de gelo da imagem dos campos de gelo. Então a técnica de curvas principais é utilizada para realizar o reagrupamento dos dados gerados pelo algoritmo EP.

Chang e Ghosh (1998b) utilizaram curvas principais para a extração de características e classificação em problemas de dimensionalidade variada e diversos números de classes e dados. As bases de dados utilizadas foram obtidas na base de dados da Universidade de Irvine (LICHMAN; BACHE, 2013), sendo elas: Satimage com 6.435 dados, 36 dimensões e 6 classes, Diabetes com 768 dados, 8 dimensões e 2 classes, Glass com 214 dados, 9 dimensões e 6 classes e

iris com 50 dados, 4 dimensões e 3 classes. Os resultados foram comparados com diversas arquiteturas de redes neurais, tendo como resultado geral uma menor taxa de erro para o método de curvas principais.

Em um segundo trabalho, Chang e Ghosh (1998a) apresentam um classificador com Curvas Principais (CCP), em que a Curva Principal no espaço *n*-dimensional é extraída para cada classe usando os rótulos dos dados de treinamento. Um novo dado é rotulado de acordo com a distância euclidiana do evento em relação à curva principal. As bases de dados utilizadas foram também obtidas na base de dados da Universidade de Irvine (LICHMAN; BACHE, 2013). Os resultados são comparados com os métodos K-vizinhos mais próximos (KNN - *k-nearest neighbor*) (SONG et al., 2007) e perceptron multicamadas (MLP - *multilayer perceptron*) (WEST, 2000) e, de modo geral, os resultados são melhores com o método CCP.

Stanford e Raftery (2000) afirmam que agrupamento de dados com curvas principais combina a modelagem paramétrica do ruído e a modelagem não paramétrica dos dados a serem agrupados, podendo ser muito útil para detecção de características curvilíneas em padrões de dados especiais, com ou sem ruído. Os autores propõem um algoritmo de agrupamento dividido em duas etapas; primeiro é aplicado o algoritmo HPCC (*hierarchical principal curve clustering*), sendo este um algoritmo hierárquico e aglomerativo; e posteriormente usa-se o algoritmo CEM-PCC (*classification expectation maximization principal curve clustering*), sendo este baseado no algoritmo EM (*expectation maximization*) (DEMPSTER; LAIRD; RUBIN, 1977). O HPCC tem como objetivo obter o número de agrupamentos. Posteriormente, aplica-se o algoritmo CEM-PCC para refinar os agrupamentos e eliminar os ruídos. Foram utilizados dados bidimensionais para testar o método, porém os autores afirmam que o método pode ser aplicado a dados multidimensionais.

Cleju, Fränti, e Wu (2005) propuseram o uso de curvas principais para realizar agrupamento de dados. Os autores usam a ideia de projetar os dados na curva e posteriormente ordená-los. O método é comparado com uma abordagem que projeta os dados em um único eixo. Para isso o autor utilizou quatro métodos de projeção: PAC (*Principal Axis Based Clustering*), um método híbrido composto pelo PAC associado ao *K-means,* RLS (*Randomized Local Search*) (FRÄNTI; KIVIJÄRVI, 2000); e por último o algoritmo *K-means*. O método foi aplicado em bases de dados reais com até 16 dimensões e dados simulados com duas dimensões. Os resultados mostraram que o agrupamento baseado em curvas principais obteve melhores resultados em comparação com as abordagens que utilizam PAC, RLS, *K-means* e PAC associado com *K-means*.

No trabalho de Cleju, Franti e Wu (2005) foi utilizado o método de curvas principias associado ao algoritmo de árvore geradora mínima (ASUNÇÃO; LAGE; REIS, 2002). O método foi aplicado às bases de dados reais e simuladas. Para as bases de dados reais foram utilizadas imagens em tons de cinza com dimensões variando entre 3 e 16. Para a base de dados simulados foi utilizada uma base bidimensional com agrupamentos de misturas de gaussianas e outra base de dados bidimensional com variação de complexidade em termos de distribuição de dados espaciais. Primeiramente utiliza-se o algoritmo árvore geradora mínima para fazer um pré-agrupamento dos dados. Posteriormente usam-se curvas principais em cada um dos subgrupos gerados anteriormente pelo algoritmo árvore geradora mínima, gerando assim os novos agrupamentos. A vantagem deste método é que o subconjunto de dados que corresponde a um ramo da árvore pode ser facilmente modelado pelo método de curvas principais. Os resultados foram ligeiramente melhores em comparação com algoritmos clássicos de agrupamento como *k-means* e agrupamento sobre eixos principais.

Faier (2006) utilizou o método de curvas principais, por meio do algoritmo K-segmentos não suave, para extração de características e classificação de dados no contexto de descargas parciais em transformadores elétricos do sistema de potência. A base de dados teve dois enfoques, o especialista, em que se enfatizou a capacidade de generalização do classificador, e o estatístico, em que o desempenho do classificador foi avaliado segundo sua eficiência de classificação. O classificador baseado em curvas principais foi utilizado a partir de uma configuração dos histogramas de descargas parciais que podia operar no espaço de 1.024 ou de 3 dimensões. Após o uso de curvas principais para a extração de características de mapas de descargas parciais, as distâncias entre os dados e os segmentos foram avaliadas segundo os critérios de classificação: distância mínima, redes neurais e votação. Para mapas de 1.024 dimensões, a eficiência total dos classificadores atingiu índices elevados. Entretanto, os classificadores configurados para atuar em três dimensões não se mostraram robustos.

Em Ferreira et al. (2010), curvas principais foram utilizadas para a detecção de distúrbios em sistemas de potência combinada com redes neurais artificiais (RNA), em que curvas principais atuaram como extratoras de parâmetros e RNA são utilizadas como ferramenta de detecção. Aplicou-se o trabalho em bancos de dados sintéticos e experimentais. Os resultados foram comparados com as técnicas propostas em (GU et al., 2004; RIBEIRO et al., 2006). Os resultados mostraram-se bons em comparação com os outros dois métodos.

Ferreira et al. (2013) utilizaram curvas principais para o monitoramento da qualidade de energia em termos de análise de dados, detecção e classificação. CP foram utilizadas para realizar a extração de parâmetros, detecção e classificação. Foram utilizados sinais de tensão sintéticos com sete classes de distúrbios. Os resultados comprovaram a eficiência da técnica proposta.

No trabalho de Liu et al. (2013), curvas principais foram usadas para estimar o tempo de viagem de um veículo em cidades onde há grandes congestionamentos. A base de dados foi construída por meio de GPSs (*Global Positioning Systems*) acoplados em taxis que circulam na região metropolitana de Beijing, 2,6 milhões de trajetórias foram geradas a partir de 20 mil veículos. Foi analisado o resultado em atraso em três cenários entre 12:00 AM e 12:15 AM, 8:15 AM e 8:30 AM, 12:00 PM e 12:15 PM. O resultado obtido com curvas principais mostrou que, em média, 60% do tempo é desperdiçado em cruzamentos.

Em Ferreira et al. (2015) foi proposto um novo índice de desvio de qualidade de energia baseado em curvas principais. Foi utilizada uma base sintética de sinais de tensão e um banco de sinais experimentais do IEEE (Instituto de Engenheiros Elétrico Eletrônicos). Bons resultados foram obtidos e o método se mostrou versátil, podendo ser aplicado a qualquer tipo de perturbação.

Fica evidente a existência de muitos trabalhos com curvas principais na área de reconhecimento de padrões e seus resultados são bastante otimistas.

2.6.2 k-segmentos não suave

O algoritmo apresentado por Verbeek, Valssis e Krose (2002), o k-segmentos não suave (geralmente referido apenas como k-seg), propõe a construção passo a passo da curva principal criando-se primeiramente um único segmento e posteriormente o número de segmentos é aumentado progressivamente de acordo com os parâmetros do algoritmo. Todos os segmentos são interligados por segmentos independentes da curva principal. Pelo fato do k-seg ser a base para o método proposto nesta dissertação, torna-se necessária uma maior explanação do mesmo, que é feita a seguir.

O método k-seg produz curvas principais que podem interceptá-lo, item que o método proposto por Hastie e Stuetzle (1989) não alcança. Além disso, consegue obter uma aproximação relativamente boa para curvas complexas com convergência garantida (VERBEEK; VLASSIS; KROSE, 2002).

Em seu trabalho, Verbeek utiliza um modelo probabilístico para encontrar as curvas principais baseado no trabalho de Tibshirani (1992) por meio da maximização da verossimilhança logarítmica. Os dados são modelados por:

$$p(\mathbf{x}) = \int_0^1 p(\mathbf{x}|t)p(t)\,\mathrm{d}t \tag{9}$$

em que t é a variável latente distribuída no comprimento do arco da curva parametrizada de comprimento l, $p(\mathbf{x}|t)$ é um modelo gaussiano esférico modelando o ruído localizado no ponto t da curva com variância sobre todo t. A variância é um parâmetro de suavização, que pode ser configurado pelo usuário.

O algoritmo inicia-se na inserção do primeiro segmento *s*, sendo que todos os dados do universo são levados em conta, tendo assim um conjunto de dados em que o centro corresponde ao valor médio dos dados. Em posse desse primeiro conjunto é construído o primeiro segmento *s,* o qual tem a direção da primeira componente principal dos dados. Este primeiro segmento é composto por apenas uma parte da primeira componente principal, corresponde à direção da primeira componente principal do conjunto de dados com comprimento equivalente a $3\sigma/2$ associado a esta componente, em que σ^2 é a variância ao longo do primeiro componente principal.

A linha ***s*** é definida como:

$$\mathbf{s} = \{\mathbf{s}(t) | t \in \mathbb{R}\} \tag{10}$$

em que $\mathbf{s}(t) = \mathbf{c} + \mathbf{u}t$. A distância de um ponto **x** à linha ***s*** é definida como:

$$d(\mathbf{x}, s) = \min_{t \in \mathbb{R}} ||s(t) - \boldsymbol{x}|| \tag{11}$$

Seja $\boldsymbol{X}_n$ um conjunto de dados em $\mathfrak{N}^d$. Definimos as regiões de Voronoi (LANGE; BISHOP; RIPLEY, 1996) $V_1, \ldots, V_k$ como:

$$V_i = \{\mathbf{x} \in X_n | i = \arg\min d(\mathbf{x}, s_j)\} \tag{12}$$

em que, V_i contém todos os pontos (conjunto de dados), de forma que a *i*-ésima linha esteja mais próxima dos pontos, como observado na Figura 8, em que se observam as regiões de Voronoi V_1 e V_2 com seus respectivos segmentos. Assim, o algoritmo busca minimizar a distância total quadrática de todos os pontos à linha *k* de acordo com:

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in V_i} D(\mathbf{x}, s_i)^2 \tag{13}$$

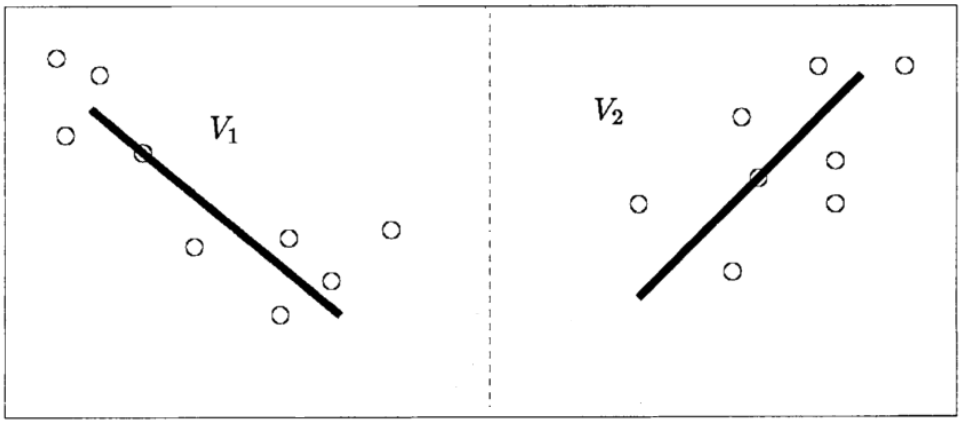


Figura 8 Regiões de Voronoi
Fonte: Adaptado de Verbeek, Vlassis e Krose (2002)

Na prática, foi observado por Verbeek que o algoritmo apresentava melhores resultados com a remoção da restrição de que os segmentos devem incluir todas as projeções dos pontos da região de Voronoi na primeira componente principal. Dessa forma, foi encontrado empiricamente, que o algoritmo apresenta melhor desempenho se forem usados segmentos da primeira componente principal, limitados ao comprimento de $3\sigma/2$ no centro da região de Voronoi. Para garantir a convergência, verifica-se para cada região de Voronoi se (13) decresce. Se (13) não decrescer, usa-se o segmento que inclui todas as projeções dos pontos na primeira componente principal ao invés de usar um segmento de comprimento $3\sigma/2$ do primeiro componente principal. Assim garante-se o decrescimento de (13). Em qualquer uma dessas situações, garante-se que todo segmento que compõe a curva principal atravesse os dados na direção da primeira componente principal de cada região de Voronoi, e a união destes segmentos resulta na curva principal poligonal que melhor se ajusta aos dados.

Para a inserção do segundo segmento é realizado um teste com todo o conjunto de dados, com o objetivo de encontrar qual é o melhor ponto a ser definido como centro do agrupamento correspondente a este segundo segmento. Dessa forma, para definir quais dados pertencem ao novo agrupamento é usado o algoritmo *k-means* ((DUDA; HART; STORK, 2012; THEODORIDIS; KOUTROUMBAS, 2009). Após a inserção do segundo segmento é realizado um cálculo das distâncias dos dados aos segmentos mais próximos, se um evento anteriormente pertencia ao primeiro segmento, é repetido o processo de agrupamento. Após este passo, estes dois segmentos são unidos por uma linha. A inserção de novos segmentos segue os passos acima até que o algoritmo convirja. Existem dois requisitos para que ocorra a convergência do k-seg; quando o número máximo de seguimentos é atingido (k = k_max) e quando a região de Voronoi tem menos de três eventos. O requisito que ocorrer primeiro leva o algoritmo à convergência.

O fluxograma demonstrado na Figura 9 ilustra o processo de extração da CP pelo algoritmo k-seg.

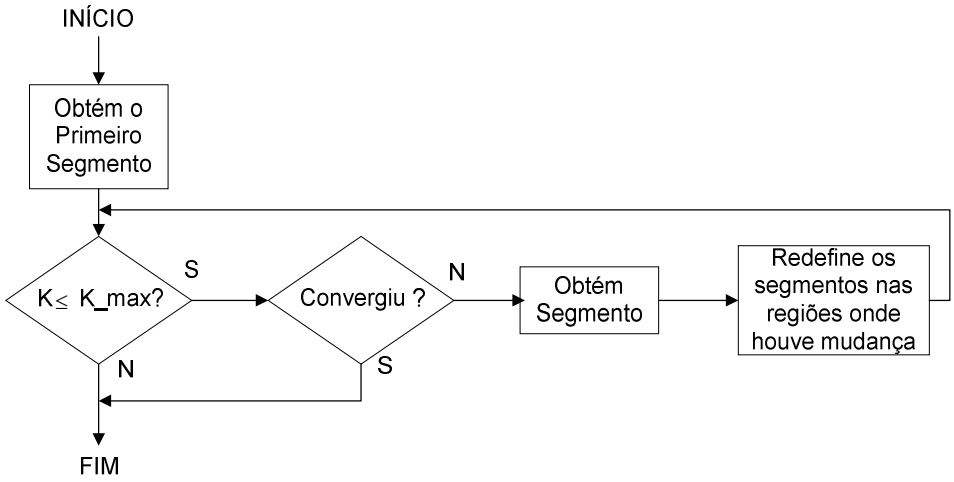


Figura 9 Fluxograma do algoritmo k-segmentos não suave

Além de ser robusto e ter a convergência garantida, o algoritmo k-seg requer poucos parâmetros iniciais. Observe a chamada da rotina que implementa o algoritmo em software MatLab:

[vertices; edges] = kseg(X; k max; alpha; lambda);

Em que:

vertices: é uma matriz que possui todas as coordenadas dos vértices que compõem a curva. Seu tamanho varia de acordo com o número de segmentos e a dimensão dos dados que estão na matriz de dados **X**.

edges: é uma matriz 2**Y* x 2**Y,* em que Y é o número de segmentos passado como parâmetro ao método. Nesta matriz encontram-se as definições de como os vértices da matriz *vertices* estão ligados. Para valores 0, não há conexão entre os vértices; para valores 1, significa que os vértices fazem parte de segmentos distintos, entretanto estão conectados; e, finalmente, para valores 2, significa que os vértices fazem parte do mesmo segmento

X: é a matriz contendo todos os dados.

kmax: é o número máximo de segmentos que a curva conterá, a ser definido pelo usuário.

alpha: é o dobro do quadrado da distância esperada entre os dados do conjunto de projeto e a curva. Muito embora este não seja um parâmetro opcional, ele não é empregado na extração da curva principal. Servindo apenas para a determinação de valores do vetor de saída.

lambda: após a obtenção dos segmentos, o algoritmo deve conectá-los, a partir de seus vértices, a fim de formar uma curva. Esta tarefa é realizada de forma a minimizar uma função de custo que considera as distâncias e os ângulos entre vértices dos segmentos. O objetivo é ligar segmentos que estejam próximos e obter uma curva, a mais suave possível. Se empregado o valor *default*, igual a 1, os dois objetivos - encurtar, ao máximo, a ligação entre segmentos e diminuir, tanto quanto possível, o ângulo formado pelas ligações

entre segmentos - contribuirão, igualmente, para minimizar o valor da função de custo. Aumentando-se o valor de *lambda*, dá-se prioridade a curvas mais suaves, enquanto que, diminuindo-se seu valor, se prioriza a obtenção de curvas de menor comprimento.

O algoritmo k-seg possui complexidade computacional $O(n^2)$, em que *n* é o número de eventos do conjunto de dados.

2.7 k-*means*

K-means é um algoritmo de agrupamento de dados, e tem por objetivo encontrar a melhor divisão de *P* dados em *K* grupos, de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja minimizada.

O algoritmo consiste em, dado um número previamente definido pelo usuário de *clusters*, calcular os pontos que representam os "centros", chamados centroides. Os centroides iniciais são formados aleatoriamente, e posteriormente é calculada a média das distâncias dos vetores de cada grupo aos centroides. Um processo iterativo é utilizado para encontrar os centroides finais, onde em cada passo os dados são agrupados ao *cluster* com o centroide mais próximo e posteriormente as médias são recalculadas. O algoritmo converge quando não houver mais alterações nas medias ou quando um número de iterações pré-determinadas for alcançado (PIMENTEL; FRANÇA; OMAR, 2003; THEODORIDIS; KOUTROUMBAS, 2009).

O critério de agrupamento do *k-means* é descrito pela equação a seguir:

$$E = \sum_{k=1}^{K} \sum_{X_i \in C_k} d(X_i, X_{0k}) \qquad (14)$$

em que K é o número de *clusters*, X_{0k} é o centroide do *cluster* C_k e $d(X_i, X_{0k})$ é a distância entre os pontos X_i e X_{0k}. Geralmente utiliza-se a distância Euclidiana, embora outras distâncias possam ser usadas.

2.8 *Self Organizing Maps*

Self Organizing Maps (SOM) ou mapa de Kohonen é um tipo de rede neural que se baseia no processo de aprendizagem competitiva, em que somente um neurônio de saída ou um grupo de neurônios representa uma resposta ao padrão de entrada na rede.

Durante o treinamento, os neurônios competem entre si, com o objetivo de quem é o neurônio vencedor. Posteriormente inicia-se o processo de atualização dos pesos deste neurônio e de seus vizinhos. Dessa forma, os neurônios se organizam topologicamente e se especializam na detecção de um conjunto de padrões de entrada, fazendo com que os padrões detectados por um neurônio estejam relacionados com a posição do mesmo. A Figura 10 representa um modelo bidimensional, um retangular e outro hexagonal, propostos por Kohonen (1998). É interessante citar que o modelo da rede SOM pode ser tanto bidimensional ou unidimensional, a escolha do modelo vai depender da natureza do problema.

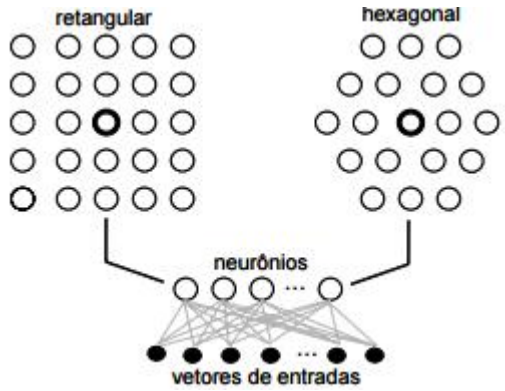


Figura 10 Modelo bidimensional da rede SOM

Os passos a seguir implementam o algoritmo de treinamento da rede SOM.

1) Inicialize os pesos e os parâmetros da rede SOM.
2) Para cada padrão de treinamento *x*,
   a) identifique o neurônio vencedor;
   b) atualize os pesos deste neurônio vencedor e de seus vizinhos.
3) Refaça o passo 2 até que não haja mudanças nas características do mapa.

Durante a fase de treinamento para cada vetor de entrada há apenas um neurônio vencedor. No entorno deste neurônio vencedor há uma cooperação topológica de neurônios, que serão excitados conforme uma função de vizinhança, assim os pesos sinápticos do neurônio vencedor e de seus vizinhos são adaptados conforme o padrão de entrada. A atualização dos pesos (passo 2(b)) é feita pela equação 17.

$$w_i(t+1) = w_i(t) + \lambda(t)h_{ij}(t)(x(t) - w_i(t)) \qquad (17)$$

em que $w_i(t)$ são os pesos sinápticos dos neurônios, $x(t)$ são os vetores de entrada, $\lambda(t)$ é uma taxa de aprendizagem monotonicamente decrescente e $h_{ij}(t)$ é a função de vizinhança, que é escolhida de forma a ter seu valor máximo do neurônio vencedor, decrescendo à medida que se afasta dele e tendo uma largura (número de neurônios abrangidos por ela) que decresça com o tempo (THEODORIDIS; KOUTROUMBAS, 2009).

No final do processo de aprendizagem cada neurônio ou grupo de neurônios representará um padrão distinto dentro do conjunto de padrões fornecidos como entrada para a rede (GONÇALVES et al., 1996).

A Figura 11 ilustra o processo de treinamento da rede SOM.

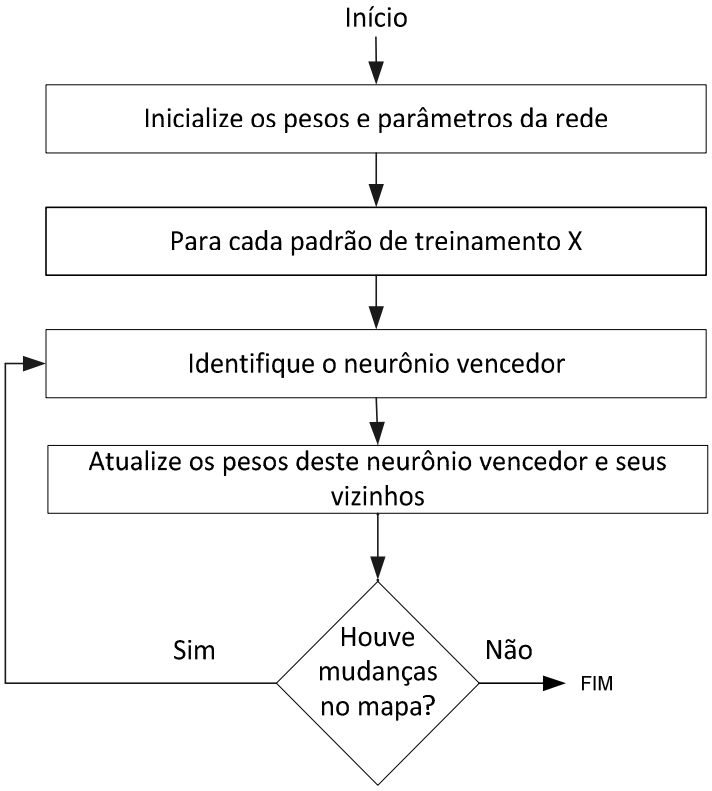


Figura 11 Fluxograma do algoritmo rede SOM

Uma forma de interpretar a rede SOM é observar o mapa semântico gerado pela atualização dos pesos dos neurônios (passo 2), pois durante o treinamento, os neurônios de certa vizinhança irão se mover para uma mesma direção, pois eventos similares tendem a ativar neurônios adjacentes, formando um mapa semântico em que eventos semelhantes são mapeados conjuntamente e os dissimilares são separados.

Por meio da análise visual do mapa semântico é possível fazer uma análise qualitativa dos dados e identificar os agrupamentos presentes na base de dados. Todavia, esta análise pode ser muito subjetiva quando os agrupamentos não estão bem definidos. Por outro lado, podem-se utilizar outros métodos de agrupamento, aplicados sobre os neurônios da rede SOM, para se obter um resultado quantitativo, como por exemplo o *k-means* empregado no trabalho de Cascao (2011).

3 MATERIAIS E MÉTODOS

3.1 Método proposto

O método de agrupamento de dados proposto neste trabalho utiliza a CP extraída pelo algoritmo k-seg para encontrar os agrupamentos. Após a extração da CP, ela é particionada em duas ou mais curvas, de acordo com o número de agrupamentos definido pelo usuário. Para isso, a interligação ou as interligações, de maior comprimento, entre os segmentos da CP são identificadas automaticamente e eliminadas. Como resultado, duas ou mais curvas principais passam a representar o conjunto de dados, de tal forma que os eventos que apresentarem a menor distância à mesma curva principal pertencerão ao mesmo agrupamento.

Como vantagem, o método proposto mantém os segmentos originalmente construídos pelo algoritmo k-seg, não alterando, portanto, a forma de representação da curva construída pelo algoritmo.

As etapas do método proposto para encontrar os agrupamentos em um conjunto de dados qualquer são:

a) Passo 1: construção da curva principal de acordo com o algoritmo k-seg;
b) Passo 2: cálculo do comprimento das interligações entre os segmentos da curva principal. O comprimento das interligações é obtido utilizando o quadrado da distância euclidiana entre os vértices das interligações;
c) Passo 3: divisão da curva principal em duas ou mais, dependendo do número de agrupamentos definido pelo usuário. Isso é feito eliminando a interligação de maior comprimento, no caso de dois agrupamentos, e eliminando as interligações de maior comprimento,

no caso de mais de dois agrupamentos. A inspeção das distâncias entre os segmentos da CP pode dar uma ideia de quantos agrupamentos existem no banco de dados;

d) Passo 4: rotulação dos dados. A distância dos dados às novas curvas principais é calculada. Os eventos que obtiverem a menor distância à mesma curva principal são rotulados como pertencentes ao mesmo agrupamento. No caso do evento estiver à mesma distância de duas ou mais curvas principais, este será rotulado como pertencente ao agrupamento que possuir o maior número de dados.

A Figura 12 ilustra as etapas do método proposto.

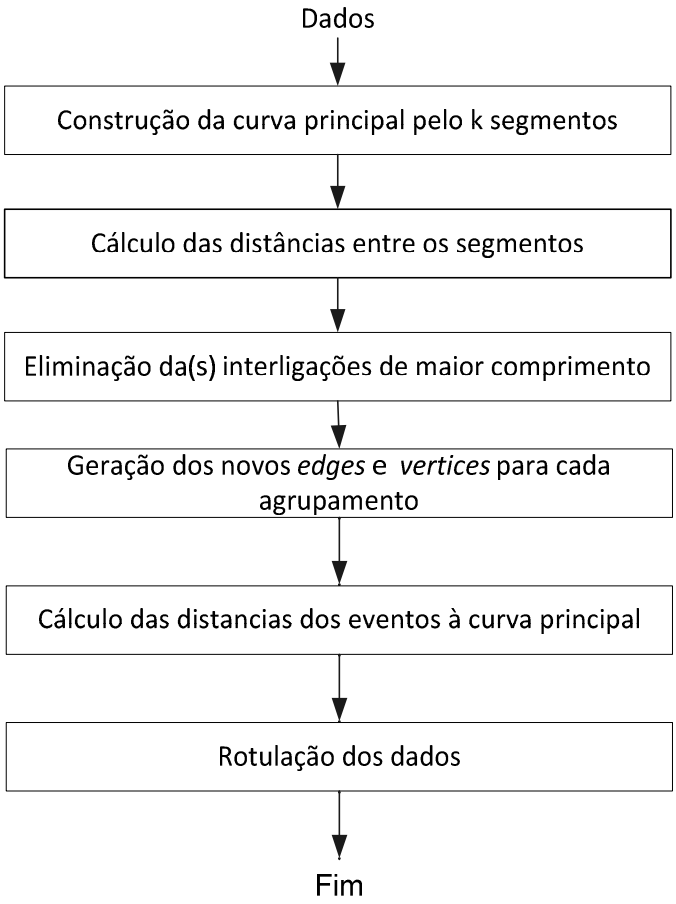


Figura 12 Fluxograma do método proposto

As etapas do método proposto são exemplificadas a seguir, a partir de um conjunto de dados simulados em duas dimensões (Figura 13). A Figura 13 ilustra a curva principal, construída pelo algoritmo k-seg, passando através do conjunto de dados. Nota-se ver que existem dois agrupamentos distintos e que a interligação de maior comprimento entre dois segmentos da curva principal une os agrupamentos. Os segmentos menos espessos representam as interligações

entre os segmentos da curva principal e os segmentos mais espessos são aqueles obtidos nas regiões de Voronoi e na direção da primeira componente principal.

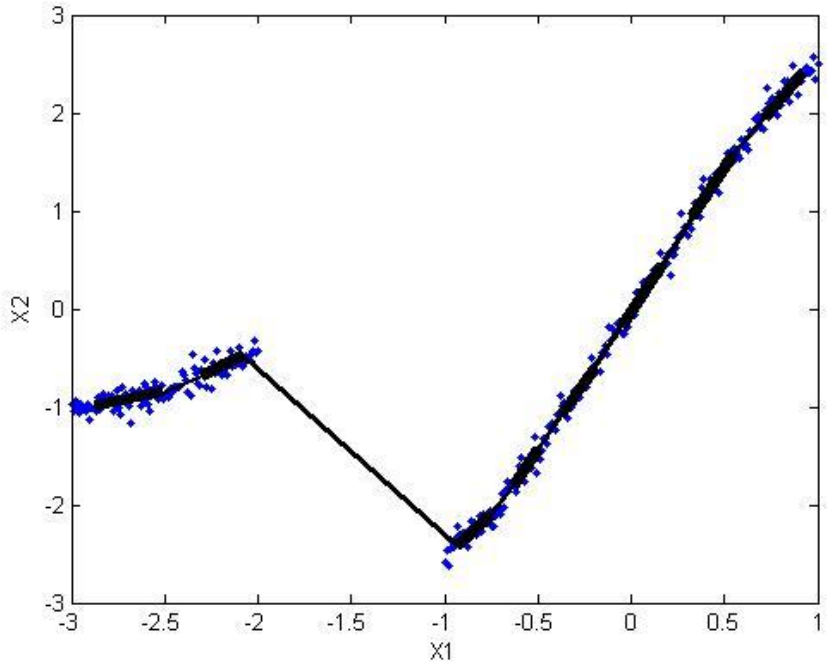


Figura 13 Curva principal original construída sobre um conjunto de dados em duas dimensões

A Figura 14 demonstra o comprimento de todas as interligações dos segmentos. Observa-se que a interligação representada no eixo das abscissas pelo número 5; apresenta o comprimento maior, e que as demais interligações apresentam comprimentos relativamente próximos entre si e consideravelmente inferiores ao comprimento de número 5.

A análise comparativa entre o comprimento das interligações é bastante útil para identificar quantos agrupamentos estão presentes no banco de dados. Comprimentos maiores indicam agrupamentos mais distantes uns dos outros.

Por outro lado, agrupamentos muito próximos no conjunto de dados podem ser difíceis de serem identificados por meio dessa análise.

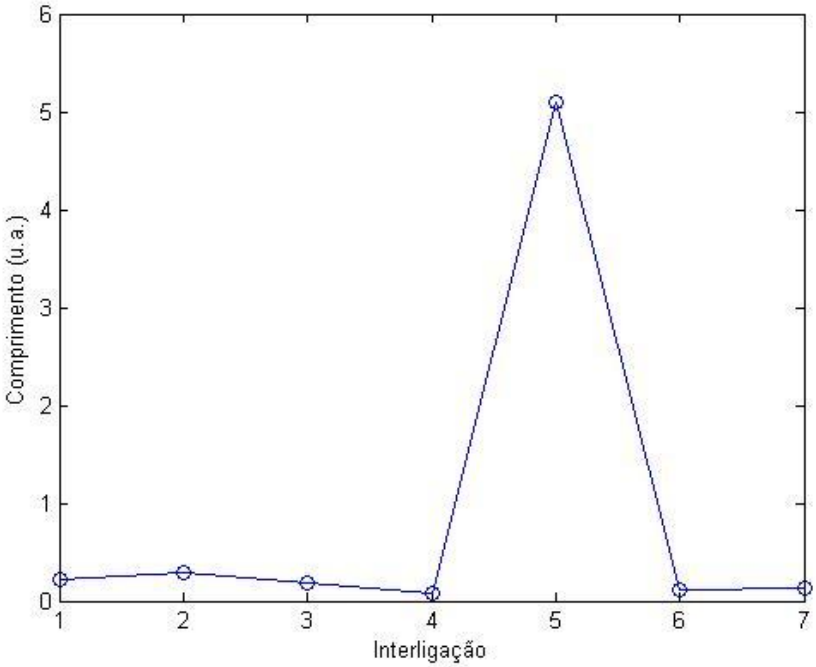


Figura 14 Distância em unidades arbitrárias (u.a.) entre os vértices das interligações da curva principal demonstrada na Figura 13

O próximo passo do método proposto consiste em obter as novas curvas principais que representam os novos agrupamentos. Isso é feito eliminando a interligação ou as interligações de comprimento maior. A Figura 15 demonstra este resultado, em que se eliminou apenas a interligação de maior comprimento.

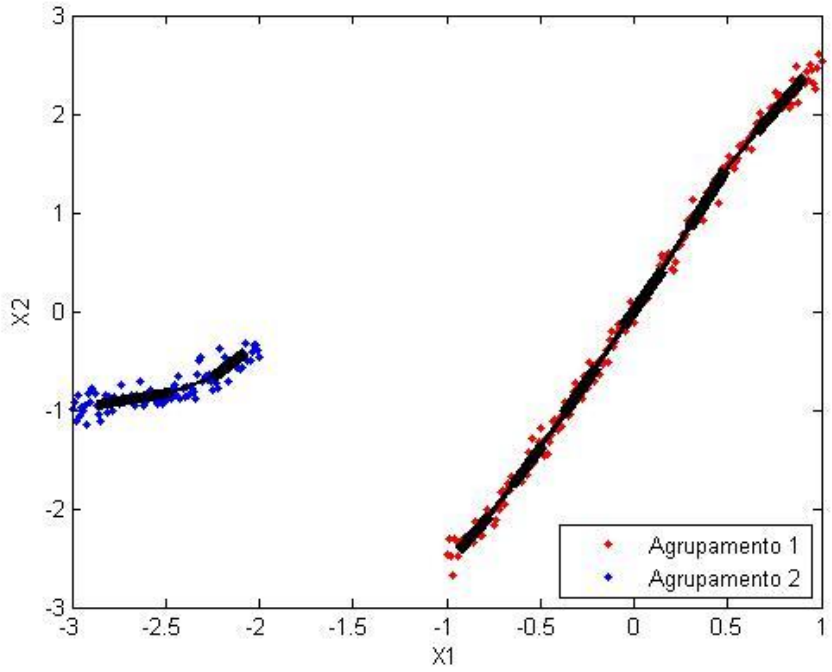


Figura 15 Exemplo de agrupamento gerado pelo método proposto

No último passo do método, os dados são rotulados seguindo a regra da menor distância à curva principal. A Figura 15 ilustra os dados rotulados e as novas curvas principais. Observa-se que os segmentos da curva original não foram alterados e os agrupamentos foram encontrados com sucesso.

Com o objetivo de verificar a qualidade dos agrupamentos obtidos propõe-se a construção de uma matriz de medidas quadrada, definida como:

$$\mathbf{M} = \begin{bmatrix} m_{11} & \cdots & m_{1J} \\ \vdots & \ddots & \vdots \\ m_{I1} & \cdots & m_{IJ} \end{bmatrix} \tag{18}$$

em que m_{ij}, para i = 1,..., I e j = 1,..., J, é a média das distâncias euclidiana dos dados do agrupamento i à curva principal j. Com isso, os termos da diagonal principal representam as medidas intraclasses e os demais representam as medidas extraclasses. A medida intraclasses indica o quanto os dados de um mesmo grupo estão próximos da curva principal que os representa enquanto a medida extraclasse indica o quanto os dados estão distantes das demais curvas principais (que representam os demais grupos). Essa informação é bastante útil para se analisar o espalhamento dos dados dos agrupamentos em espaços de alta dimensão, em que a análise visual não é possível e, consequentemente, inferir sobre os agrupamentos obtidos em termos de homogeneidade e heterogeneidade.

A matriz de medidas mostrada em (19) foi obtida para os agrupamentos da Figura 15 e mostra que os dados estão bem representados pelos seus respectivos agrupamentos. É importante enfatizar que foram obtidas distâncias iguais a zero para as medidas intraclasses devido ao número de casas decimais ter sido limitado a dois, uma vez que, para se obter um valor nulo seria necessário que os dados estivessem exatamente sobre a curva, o que não é o caso deste exemplo.

$$\mathbf{M} = \begin{bmatrix} 0{,}00 & 7{,}38 \\ 6{,}01 & 0{,}00 \end{bmatrix} \tag{19}$$

Para bases de dados com vários agrupamentos, analisar a matriz de medidas conforme (18) pode ser subjetivo. Uma alternativa é quantificar a matriz de medidas usando um único índice. Com este objetivo, foi proposto o índice equacionado em (18):

$$K_m = \frac{\frac{1}{m}\sum_{i=1}^{m} \mathbf{M}(i,i)}{z} \tag{20}$$

em que m é o número de agrupamentos e z é uma constante obtida pela média dos valores da matriz de medidas $\mathbf{M}(i,j)$ para os quais $i \neq j$ (valores fora da diagonal principal). Valores menores de K_m indicam que os grupos estão melhores definidos e valores maiores indicam que os objetos daquele grupo não estão bem agrupados. Este índice pode também ser utilizado para se ter uma ideia de quantos agrupamentos existe no banco de dados.

Toda implementação do método proposto, testes e demais análises foi realizada utilizando a ferramenta MatLab versão 7.8.0.347 (R2009a).

3.2 Parâmetros do método proposto

O método proposto possui, basicamente, três parâmetros de entrada: número de agrupamentos a serem obtidos, comprimento dos segmentos e número de segmentos da curva principal.

A chamada da função que implementa o método proposto em código MatLab possui a seguinte forma:

[ED, VR, C, M, D, edges, vertices] = ksegcluster(Xnor, numberofsegments, lengthofsegment, numberofcluster);

Em que:

ED: é uma matriz de células contendo referências às matrizes *edges* do algoritmo k-seg (veja a Seção 2.6.2), sendo que cada referência modela um único agrupamento realizado pelo método proposto.

VR: é uma matriz de células contendo referências às matrizes *vértices* do algoritmo k-seg (veja a Seção 2.6.2), sendo que cada referência modela um único agrupamento realizado pelo método proposto.

C: é um vetor contendo a rotulação dos dados obtida pelo método proposto.

M: é a matriz de medidas (veja a Seção 3.1).

D: é o vetor de distâncias entre as interligações dos segmentos.

vértices e *edges*: são as matrizes retornadas pelo algoritmo k-seg, conforme discutido na seção 2.6.2. Estas matrizes carregam informações a respeito da curva principal original do conjunto de dados.

Xnor: é a matriz contendo a base de dados.

numberofsegments: é o número máximo de segmentos que a CP conterá, definido pelo usuário.

lengthofsegment: é o comprimento do segmento da CP, definido pelo usuário.

numberofcluster: é o número de agrupamentos que a base de dados possui, definido pelo usuário.

3.3 Definição dos parâmetros do método proposto

Em seu trabalho, Verbeek, Vlassis e Krose (2002) citam que a obtenção do comprimento de segmento adequado para o método k-seg em uma dada base de dados é uma tarefa que pode ser computacionalmente complexa, se algum algoritmo de otimização for utilizado para esse fim. Dessa forma, Verbeek conclui, experimentalmente, que um bom comprimento para o segmento é o equivalente a 3/2 do desvio padrão (σ) dos dados ao longo do segmento.

Dessa forma com o objetivo de extrair o máximo possível do método proposto em termos de resultado, buscou-se obter os parâmetros do método proposto de forma que a taxa de acerto fosse a maior possível. Assim, buscou-se o comprimento e o número de segmentos para cada base de dados em um espaço de soluções, limitado entre 2 a 20 segmentos e comprimento de segmento de 0.1 a 2.5 do desvio padrão (σ) dos dados ao longo do segmento.

Uma análise mais profunda em relação ao comprimento e o número de segmentos é apresentada na seção 5.6 análise do comprimento do segmento e número de segmentos.

4 BASE DE DADOS

As bases de dados utilizadas para avaliar o método proposto foram obtidas no *UCI Machine Learning Repository* (LICHMAN; BACHE, 2013). Este repositório foi escolhido por disponibilizar bases de dados muito utilizadas pela comunidade científica para a análise empírica de algoritmos de aprendizado e devido às bases de dados serem compostas por dados multidimensionais, dados reais e simulados, envolvendo problemas clássicos de agrupamento e classificação.

Além dessas bases de dados, foram também utilizadas bases de dados sintéticas em duas dimensões, o que permite uma análise visual dos resultados obtidos pelo método proposto.

As bases de dados foram normalizadas pelo valor máximo absoluto (Equação 21), com o objetivo de minimizar problemas de dispersões distintas entre as variáveis.

$$\mathbf{x}_n(i) = \frac{\mathbf{x}(i)}{\left\|\mathbf{x}(i)\right\|_\infty} \tag{21}$$

em que $\mathbf{x}(i)$ é o vetor que carrega as variáveis que descrevem o padrão i da base de dados.

5 RESULTADOS

O método proposto foi aplicado às bases de dados com atributos, número de classes e dados variados. O número de segmentos e comprimento dos segmentos foi definido experimentalmente. No entanto, a Seção 5.6 apresenta um estudo, cujo objetivo é mostrar o comportamento do método proposto em função destes parâmetros.

Os resultados foram comparados com os resultados obtidos a partir dos algoritmos de agrupamento *k-means,* usando a distância Euclidiana (KM-E) e Manhattan (KM-M) e o algoritmo SOM na configuração unidimensional.

Para a implementação do algoritmo SOM, o *k-means* foi utilizado para agrupar os neurônios da rede SOM e, finalmente, obter a divisão entre grupos. O número de neurônios da rede SOM foi definido experimentalmente. Tentou-se adotar o número de neurônios próximo ao número de vértices da curva principal usada pelo método proposto.

Com o objetivo de verificar a capacidade de generalização do método proposto, as bases de dados foram divididas, de forma aleatória (seguindo uma distribuição uniforme no sorteio dos dados), em dados de treinamento (70% do total dos dados) e dados de teste (30% do total dos dados). Com o conjunto de dados de treinamento foram obtidas as CP para o método proposto, os centroides para o *k means* e a distribuição dos neurônios para a rede SOM. Posteriormente, os métodos foram aplicados ao conjunto de dados de teste. Este processo foi realizado 30 vezes e a média e o desvio padrão dos acertos para os dados de treinamento e para os dados de teste foram obtidos.

5.1 Base de dados espiral dupla

O método proposto foi aplicado a uma base de dados composta por dois agrupamentos na forma de duas espirais, sendo 206 dados no total. A base de dados modelada pela curva principal é demonstrada na Figura 16, e os agrupamentos obtidos pelo método são demonstrados na Figura 17(a). Observa-se que o método obteve 100% de acerto para esta base de dados. A Figura 17(b) representa os agrupamentos obtidos pelo algoritmo do *k-means* utilizando a distância Manhattan.

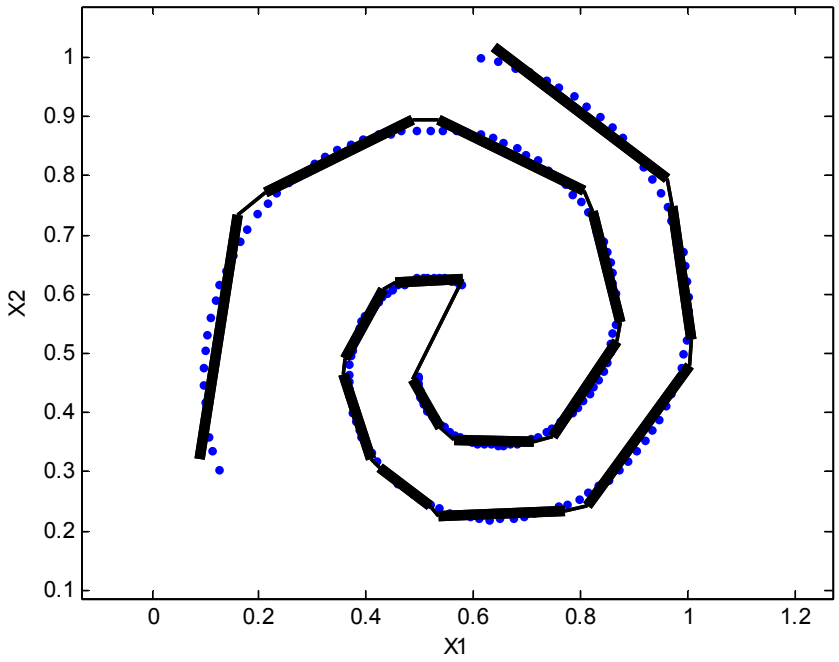
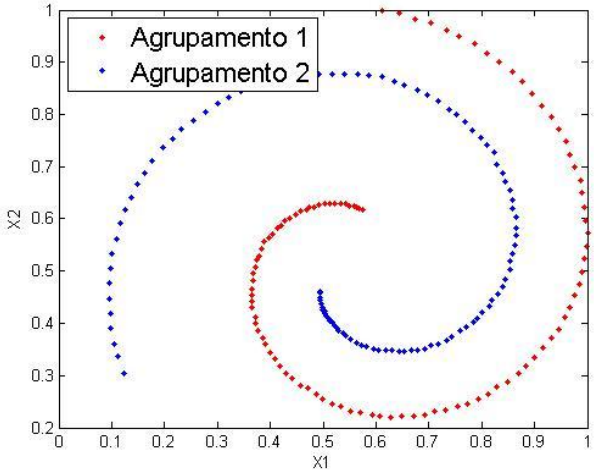
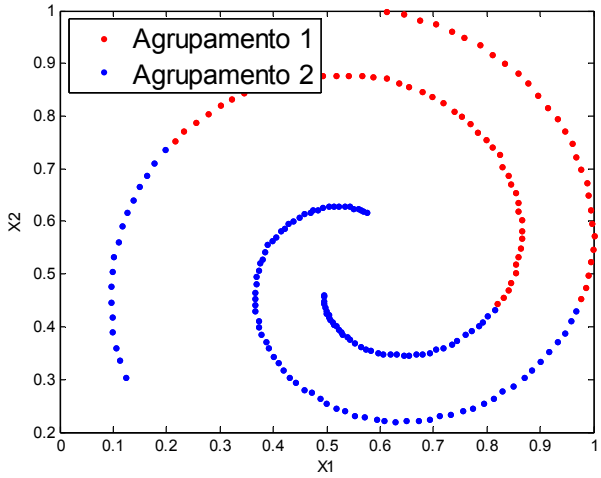


Figura 16 Curva principal obtida para a base de dados com duas espirais



(a)



(b)

Figura 17 Resultado do método proposto para a base de dados espiral dupla: (a) pelo método proposto (b) pelo método *k-means* com medida de similaridade Manhattan

A Figura 18 demonstra as distâncias entre os vértices das interligações da curva principal demonstrada na Figura 16. Pode-se concluir que existem apenas dois agrupamentos, uma vez que apenas um segmento de interligação com maior comprimento em relação aos outros segmentos foi encontrado.

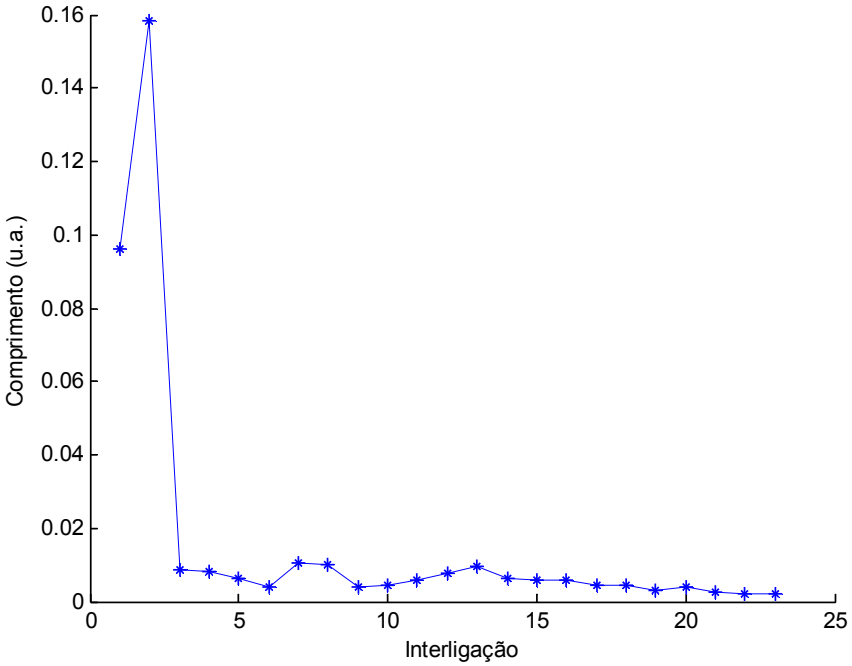


Figura 18 Distância em unidade arbitrária entre os vértices para a base de dados espiral dupla

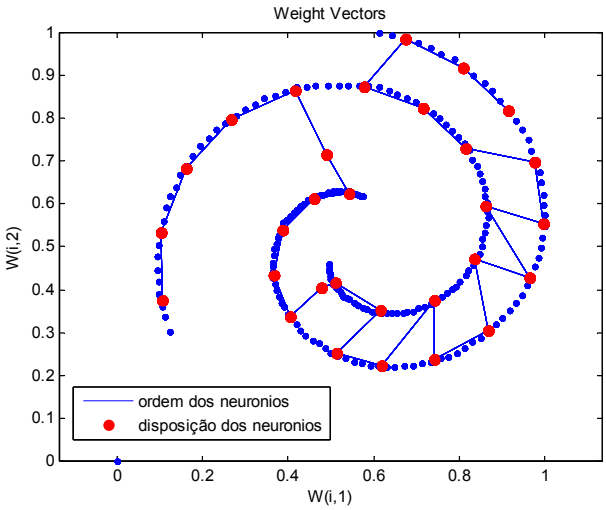
A taxa de erro dos algoritmos de agrupamento para a base de dados espiral dupla é representada na Tabela 1. O melhor resultado (menor taxa de erro) foi alcançado pelo método proposto, em que 0% de taxa de erro foi obtida. Os demais métodos obtiveram taxas de erro superiores a 40%. Os resultados obtidos pelo método proposto utilizaram os parâmetros de 15 segmentos e comprimento de 1,5σ.

Tabela 1 Desempenho dos algoritmos de agrupamento em função da taxa de erro para a base de dados espiral dupla

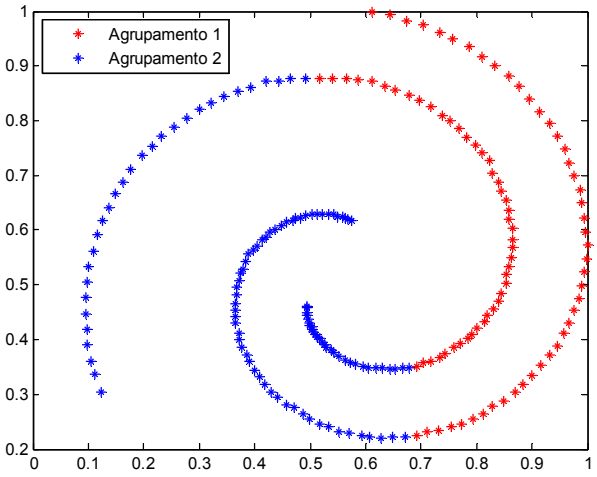
Dados	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede SOM
Espiral Dupla	49,51%	41,74%	0%	52,42%

Deve-se observar que o baixo índice de acerto da rede SOM é devido à distribuição dos neurônios sobre a base de dados (Figura 19(a)), resultando na classificação demonstrada na Figura 19(b).

Os dados da base espiral dupla têm uma característica parecida com a dos *clusters* alongados e esféricos e é uma das bases de dados mais desafiantes para os métodos de agrupamentos, apesar de ser uma base de dados bidimensional. Como os dados são dispostos em fila, a curva principal pôde modelá-los perfeitamente (veja a Figura 16), o que levou ao bom resultado obtido.



(a)



(b)

Figura 19 Resultado da rede SOM para a base de dados espiral dupla: (a) distribuição dos neurônios sobre a base de dados (b) classificação da rede SOM

A Tabela 2 representa os resultados para o teste de generalização dos métodos. Observa-se que o método rede SOM obteve a melhor taxa de erro para os dados de treino e teste. O método proposto obteve resultados próximos ao KM-Euclidiana e KM-Manhattan.

Tabela 2 Capacidade de generalização dos métodos em função da taxa de erro (média ± desvio padrão) para a base de dados espiral dupla

Base Espiral Dupla	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede SOM
Média de acerto (treinamento)	47,98±1,29%	44,49%±4,28%	47,10±2,09%	29,30±15,62%
Média de acerto (teste)	45,21±3,70%	44,73%±3,64%	45,37±3,51%	28,49±15,50%

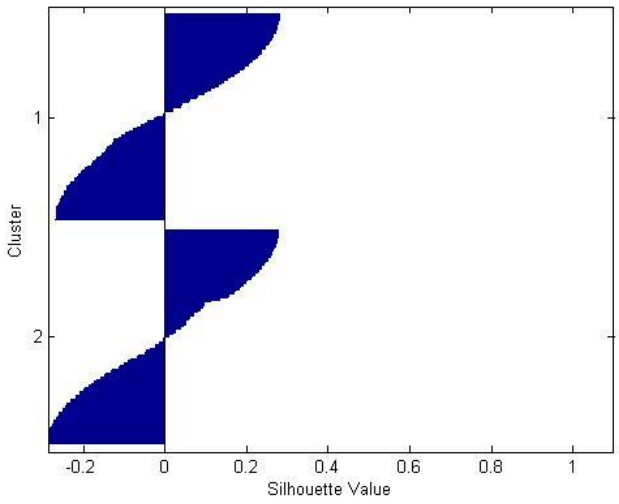
Nota-se na matriz de medidas em (22), obtida para esta base de dados, que a média das distâncias dos eventos à curva que os representa (diagonal principal) é da ordem de 10^{-4}, já que os eventos estão quase todos situados na curva principal, como demostrado na Figura 16. Por outro lado, as medidas extraclasses (valores fora da diagonal principal) mostram o quanto os eventos estão distantes das curvas principais que representam os outros grupos. Isso mostra que os eventos estão bem representados pelo método, o que leva ao índice K_m = 0,0035, calculado conforme (18).

$$\mathbf{M} = \begin{bmatrix} 0{,}0001 & 0{,}0306 \\ 0{,}0189 & 0{,}0001 \end{bmatrix} \tag{22}$$

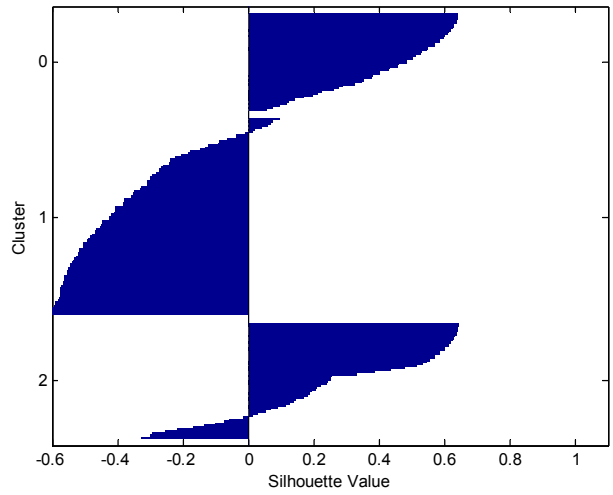
A Figura 20 demonstra o resultado do silhouettes aplicado aos agrupamentos obtidos pelo método proposto e pelos métodos KM-E, KM-M e rede SOM com silhouettes médio (Sil_m) de -0,0019, 0,5498, 0,5079 e -0,0017,

respectivamente. Percebe-se que os valores de silhouettes obtidos para o KM-E são superiores e todos positivos, indicando uma qualidade de agrupamento superior, tendo assim uma maior coesão no interior dos agrupamentos e maior separação intergrupos, embora possua uma taxa de erro maior que a do método proposto. Este resultado é interessante e passa uma ideia de que há um ou mais eventos que estão bem situados em seus grupos, de acordo com o critério do algoritmo *K-means* (menor distância ao centroide), mas que na verdade pertencem a outro agrupamento. Neste caso, o método proposto conseguiu alocar corretamente todos os eventos.

Pode-se entender também, a partir dos resultados da Figura 20, que o silhouettes não é um método de validação de agrupamentos indicado para o tipo de agrupamento espiral dupla, já que pode levar o projetista a interpretações equivocadas acerca dos agrupamentos encontrados.



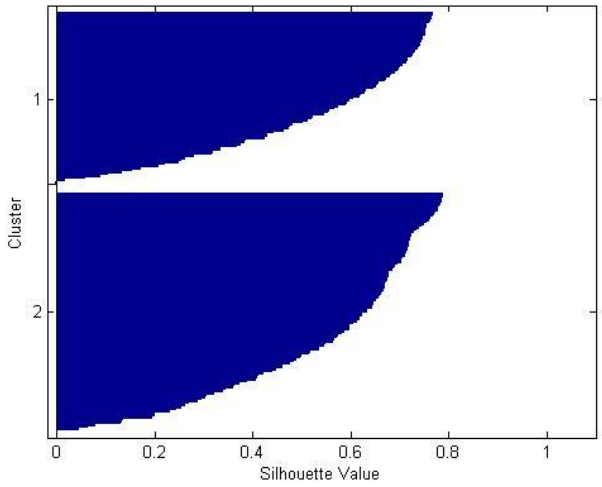
(a)



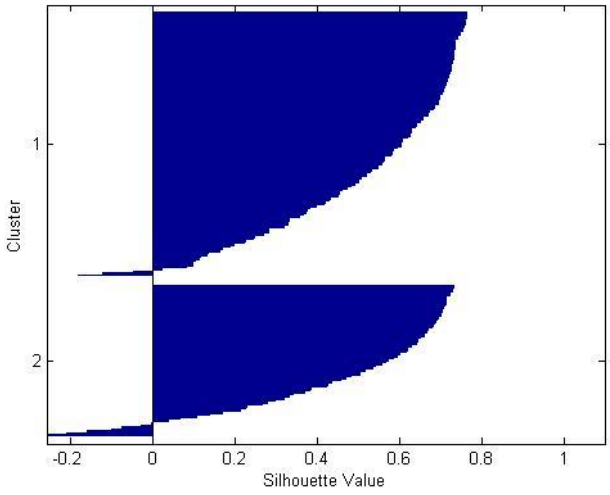
(b)

Figura 20 Silhouettes dos agrupamentos obtidos para a base de dados espiral dupla: (a) com o método proposto; (b) com a rede SOM; (c) com o KM-E; (d) com o KM-M

(...continua...)


Cluster
1
2
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(c)


Cluster
1
2
-0.2
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(d)

5.2 Base de dados *half-rings*

O método proposto foi aplicado à uma base de dados conhecida como *half-rings*, composta por dois agrupamentos e 373 dados no total. A base de dados modelada pela curva principal é ilustrada na Figura 21. Os agrupamentos obtidos pelo método proposto estão demonstrados na Figura 22(a), em que um acerto na classificação dos dados de 100% foi alcançado. A Figura 22(b) representa o resultado do agrupamento de dados alcançado pelo algoritmo *k-means* utilizando a distância Euclidiana.

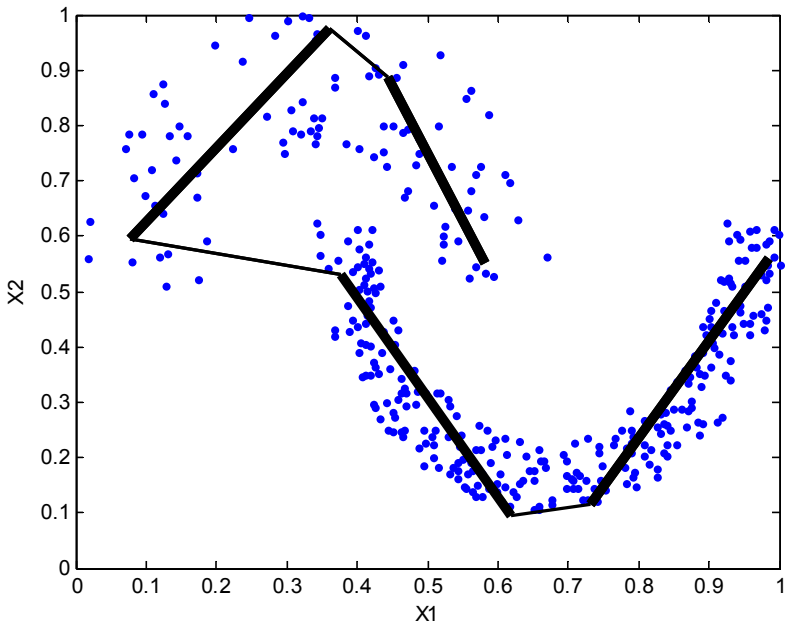
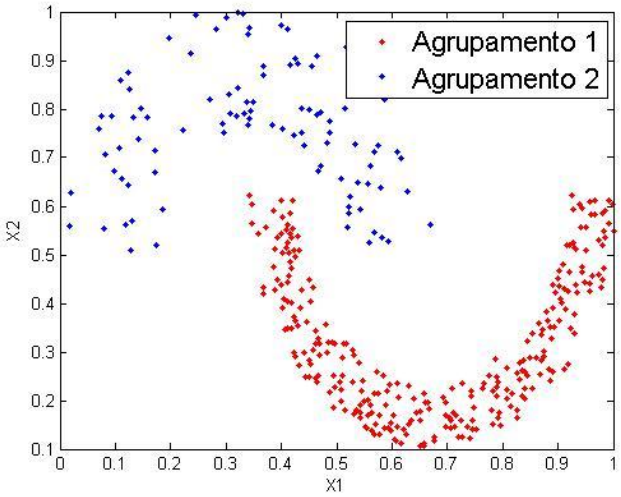
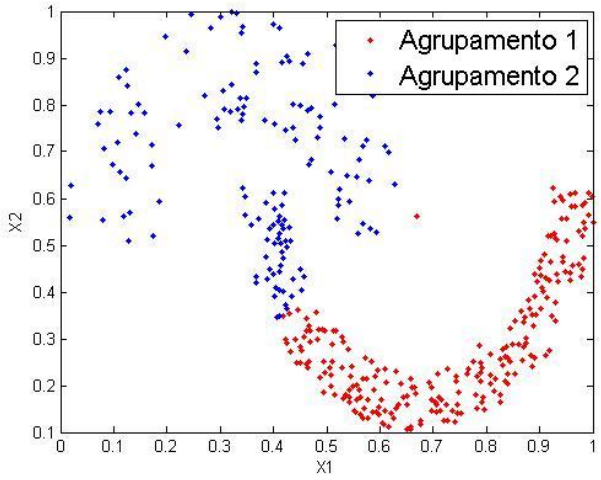


Figura 21 Curva principal original obtida para a base de dados *half-rings*



(a)



(b)

Figura 22 Resultado do método proposto para a base de dados *half-rings*: (a) pelo método proposto (b) pelo método *K-means* com medida de similaridade Euclidiana

A Figura 23 demonstra as distâncias entre os vértices das interligações da curva principal demonstrada na Figura 21. Pode-se concluir a presença de apenas dois agrupamentos, uma vez que um segmento de interligação com maior comprimento em relação aos outros segmentos se destaca.

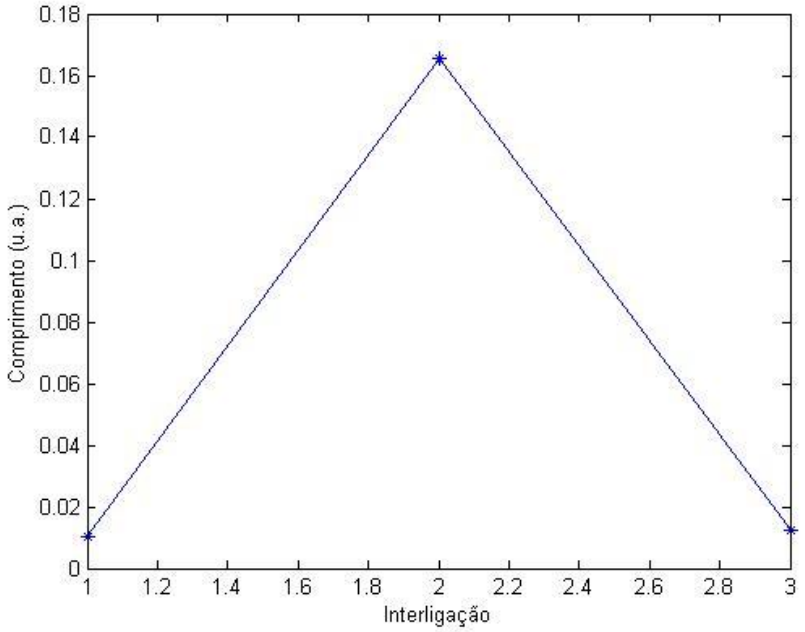


Figura 23 Distância em unidade arbitrária entre os vértices para a base de dados *half-rings*

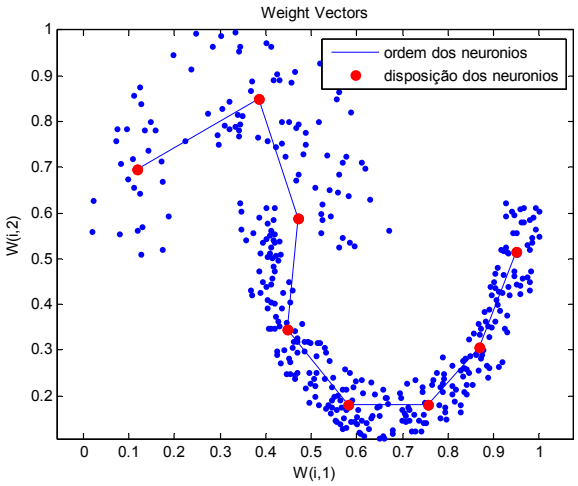
A taxa de erro de algoritmos de agrupamento para a base de dados *half-rings* está representada na Tabela 3. O melhor resultado (menor taxa de erro) foi alcançado pelo método proposto (0% de erro). Os resultados obtidos pelo método proposto utilizaram os parâmetros de 4 segmentos e comprimento de 1,5σ.

Tabela 3 Resultados dos algoritmos de clusterização, taxa de erro para a base de dados *half rings*

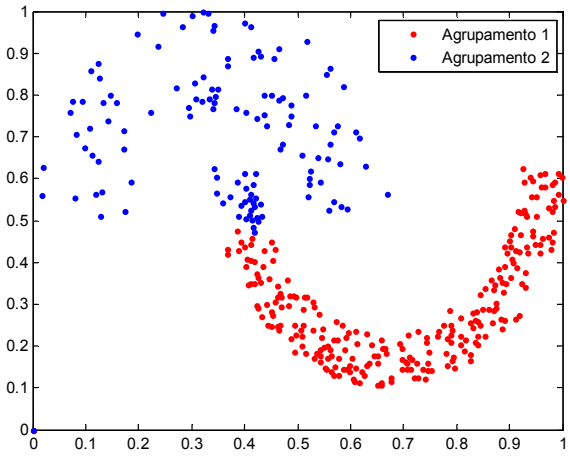
Dados	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede Som
Half-rings	13,40%	17,42%	0%	7,77%

Verifica-se que o erro de 7,77% obtido pela rede SOM é devido à distribuição dos neurônios sobre a base de dados, o que pode ser observado na Figura 24(a), resultando na classificação demonstrada na Figura 24(b). Observa-se que há um neurônio posicionado fora da região dos dados e entre os dois agrupamentos, gerando confusão na classificação.

Os dados da *half-rings* têm uma característica parecida com a dos *clusters* compactos e semiesféricos e está entre as bases de dados mais desafiantes para os métodos de agrupamentos, apesar de ser uma base de dados bidimensional. Como os dados são dispostos em forma de dois semicírculos, a curva principal pôde modelá-los perfeitamente (veja a Figura 22(a)), o que levou ao resultado obtido.



(a)



(b)

Figura 24 Resultado da rede SOM para a base de dados *half rings*: (a) distribuição dos neurônios sobre a base de dados; (b) classificação

Na verificação de generalização do método proposto, observa-se que o resultado para a base de treinamento e teste da rede SOM é superior aos demais métodos. O método proposto apresenta a maior taxa de erro para os dados de treinamento e teste. Na Tabela 4 está representada a taxa de erro.

Tabela 4 Capacidade de generalização dos métodos em função da taxa de erro (média ± desvio padrão) para a base de dados *half rings*

Base *half rings*	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede SOM
Média de acerto (treinamento)	12,95±1,82%	16,33±2,04%	42,52±3,15%	7,24±9,88%
Média de acerto (teste)	13,09±2,63%	15,02±2,49%	42,91±4,47%	7,05±9,49%

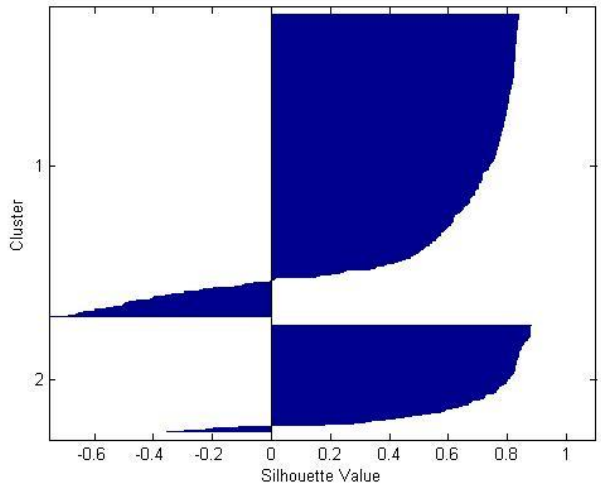
Percebe-se que na matriz de medidas em (23), obtida para esta base de dados, a média das distâncias dos eventos à curva que os representa (diagonal principal) é da ordem de 10^{-3}, já que os eventos estão quase todos situados ao redor da curva principal, como demonstrado na Figura 22. Por outro lado, as medidas extraclasses (valores fora da diagonal principal) mostram o quanto os eventos estão distantes das curvas principais que representam os outros grupos. Isso mostra que os eventos estão bem representados pelo método, o que leva ao índice K_m = 0,0261, calculado conforme (18).

$$\mathbf{M} = \begin{bmatrix} 0{,}0015 & 0{,}1213 \\ 0{,}0970 & 0{,}0042 \end{bmatrix} \tag{23}$$

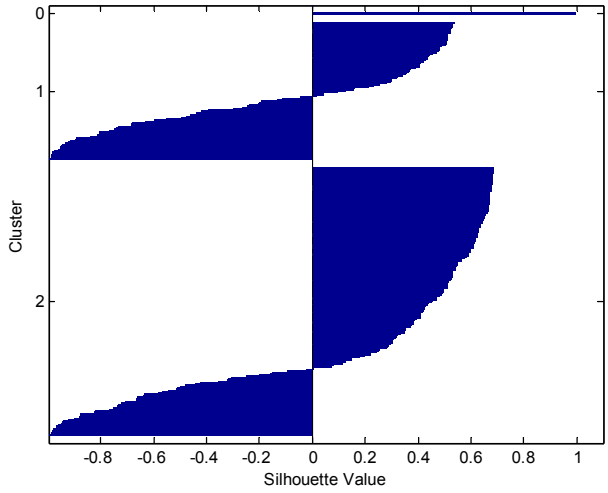
A Figura 25 demonstra o resultado do silhouettes aplicado aos agrupamentos obtidos pelo método proposto e pelos métodos KM-Euclidiana, KM-Manhattan e rede SOM com silhouettes médios (Sil_m) de 0,1160, 0,6853, 0,6460 e 0,1160, respectivamente. Observa-se que os valores de silhouettes

obtidos para o KM-Euclidiana são superiores e todos positivos, indicando uma qualidade de agrupamento superior, tendo assim uma maior coesão no interior dos agrupamentos e maior separação intergrupos, embora possua uma taxa de erro maior que a do método proposto. Este resultado é interessante e passa uma ideia de que há um ou mais eventos que estão bem situados em seus grupos, de acordo com o critério do algoritmo *K-means* (menor distância ao centroide), mas que na verdade pertencem a outro agrupamento. Neste caso, o método proposto conseguiu alocar corretamente todos os eventos.

Pode-se entender também, a partir dos resultados da Figura 25, que o silhouettes não é um método de validação de agrupamentos indicado para o tipo de agrupamento espiral dupla, já que pode levar o projetista a interpretações equivocadas acerca dos agrupamentos encontrados.



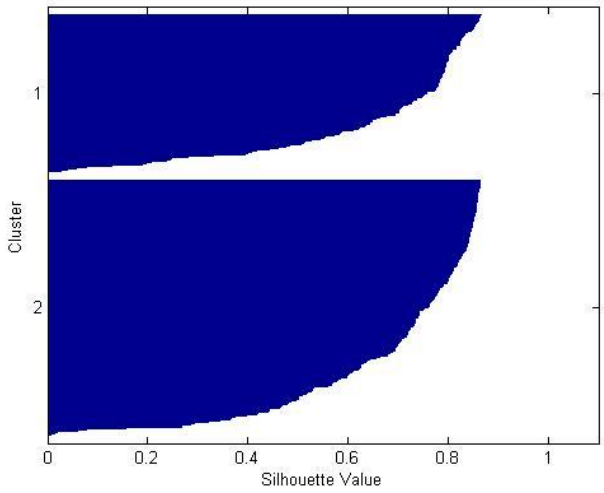
(a)



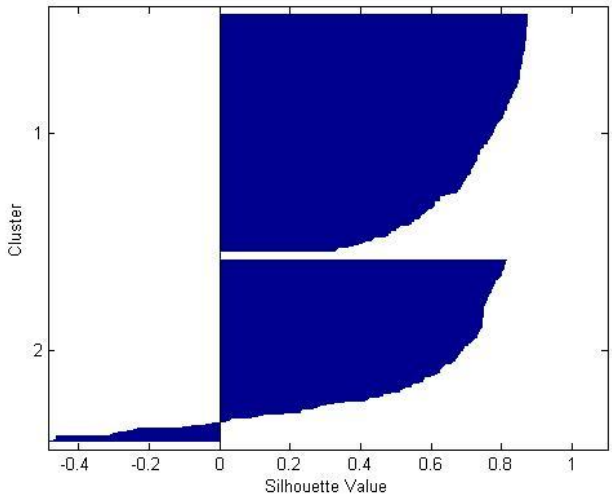
(b)

Figura 25 Silhouettes dos agrupamentos obtidos para a base de dados *half-rings*: (a) com o método proposto; (b) com a rede SOM; (c) com o KM-E; (d) com o KM-M

(...continua...)


1
2
Cluster
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(c)


1
2
Cluster
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(d)

5.3 Base de dados Iris

A base de dados conhecida como Iris é uma base de dados muito conhecida em problemas de reconhecimento de padrões, ela é composta por três classes referentes a três espécies de plantas (i*ris setosa, iris virginica e iris versicolor*), cada agrupamento possui 50 espécies com quatro características: comprimento e largura das sépalas e das pétalas em centímetros.

A Tabela 5 demonstra a matriz de confusão dos resultados obtidos pelo método proposto, em que o erro total foi de 3,33%. A Tabela 6 demonstra a comparação dos resultados com os métodos *k-means* e SOM. Para esta base de dados a rede SOM obteve o melhor resultado, com taxa de erro de 0%. Os resultados obtidos pelo método proposto utilizaram os parâmetros de 3 segmentos e comprimento de 0,9σ.

Tabela 5 Matriz de confusão obtida pelo método proposto para a base de dados Iris

Classe	1	2	3
1	100%	0%	0%
2	0%	98%	2%
3	0%	2%	98%
		Erro total: 3,33%	

Tabela 6 Desempenho dos algoritmos de agrupamento em função da taxa de erro para a base de dados Iris

Dados	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede Som
Iris	4%	6%	3,33%	0%

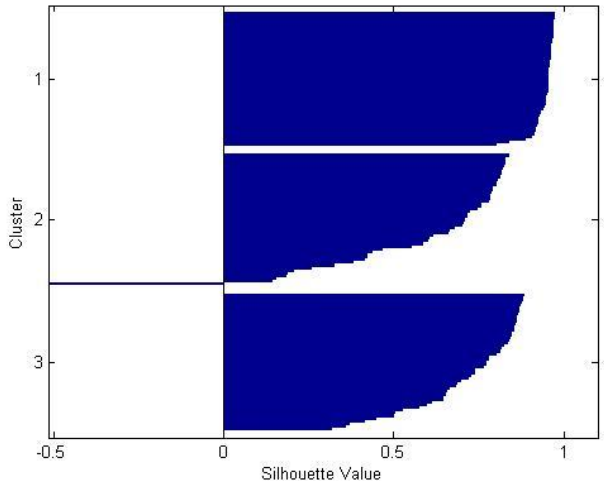
Na verificação de generalização do método proposto, apresentado na Tabela 7, observa-se que o resultado para a base de treinamento e teste o método rede SOM é superior aos demais métodos, entretanto seu desvio padrão é o alto. Nota-se que a capacidade de generalização do método proposto obtém o segundo melhor resultado, superando os métodos KM-Euclidiana e KM Manhattan.

Tabela 7 Capacidade de generalização dos métodos em função da taxa de erro (média ± desvio padrão) para a base de dados Iris

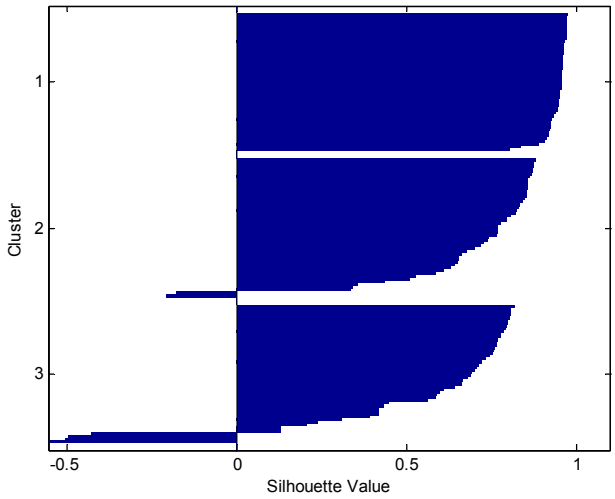
Base Iris	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede SOM
Média de acerto (treinamento)	74,38±25,88%	71,96±24,70%	67,30±5,07%	33,84±23,65%
Média de acerto (teste)	75,25±9,25%	72,74±25,13%	65,92±9,25%	33,33±23,19%

Nota-se na Figura 26 que, de acordo com o índice de silhouettes, o KM-E e o método proposto têm uma qualidade de agrupamento superior. Um pequeno conjunto de eventos foi agrupado de forma incorreta pelo método proposto.

A fim de analisar os resultados de forma mais quantitativa, o silhouettes médio foi calculado para cada agrupamento, de acordo com a Equação (7). Foram obtidos silhouettes médios (Sil_m) de 0,7565, 0,7648, 0,7033 e 0,7250 para o método proposto, para o método KM-E, KM-M e para a rede SOM, respectivamente. Estes resultados mostram que os quatro algoritmos obtiveram agrupamentos fortes, com destaques para o KM-E e o método proposto.



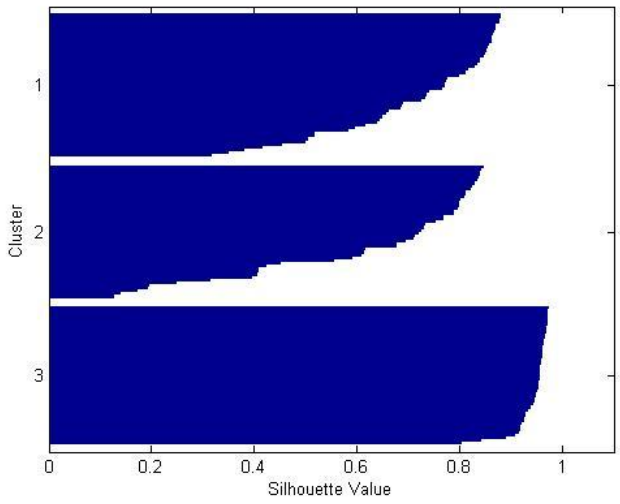
(a)



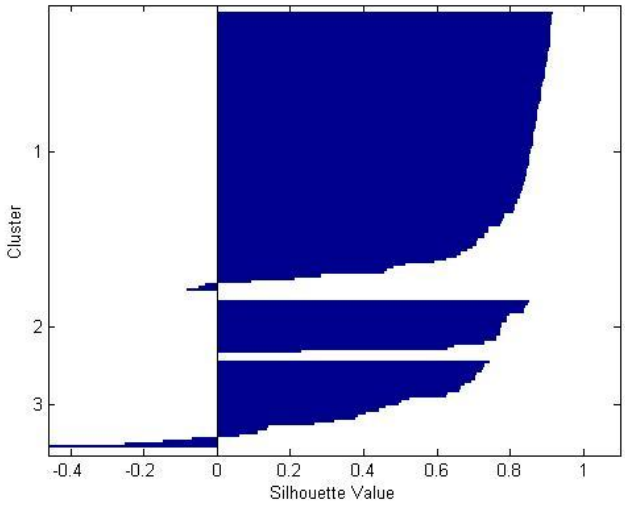
(b)

Figura 26 Silhouettes dos agrupamentos obtidos para a base de dados Iris: (a) com o método proposto; (b) com a rede SOM; (c) com o KM-E; (d) com o KM-M

(...continua...)


Cluster
1
2
3
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(c)


Cluster
1
2
3
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(d)

Observa-se na matriz de medidas em (24) que a média das distâncias dos eventos à curva que os representa (diagonal principal) é menor do que os demais valores, indicando que os eventos estão distantes das curvas principais que representam os outros grupos. O índice K_m obtido a partir de (24) é 0,0203.

$$\mathbf{M} = \begin{bmatrix} 0{,}0046 & 0{,}7641 & 0{,}3182 \\ 0{,}9285 & 0{,}0136 & 0{,}0988 \\ 0{,}4022 & 0{,}0876 & 0{,}0081 \end{bmatrix} \tag{24}$$

A Figura 27 demonstra o comprimento de todas as interligações dos segmentos. Observa-se que as interligações representadas no eixo das abscissas pelos números 1 e 2 são as únicas presentes na base de dados. Isso se dá pelo fato de o método proposto ter obtido o melhor resultado com a configuração de três segmentos, consequentemente tem-se apenas duas interligações.

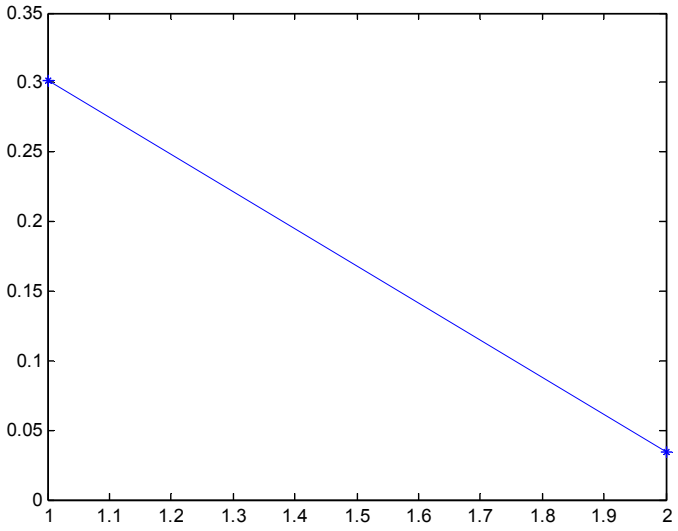


Figura 27 Distância em unidade arbitrária entre os vértices para a base de dados Iris

5.4 Base de dados Diabetes

A base de dados conhecida como Diabetes, construída na década de 90, é baseada na análise de exames de pacientes da região de Phoenix, Arizona, EUA. Ela é composta por 2 classes, diabético e não diabético. Cada agrupamento possui 500 e 268 eventos, respectivamente. Esta base de dados possui 8 atributos: Número de vezes que engravidou, concentração de glicose posteriormente à ingestão de carboidratos, pressão arterial, espessura da prega cutânea, índice de massa corporal, idade, índice de insulina posterior à injeção de carboidratos e questão genética.

A Tabela 8 demonstra a matriz de confusão obtida pelo método proposto, em que um erro total de 33,98% foi obtido. A Tabela 9 demonstra a comparação com os outros métodos. Neste caso o método proposto foi ligeiramente inferior comparado ao KM-E. A rede SOM obteve o melhor desempenho, com taxa de erro de apenas 1,82%. Os resultados obtidos pelo método proposto utilizaram os parâmetros de 12 segmentos e comprimento de 1,0σ.

Tabela 8 Matriz de confusão obtida pelo método proposto para a base de dados Diabetes

Classe	1	2
1	75,8%	24,2%
2	52,23%	47,76%
	Erro total: 33,98%	

Tabela 9 Desempenho dos algoritmos de agrupamento em função da taxa de erro (%) para a base de dados Diabetes

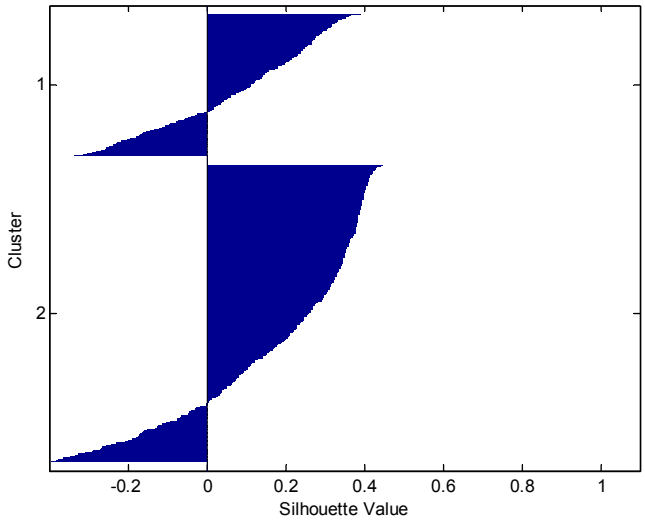
Dados	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede Som
Diabetes	33,07%	35,80%	33,98%	1,82%

Na verificação de generalização do método proposto, apresentado na Tabela 10, observa-se que o resultado para a base de treinamento e teste do método KM-Euclidiana é superior aos demais métodos. O método proposto obtém a maior taxa de erro entre os métodos analisados.

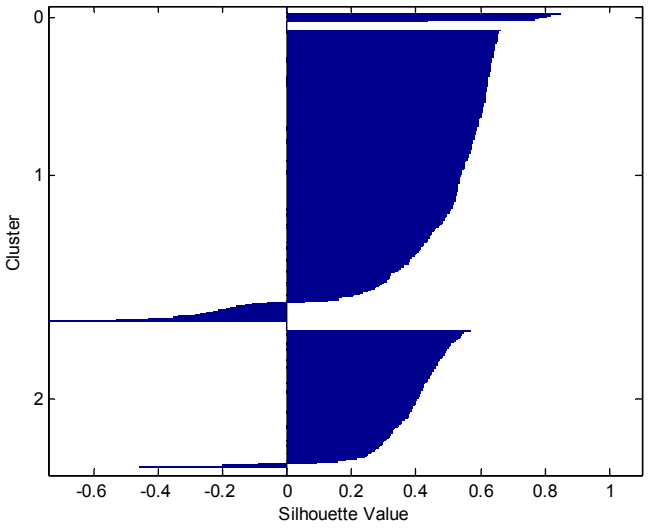
Tabela 10 Capacidade de generalização dos métodos em função da taxa de erro (média ± desvio padrão) para a base de dados Diabetes

Base Diabetes	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede SOM
Média de acerto (treinamento)	32,66±1,52%	37,42±4,64%	42,29±4,29%	36,71±3,65%
Média de acerto (teste)	33,13±2,43%	39,86±4,94%	41,76±5,04%	36,88±4,72%

Verifica-se na Figura 28 que de acordo com o índice de silhouettes o KM-M obteve o melhor resultado com silhouettes médio (Sil_m) de 0,2606. O método proposto obteve um silhouettes médio 0,1516, e o KM-E de 0,3651. O silhouettes médio para a rede SOM foi de 0,4433. Assim, o KM-M alcançou maior coesão no interior dos agrupamentos e maior separação intergrupos em comparação com os outros métodos, do ponto de vista do silhouettes.



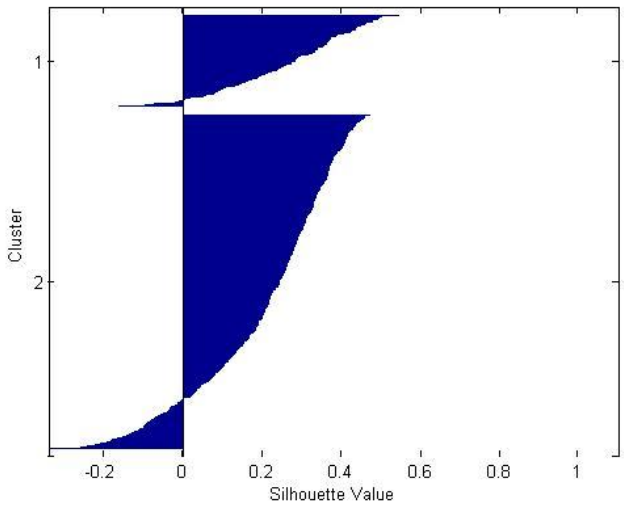
(a)



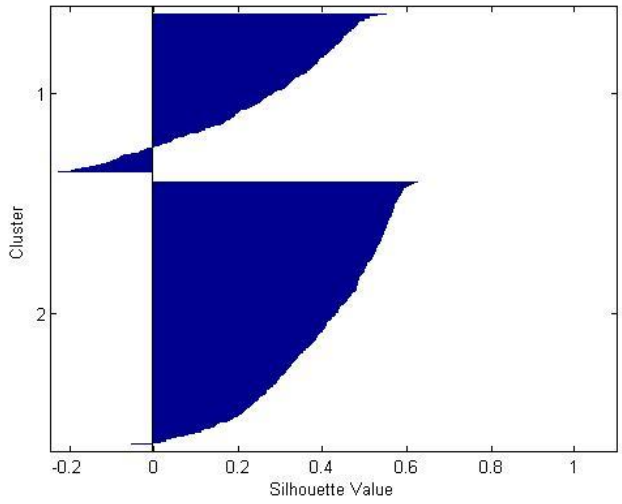
(b)

Figura 28 Silhouettes dos agrupamentos obtidos para a base de dados Iris: (a) com o método proposto; (b) com a rede SOM; (c) com o KM-Euclidiana; (d) com o KM-Manhattan

(...continua...)


Cluster
1
2
-0.2
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(c)


Cluster
1
2
-0.2
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(d)

$$\mathbf{M} = \begin{bmatrix} 0{,}0619 & 0{,}1316 \\ 0{,}1610 & 0{,}0476 \end{bmatrix} \tag{25}$$

Observa-se na matriz de medidas em (25) que a média das distâncias dos eventos à curva que os representa (diagonal principal) é claramente menor do que os demais valores, indicando que os eventos estão distantes das curvas principais que representam os outros grupos e próximos da curva principal que os representa. Apesar de o método proposto obter um erro de classificação de 33,98%, pode-se afirmar que os dados estão bem representados pelas curvas principais obtidas. O índice K_m obtido a partir de (25) é 0,3686.

A Figura 29 demonstra o comprimento de todas as interligações dos segmentos. Observa-se que as interligações representadas no eixo das abscissas pelos números 6 e 11 apresentam o comprimento maior, e que as demais interligações apresentam comprimentos relativamente próximos entre si e inferiores aos comprimentos de números 6 e 11. Este é um indicativo de que há três agrupamentos no banco de dados. Seguindo essa abordagem, a separação encontrada pelos três agrupamentos leva um erro de aproximadamente 33,85 % no que tange a separação entre diabéticos e não diabético. Os resultados demonstrados na Tabela 5 foram obtidos considerando-se dois agrupamentos na aplicação do método proposto, em que apenas a interligação de número 6 (Figura 29) foi eliminada.

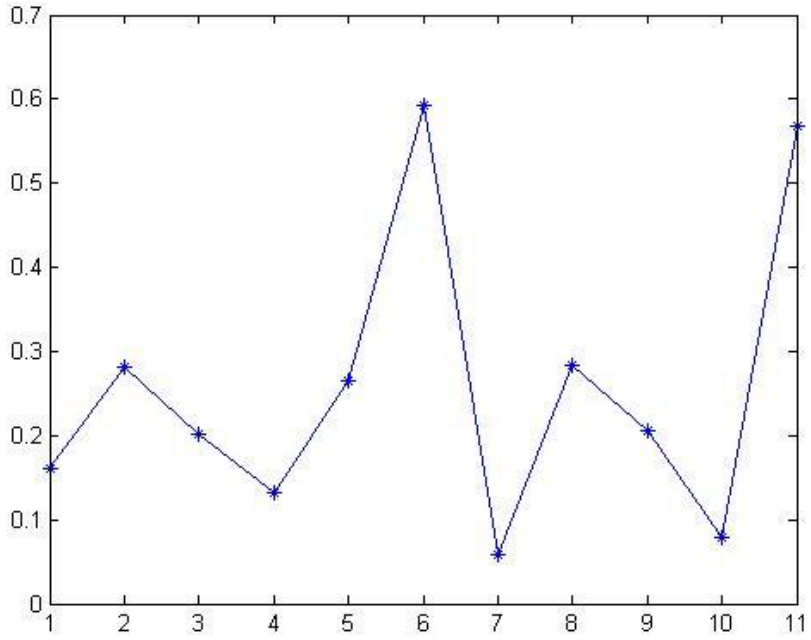


Figura 29 Distância em unidade arbitrária entre os vértices para a base de dados Diabetes

5.5 Base de dados *Wine*

A base de dados conhecida como *Wine* é resultado de uma análise de vinhos cultivados em uma mesma região da Itália e é composta por três classes, 178 dados no total, sendo que a classe 1 possui 59 dados, a classe 2 possui 71 dados e a classe 3 possui 48 dados. A base de dados possui 13 atributos: taxa alcoólica, taxa de ácido málico, taxa de ash, taxa de alcalinidade, taxa de magnésio, total de fenóis, taxa de flavonoides, taxa de fenóis não flavonoides, taxa de proantocianidinas, intensidade da cor, taxa de HUE, taxa de diluição e taxa de prolina.

A Tabela 11 representa a matriz de confusão obtida com o método proposto, em que o erro foi de apenas 7,86%. A Tabela 12 demonstra a comparação com os outros métodos. Os resultados obtidos pelo método proposto utilizaram os parâmetros de 3 segmentos e comprimento de 0,2σ.

Tabela 11 Matriz de confusão obtida pelo método proposto para a base de dados *Wine*

Classe	1	2	3
1	100%	0%	0%
2	2,81%	81,69%	15,49%
3	2,09%	0%	97,91%
	Erro total: 7,86%		

Tabela 12 Desempenho dos algoritmos de agrupamento em função da taxa de erro (%) para a base *Wine*

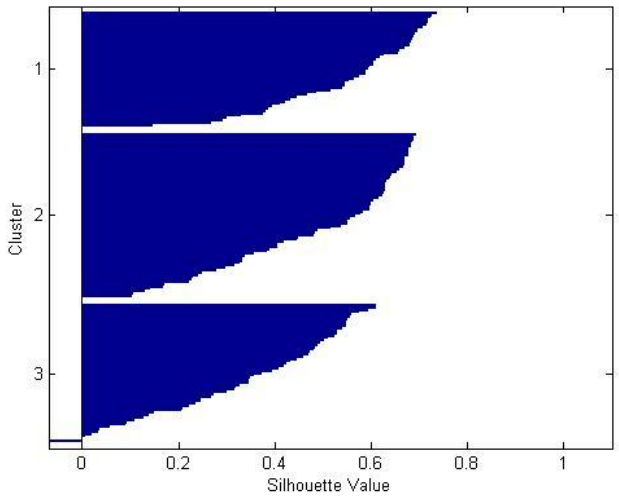
Dados	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede SOM
Wine	29,77%	6,74%	7,86%	0%

Na verificação de generalização do método proposto, representado na Tabela 13, observa-se que os resultados de todos os métodos foram muito próximos e com uma taxa de erro elevada.

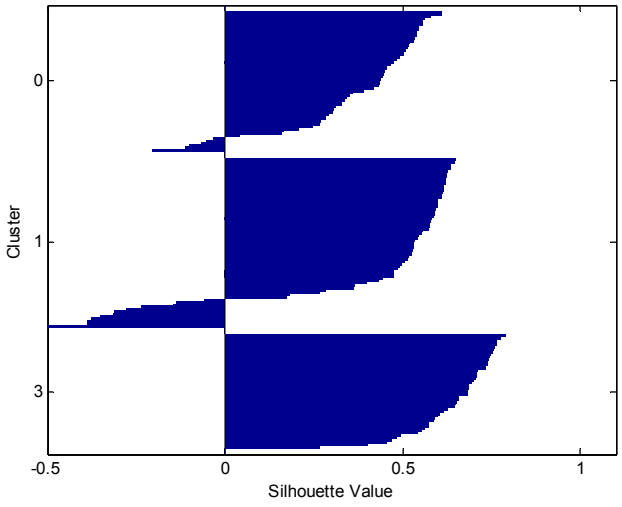
Tabela 13 Capacidade de generalização dos métodos em função da taxa de erro (média ± desvio padrão) para a base de dados *Wine*

Base *Wine*	KM-Euclidiana	KM-Manhattan	Método Proposto	Rede SOM
Média de acerto (treinamento)	67,66±7,52%	63,36±4,64%	66,63±4,70%	66,16±5,57%
Média de acerto (teste)	69,13±8,43%	62,65±4,94%	66,41±7,89%	67,09±8,72%

Observa-se na Figura 30, que de acordo com o índice de silhouettes o método proposto e o KM-Manhattan possuem uma qualidade de agrupamento superior em comparação com os outros métodos, tendo assim uma maior coesão no interior dos agrupamentos e maior separação intergrupos, com silhouettes médio 0,4618 para o método proposto e de 0,4554 para o KM-Manhattan. O silhouettes médio para o KM-Euclidiana foi Sil_m = 0,1625 e o silhouettes médio para a rede SOM Sil_m = 0,4526.



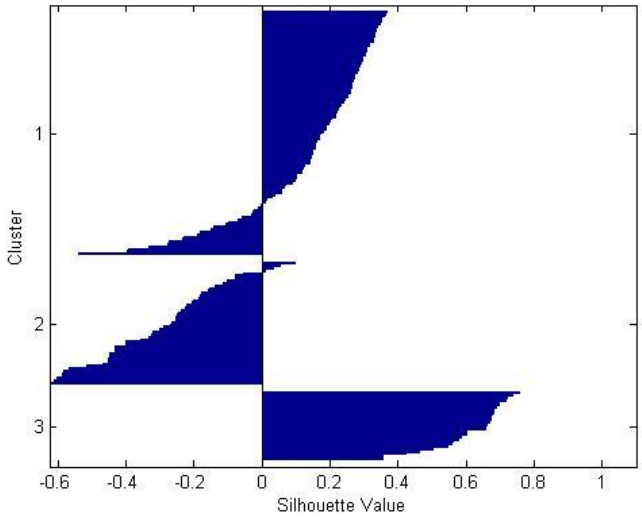
(a)



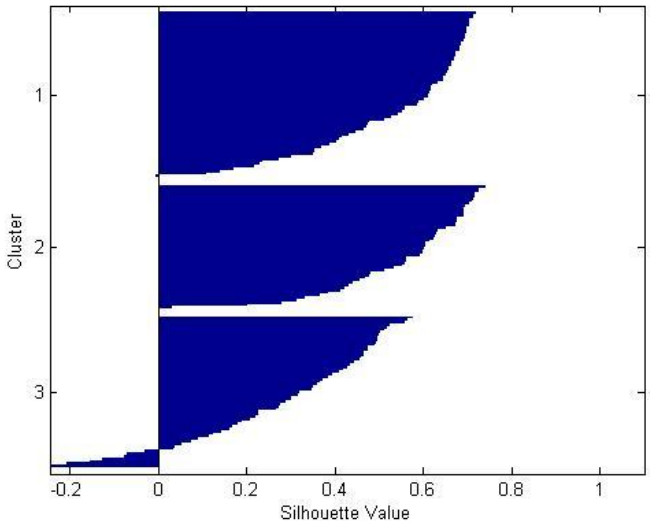
(b)

Figura 30 Silhouettes para o método proposto com a base de dados *Wine*: (a) para o método proposto; (b) com a rede SOM; (c) para o KM-Euclidiana; (d) para o KM-Manhattan

(...continua...)


Cluster
1
2
3
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(c)


Cluster
1
2
3
-0.2
0
0.2
0.4
0.6
0.8
1
Silhouette Value

(d)

Nota-se na matriz de medidas (26) que a média das distâncias dos eventos à curva que os representa é da ordem de quatro vezes menor que as medidas fora da diagonal, levando a um K_m de 0,2307.

$$\mathbf{M} = \begin{bmatrix} 0{,}1354 & 0{,}8716 & 0{,}5287 \\ 0{,}8698 & 0{,}1319 & 0{,}4181 \\ 0{,}5562 & 0{,}4348 & 0{,}1570 \end{bmatrix} \tag{26}$$

A Figura 31 demonstra o comprimento de todas as interligações dos segmentos. Observa-se que as interligações representadas no eixo das abscissas pelo número 1 e 2 são as únicas presentes na base de dados, devido ao fato de o método proposto ter obtido este resultado com a configuração de três segmentos, consequentemente, existem apenas duas interligações.

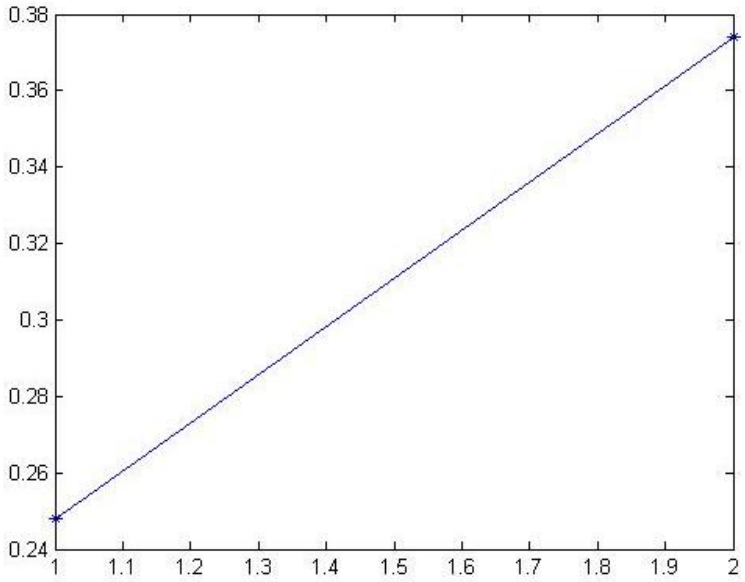


Figura 31 Distância em unidade arbitrária entre os vértices para a base de dados *Wine*

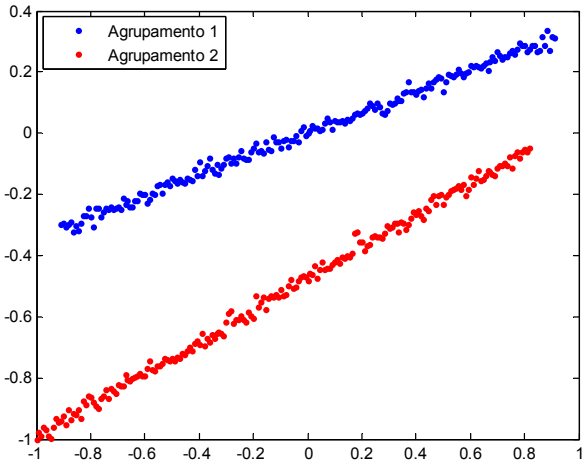
5.6 Análise do comprimento do segmento e número de segmentos

A fim de avaliar a influência do comprimento do segmento (*c*) no desempenho do algoritmo, foi verificado o número de acertos do método proposto para os três tipos de agrupamentos definidos na Seção 2.4: *clusters* alongados, esféricos e compactos. Consideraram-se os comprimentos iguais a 0,1, 0,3, 0,5, 1,0, 1,5, 2,0 e 2,5 do desvio padrão dos dados ao longo do segmento.

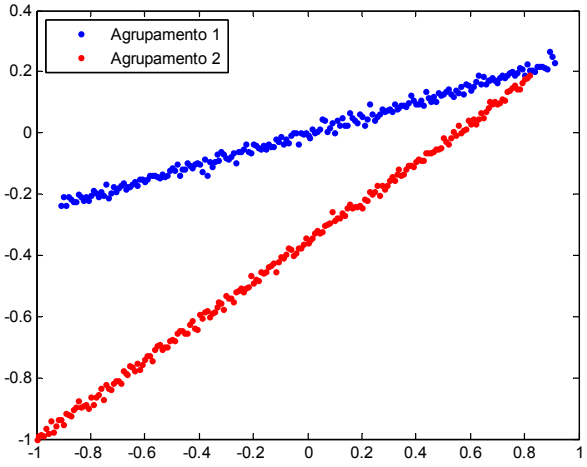
Foi variado também o número de segmentos a ser utilizado, visto que este pode também oferecer influência nos resultados. Verbeek, Vlassis e Krose (2002) mencionam que na grande maioria das aplicações não é possível saber com antecedência qual é o número ideal de segmentos para modelar corretamente a base de dados. Em problemas supervisionados, o conjunto de dados de treinamento pode ser usado para ajustar o número de segmentos ideal para o banco de dados. Em Ferreira et al. (2013, 2015) essa abordagem foi empregada para a detecção e classificação de distúrbios elétricos em sistemas de potência com bastante sucesso.

5.6.1 *Clusters* alongados

A Figura 32 demonstra duas bases de dados compostas por dois *clusters* alongados, cada um com 201 eventos, sem interseção (a) e com interseção (b) de *clusters*.



(a)



(b)

Figura 32 Distribuição dos dados para uma base de dados com *clusters* alongados

As Figuras 33 e 34 demonstram o percentual de acertos do método proposto para a base de *clusters* alongados sem interseção (Figura 32(a)) em função do número de segmentos para os diferentes comprimentos de segmento (*c*). Observa-se que quanto menor o comprimento do segmento, mais segmentos podem ser incluídos na CP até que o algoritmo k-seg convirja (Seção 2.6.2). Como as informações da CP ficam armazenadas em duas matrizes, *edges* e *vertices*, que contêm as informações dos vértices que compõem a CP e de como eles estão interligados, é fácil ver que mais segmentos implicam no aumento da dimensão dessas duas matrizes e, consequentemente, em aumento de complexidade e de memória de *hardware* para armazenamento e processamento de dados. Portanto, a obtenção de curvas principais com menor número de segmentos é desejável do ponto de vista de implementação prática do método. Por outro lado, poucos segmentos podem levar a um projeto subdimensionado e prejudicar a eficiência do método.

As Figuras 33 e 34 demonstram que para a base de dados de *clusters* alongados, acertos de 100% podem ser alcançados utilizando-se segmentos menores ou iguais a 1,5σ para os *clusters* da Figura 32(a), em que σ é o desvio padrão associado à componente. Para comprimentos de 0,1σ e 0,3σ, acertos de 100% são alcançados mais vezes com a variação do número de segmentos. Para os demais comprimentos de segmento, o número de acertos apresenta certa variação em função do número de segmentos. Todavia, como esta base de dados é uma base bidimensional, podem-se ajustar os parâmetros da CP observando-se o melhor ajuste aos dados, o que facilita o uso do método proposto.

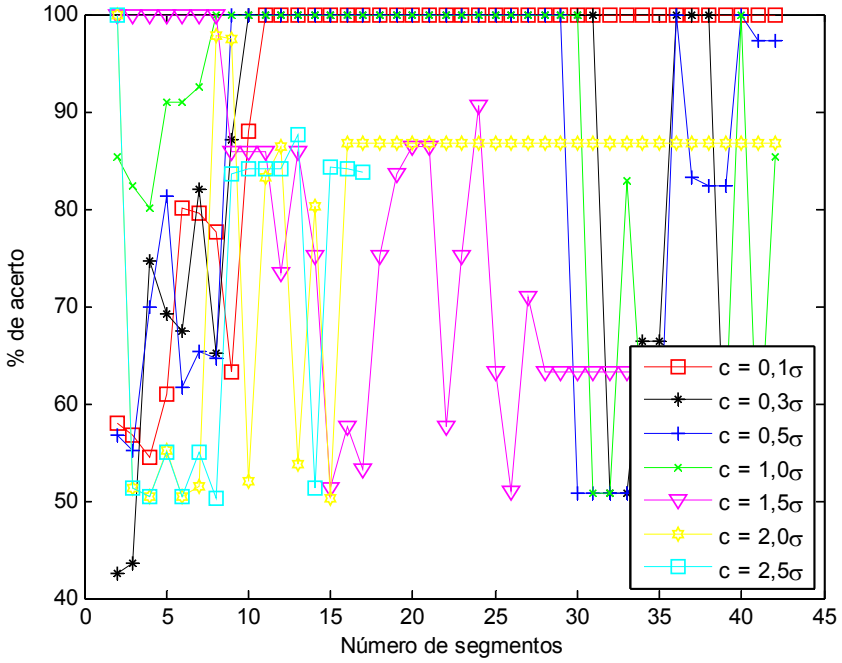


Figura 33 Taxa de acerto em função do comprimento e número de segmentos para o agrupamento alongado sem interseção

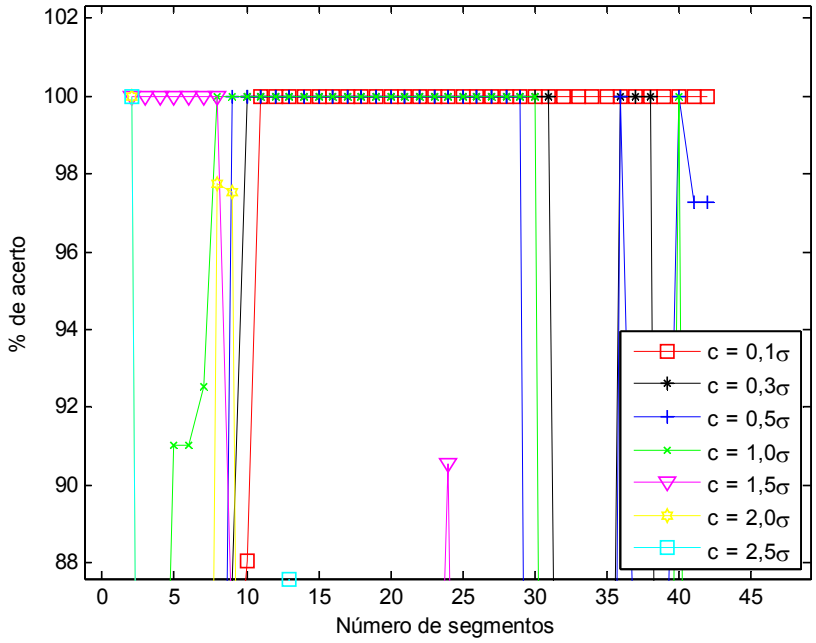


Figura 34 Taxa de acerto em função do comprimento e número de segmentos para o agrupamento alongado sem interseção (Zoom da Figura 33)

A Figura 35 demonstra o valor do índice K_m em função do número de segmentos para os *clusters* da Figura 32(a). Observa-se que os valores de K_m aproximam-se de 0 à medida que o número de segmentos aumenta.

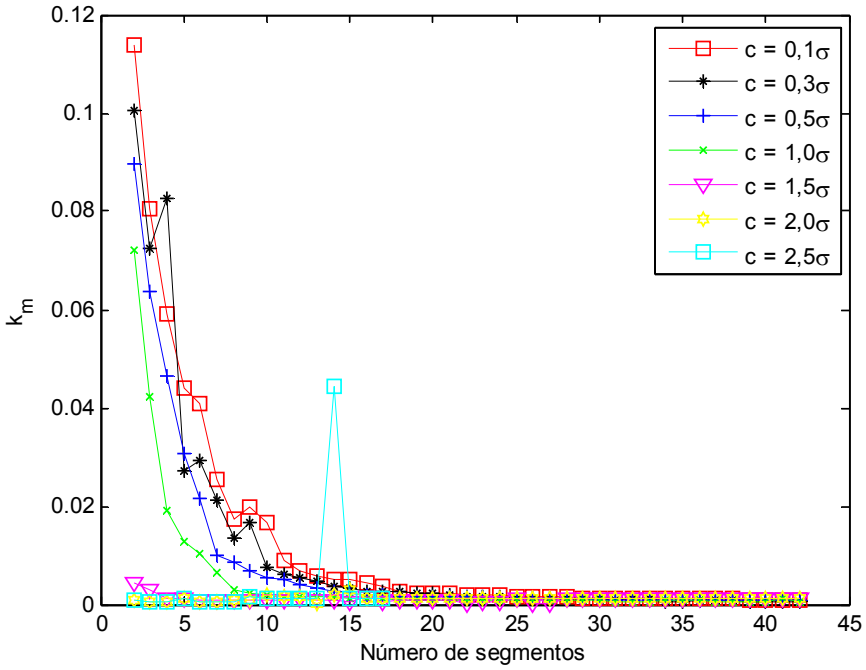


Figura 35 Valor de K_m em função do número de segmentos para o agrupamento alongado sem interseção

A Figura 36, que representa um zoom da Figura 35, demonstrando mais claramente que os menores valores de K_m, que são indicativos de agrupamentos mais fortes, são encontrados utilizando segmentos mais longos.

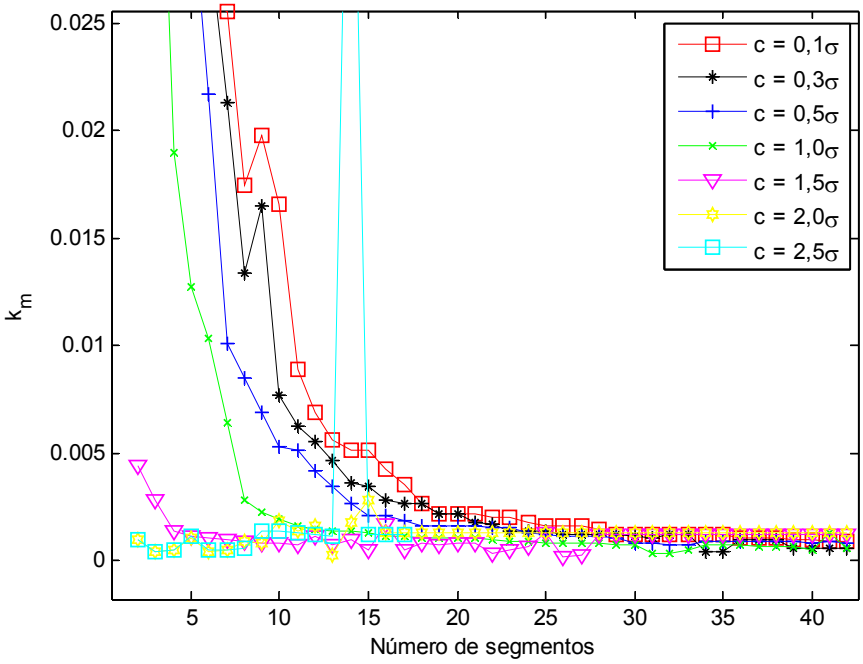


Figura 36 Valor de K_m em função do número de segmentos para o agrupamento alongado (Zoom da Figura 35)

O fato de menores valores de K_m serem encontrados em segmentos mais longos se deve à característica do agrupamento, que se trata de agrupamentos alongados, e devido ao fato de que a distância dos dados à curva principal, implementada por Verbeek, Vlassis e Krose (2002), é medida em relação ao segmento e não leva em consideração as interligações de segmento. Portanto, segmentos mais curtos geram mais interligações, o que faz com que os dados que estiverem mais próximos das interligações apresentem maiores distâncias à CP. A distância do evento à CP é calculada conforme descrito a seguir.

A Figura 37 demonstra dois eventos $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$, definidos em $\Re^2$ e representados por dois pontos no plano cartesiano, e uma curva poligonal, também definida no plano, a qual possui apenas dois segmentos s_1 e s_2, sendo o

primeiro definido pelos vértices v_1 e v_2 e o segundo por v_2 e v_3. Nota-se que os dois eventos estão mais próximos do primeiro segmento, portanto, é a partir dele que devem ser obtidas as distâncias. O evento $\boldsymbol{x}_1$ projeta, ortogonalmente, sobre o primeiro segmento. Neste caso, a distância é medida do ponto de projeção ao evento em questão. O evento $\boldsymbol{x}_2$, no entanto, não projeta sobre o segmento, sendo o vértice v_1 o ponto pertencente ao segmento que está mais próximo dele. Neste caso, a distância será tomada entre o evento $\boldsymbol{x}_2$ e o vértice v_1.

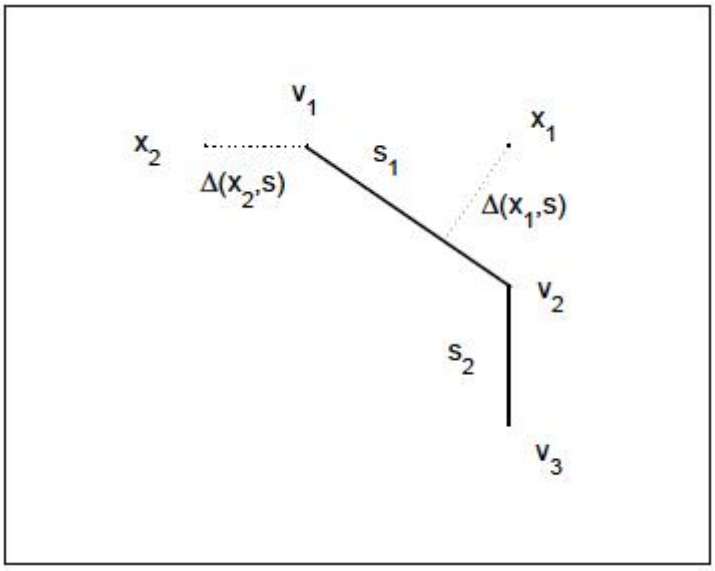


Figura 37 Distância de um ponto a uma curva poligonal

Matematicamente, a distância do evento **x** ao segmento **s** é definida como

$$\Delta(\boldsymbol{x},\boldsymbol{s}) = \begin{cases} ||\mathbf{x} - v_1||^2 & Se\ \mathbf{S}\left(t_s(\boldsymbol{X})\right) = v_1, \\ ||\mathbf{x} - v_2||^2 & Se\ \mathbf{S}\left(t_s(\boldsymbol{X})\right) = v_2, \\ ||\mathbf{x} - \boldsymbol{c}||^2 - \left((\mathbf{x}-\mathbf{c})^{\mathrm{T}}\mathbf{u}\right)^2 & para\ os\ demais\ casos, \end{cases} \quad (27)$$

A Figura 38 demonstra o percentual de acertos do método proposto em função do número de segmentos para os diferentes comprimentos de segmento (*c*) para os *clusters* da Figura 32(b). Embora o resultado varie muito em função do comprimento do segmento (*c*), pode-se concluir que o método proposto apresenta, de modo geral, maior taxa de acerto para comprimentos até 1,5σ com número de segmentos até 12, após 12 segmentos a melhor taxa de acerto é obtida com 1,0σ de comprimento. Os resultados obtidos para o índice K_m, demonstrados nas Figuras 41 e 42, em que se observa os melhores resultados para 2,0σ 1,5σ e 1,0σ de comprimento, o que demonstra, de modo geral, coerência com a taxa de acerto.

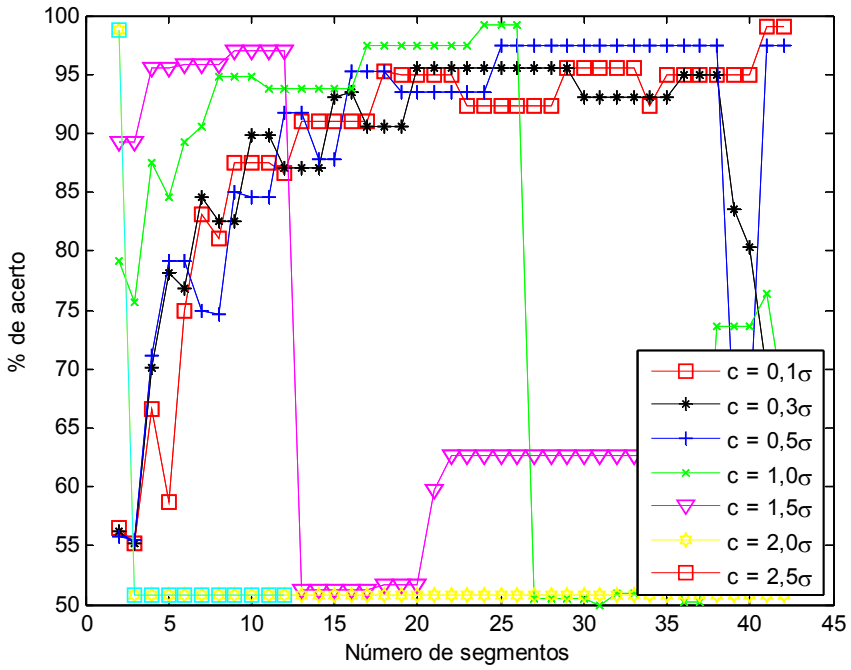


Figura 38 Taxa de acerto em função do comprimento e número de segmentos para o agrupamento alongado com interseção

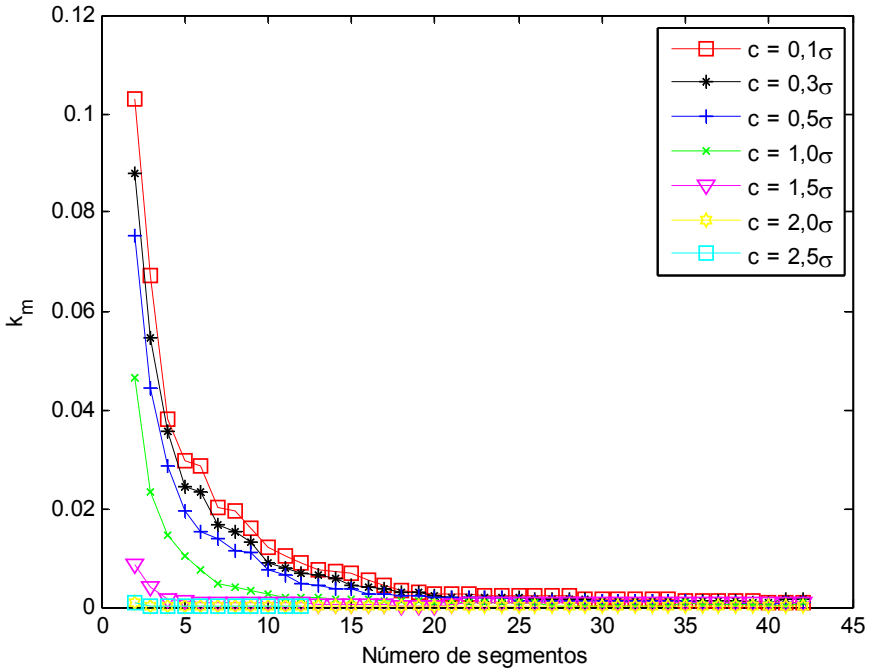


Figura 39 Valor de K_m em função do número de segmentos para o agrupamento alongado com interseção

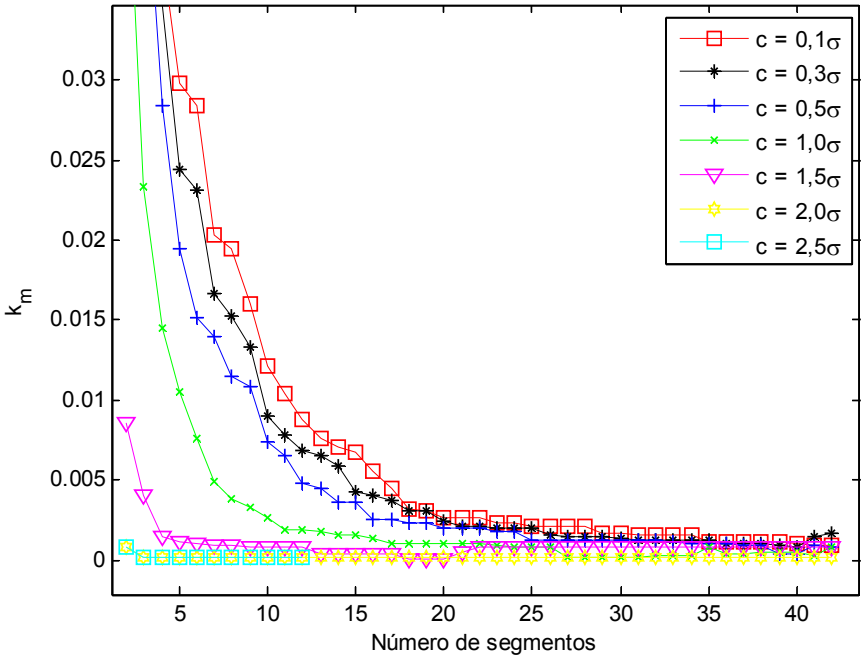
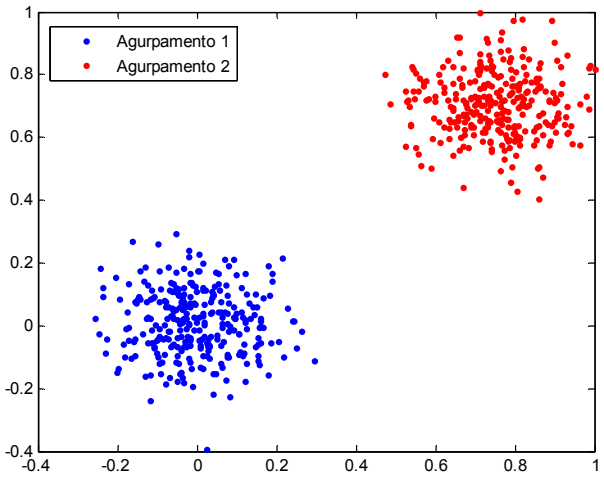


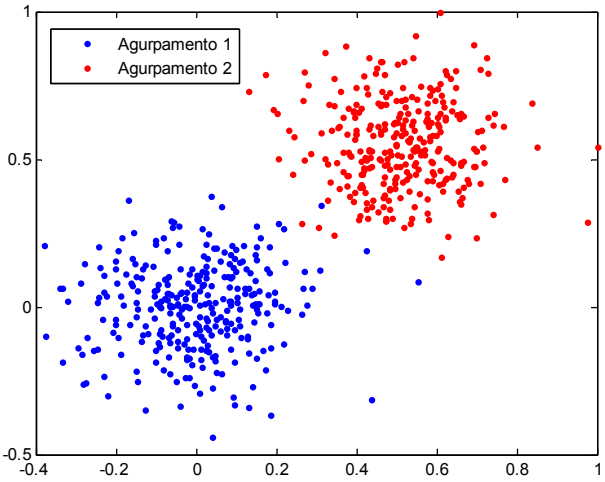
Figura 40 Valor de K_m em função do número de segmentos para o agrupamento alongado com interseção (Zoom da Figura 39)

5.6.2 *Clusters* compactos

Para este tipo de *cluster* foi utilizada a base de dados com as distribuições demonstradas na Figura 41. Ambas as bases são compostas por dois *clusters* de 300 eventos cada. Os agrupamentos demonstrados em (a) estão mais distantes um do outro no espaço, enquanto que os agrupamentos demonstrados em (b) estão mais próximos.



(a)



(b)

Figura 41 Distribuição dos dados para uma base de dados com *clusters* compactos

As Figuras 42 e 43 demonstram o percentual de acertos do método proposto em função do número de segmentos para os diferentes comprimentos de segmento (*c*) para os *clusters* da Figura 41(a). Uma maior taxa de acerto para comprimentos de segmento 0,3σ, 0,5σ, 1,0σ e 1,5σ é obtida como pode ser observado na Figura 42 e 43, em que 100% de acerto é alcançado para diferentes números de segmentos.

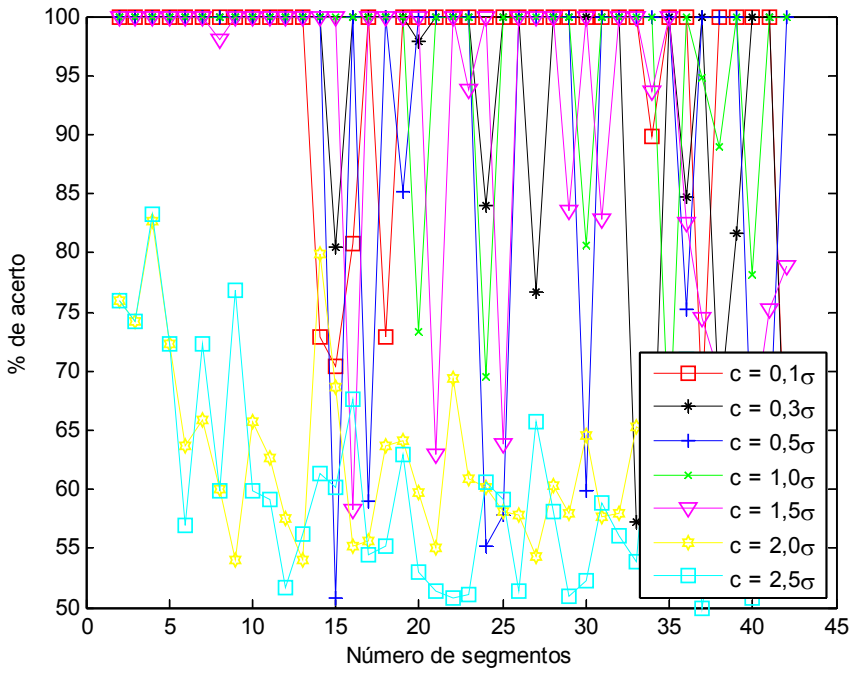


Figura 42 Taxa de acerto em função do comprimento e número de segmentos para a base de dados *clusters* compactos mostrada na Figura 41(a)

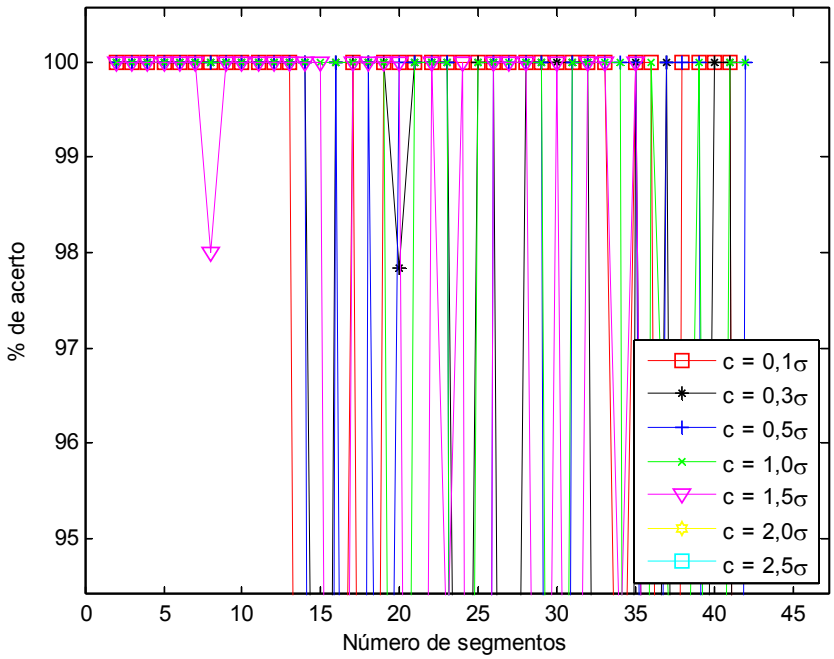


Figura 43 Taxa de acerto em função do comprimento e número de segmentos para a base de dados *clusters* compactos distantes (Zoom da Figura 42)

Os menores índices de K_m são obtidos para os comprimentos 1,0σ e 1,5σ, conforme demonstram as Figuras 44 e 45. Observa-se que o índice K_m apresenta coerência com os resultados de desempenho, visto que os comprimentos de segmento que apresentam os melhores desempenhos levam aos menores índices K_m. Observa-se que para 15 segmentos e c = 0,5σ o desempenho é de aproximadamente 50%. A Figura 46 demonstra um aumento no índice K_m para este mesmo comprimento de segmento e 15 segmentos. Para este tipo de *cluster*, o efeito em K_m do número de interligações de segmento não afeta tanto os resultados quanto para o *cluster* alongado.

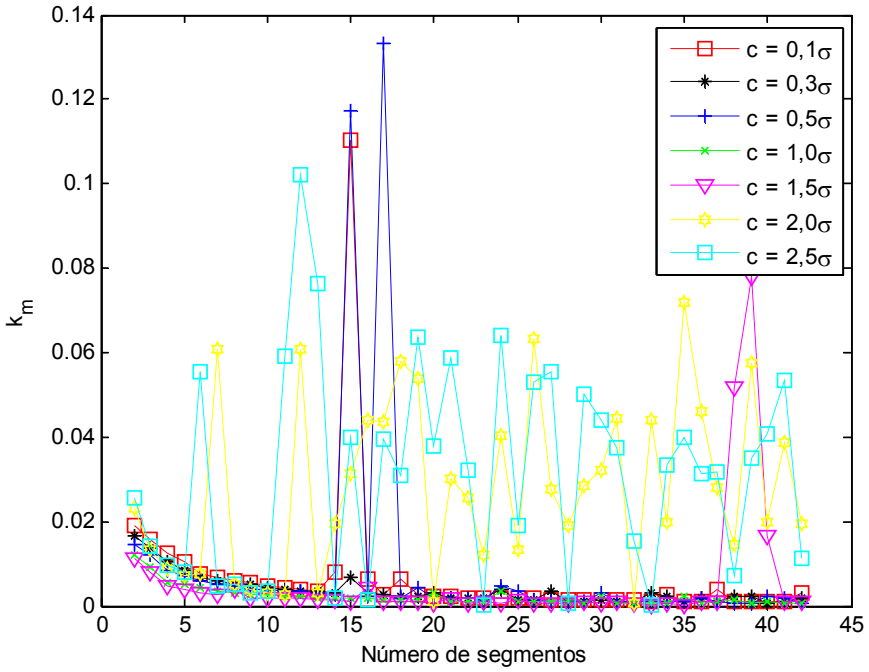


Figura 44 Valor de K_m em função do número de segmentos para a base de dados *clusters* compactos demonstrada na Figura 41(a)

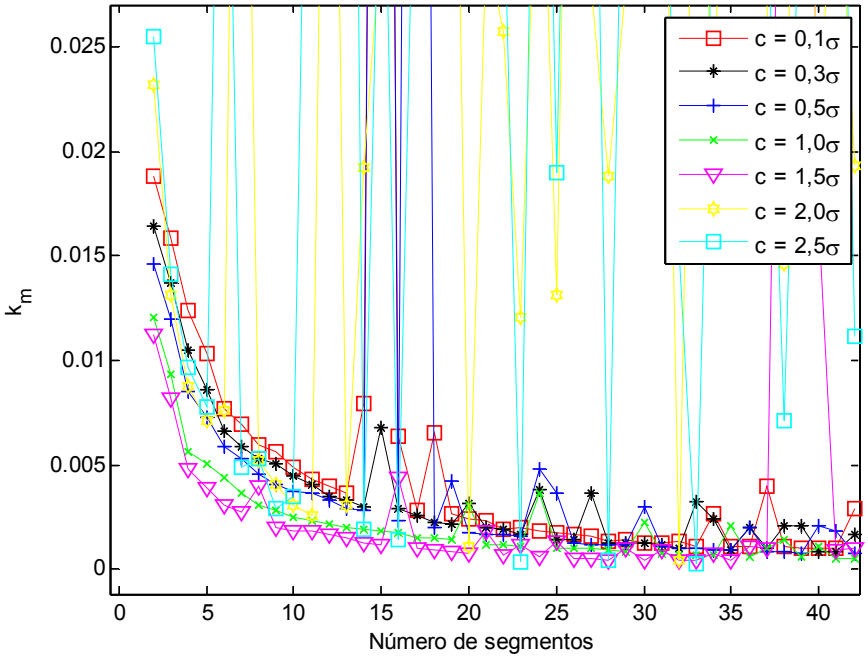


Figura 45 Valor de K_m em função do número de segmentos para a base de dados *clusters* compactos (Zoom da Figura 45)

As Figuras 46 e 48 demonstram o percentual de acertos do método proposto em função do número de segmentos para os diferentes comprimentos de segmento (*c*) para os *clusters* da Figura 41(b).

Nota-se o aumento da variação do desempenho para os diferentes comprimentos em função do número de comprimentos. As melhores taxas de acerto foram obtidas com comprimentos do segmento iguais a 0,1σ, 0,5σ e 1,0σ para poucos segmentos (de 2 a 13), embora o resultado não alcance 100% de acerto. Isso se dá devido à pequena distância entre as classes, em que o grau de dificuldade em se modelar cada uma delas aumenta e, portanto, o desempenho diminui.

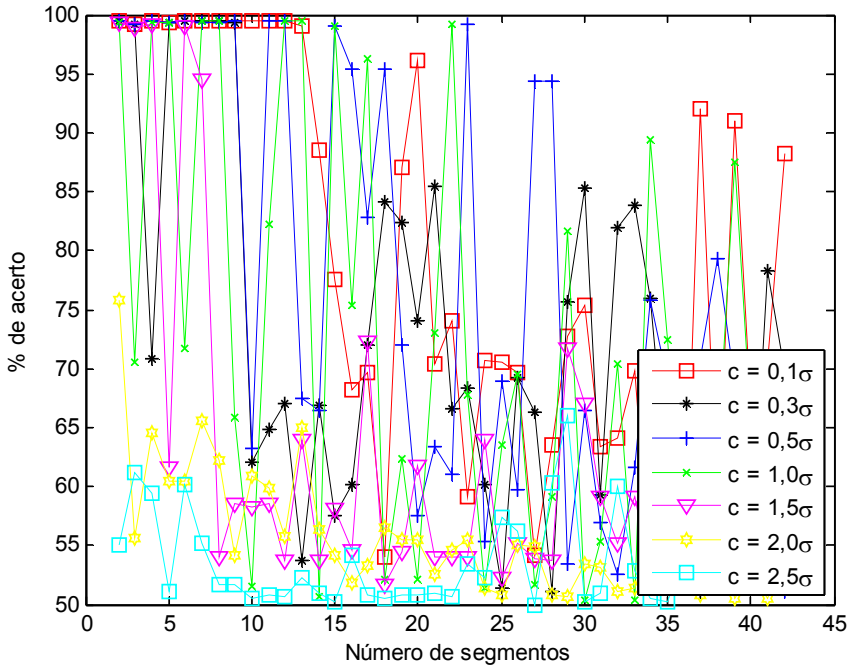


Figura 46 Taxa de acerto em função do comprimento e número de segmentos para a base de dados *clusters* compactos demonstrados na Figura 41(b)

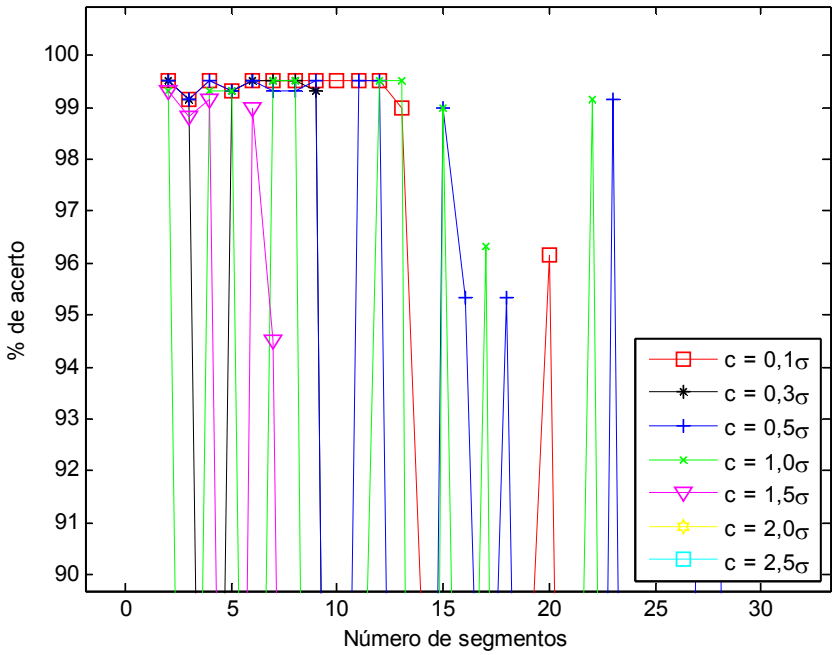


Figura 47 Taxa de acerto em função do comprimento e número de segmentos para a base de dados *clusters* compactos (Zoom da Figura 46)

Os menores índices K_m são encontrados para comprimentos de 1,5σ, que leva a um desempenho da ordem de 99% com poucos segmentos (vide Figura 46), conforme Figuras 48 e 49.

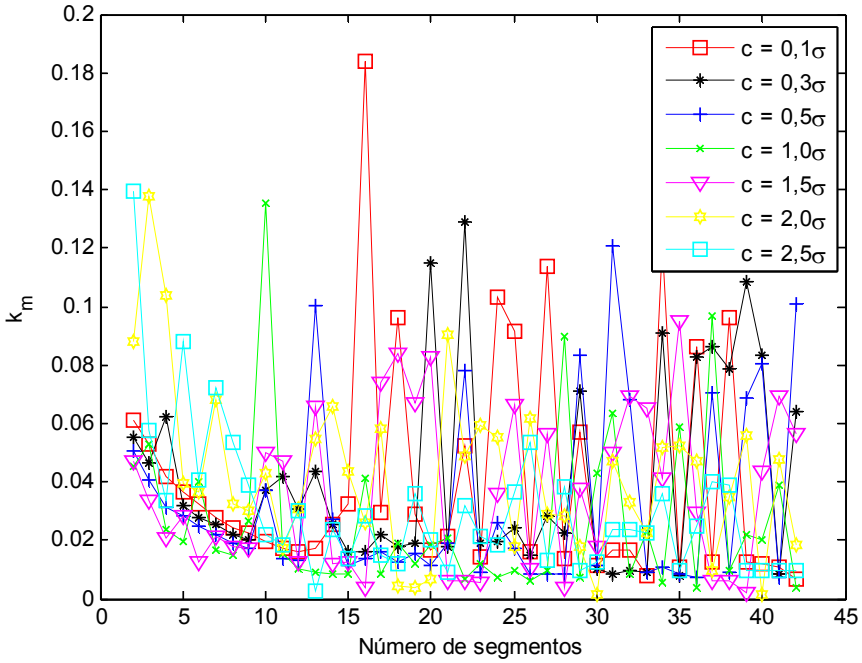


Figura 48 Valor de K_m em função do número de segmentos para a base de dados *clusters* compactos demonstrados na Figura 41(b)

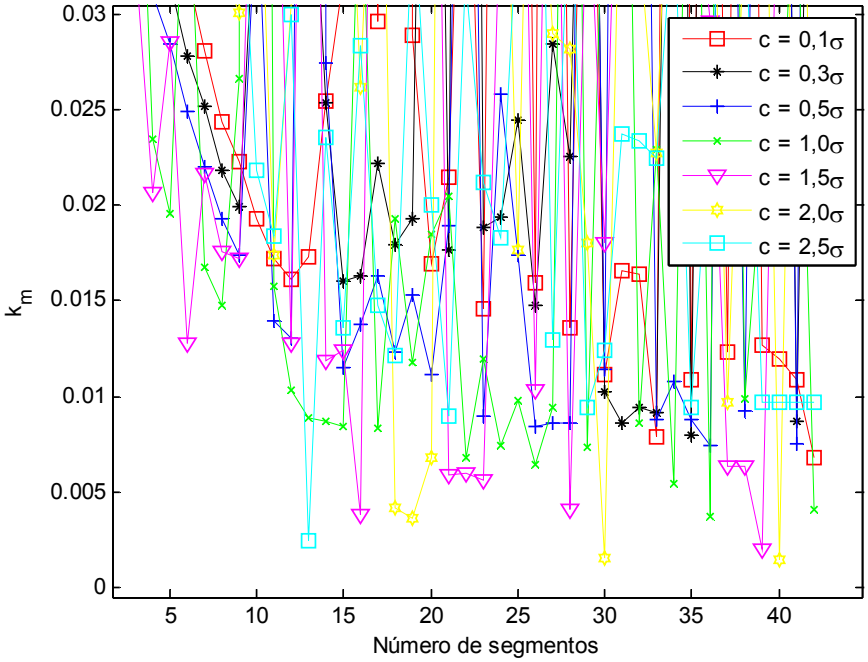
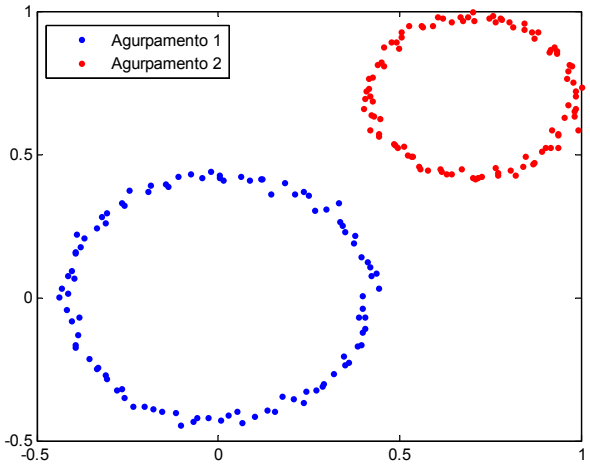


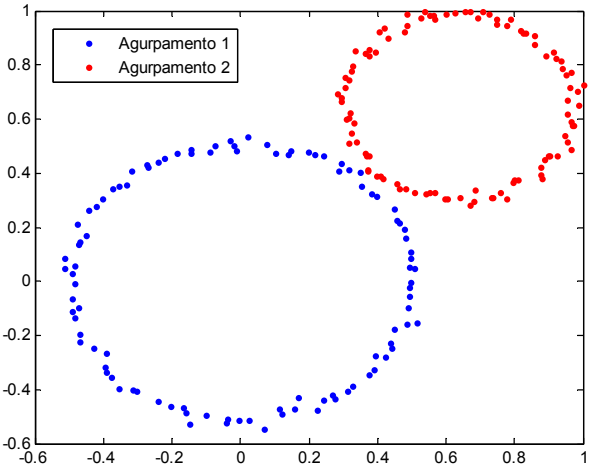
Figura 49 Valor de K_m em função do número de segmentos para a base de dados *clusters* compactos (Zoom da Figura 48)

5.6.3 *Cluster* esféricos

Para este tipo de *cluster* foi utilizada a base de dados com a distribuição conforme a Figura 50, que demonstra duas bases de dados: *clusters* distantes (a) e *clusters* próximos (b). Cada base de dados é composta por dois *clusters* de 101 eventos cada.



(a)



(b)

Figura 50 Distribuição dos dados para uma base de dados com *clusters* esféricos

A Figura 51 demonstra o percentual de acertos do método proposto em função do número de segmentos para os diferentes comprimentos de segmento (*c*) para os *clusters* da Figura 50(a). As maiores taxas de acerto foram obtidas para comprimentos de segmentos de 0,1σ, 0,5σ e 1.0 σ para números de segmentos maiores do que 15, como pode ser observado nas Figuras 51 e 52.

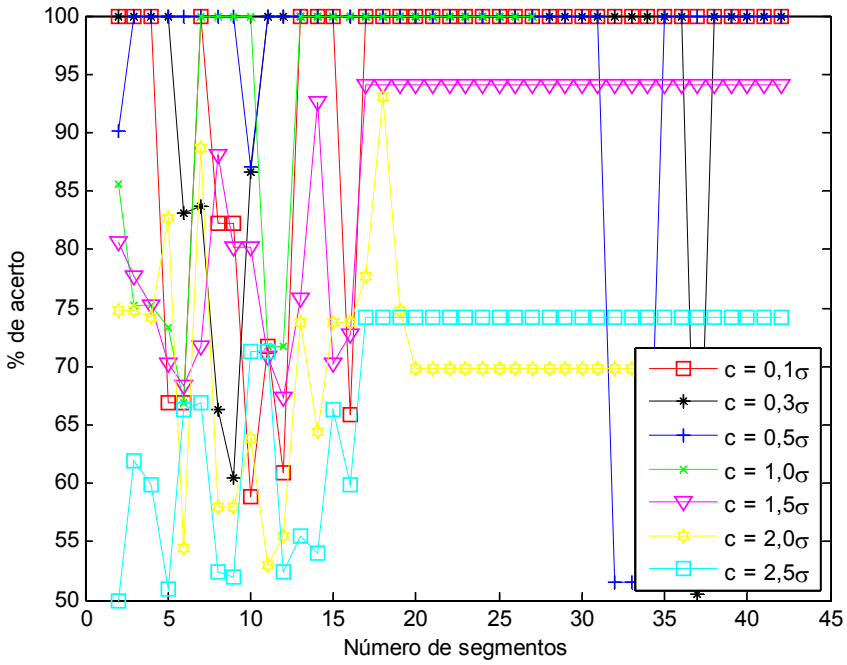


Figura 51 Taxa de acerto em função do comprimento e número de segmentos para a base de dados *clusters* circulares demonstrados na Figura 50(a)

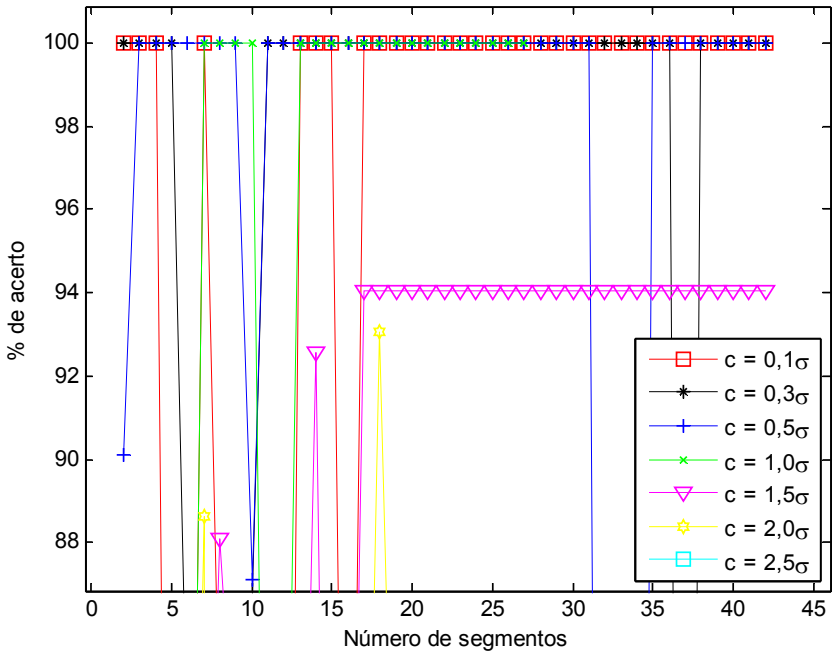


Figura 52 Taxa de acerto em função do comprimento e número de segmentos para a base de dados *clusters* circulares (Zoom da Figura 51)

Observamos na Figura 53 e 54 que o índice K_m apresenta certa coerência com os resultados de desempenho, visto que o comprimento de segmento , c = 1,0σ apresenta uma das melhores taxas de acerto.

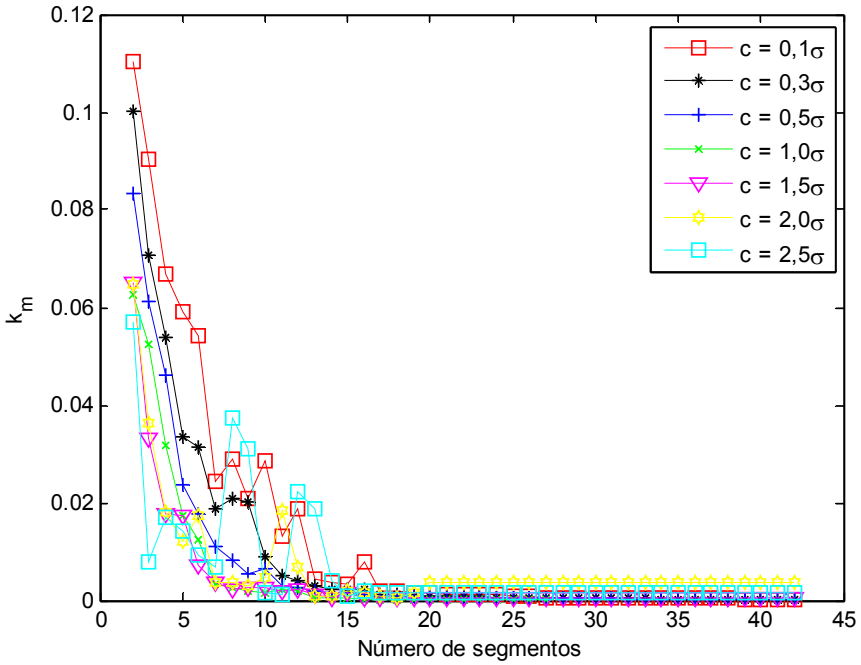


Figura 53 Valor de K_m em função do número de segmentos para a base de dados *clusters* circulares demonstrada na Figura 50(a)

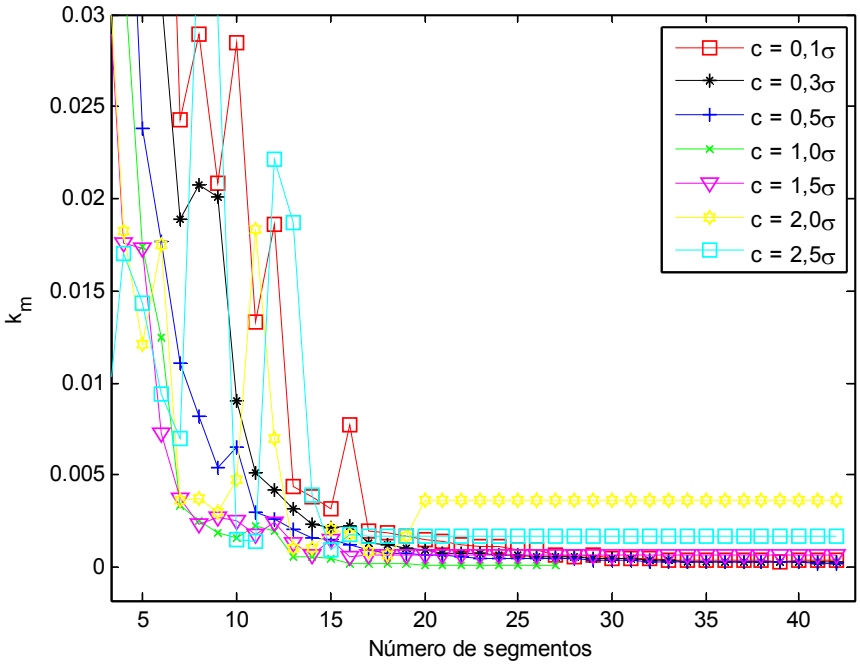


Figura 54 Valor de K_m em função do número de segmentos para a base de dados *clusters* circulares (Zoom da Figura 53)

Por outro lado, observa-se que para a base de dados mostrada na Figura 50(b) a taxa de acerto varia bastante em função do número e comprimento de segmentos, como pode ser observado na Figura 55. Isso se dá pelo fato dos agrupamentos estarem muito próximos um do outro.

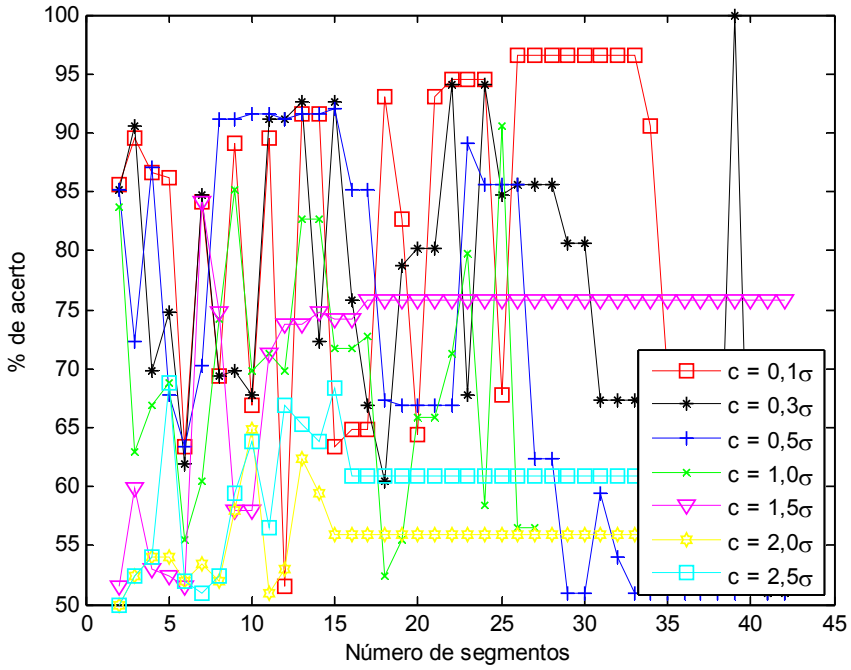


Figura 55 Taxa de acerto em função do comprimento e número de segmentos para a base de dados *clusters* circulares demonstrada na Figura 50(b)

É possível notar por meio das Figuras 56 e 57 que o índice K_m apresenta certa harmonia com os resultados de desempenho. Em seguimentos c = 2,0σ e 2,5σ apresenta maiores valores, isso se da devido à distribuição dos dados, em que seguimentos maiores possuem um índice K_m maior, embora exista certa divergência com o desempenho. Esse fato ocorre devido à distância dos eventos aos segmentos aumentarem no caso de segmentos maiores.

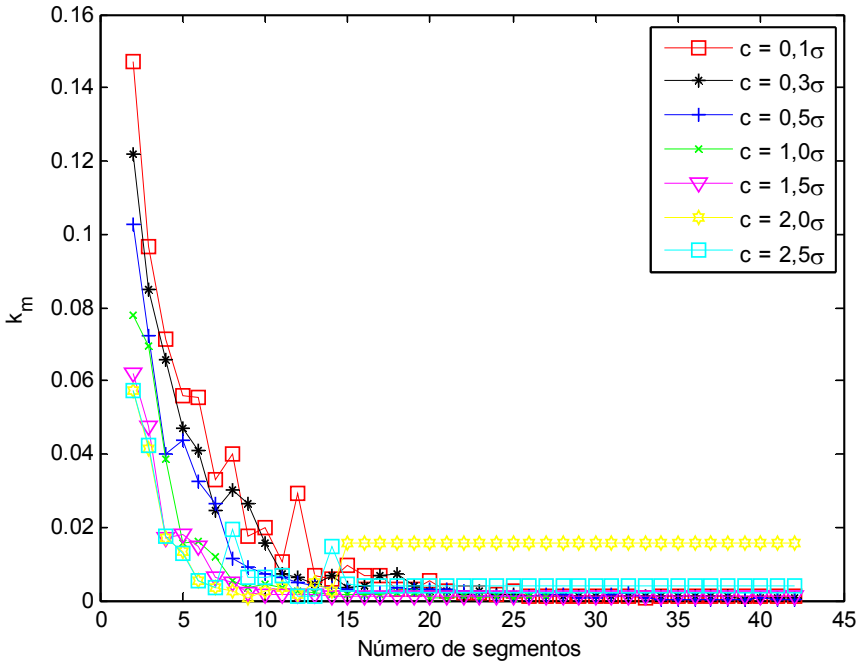


Figura 56 Valor de K_m em função do número de segmentos para a base de dados *clusters* circulares

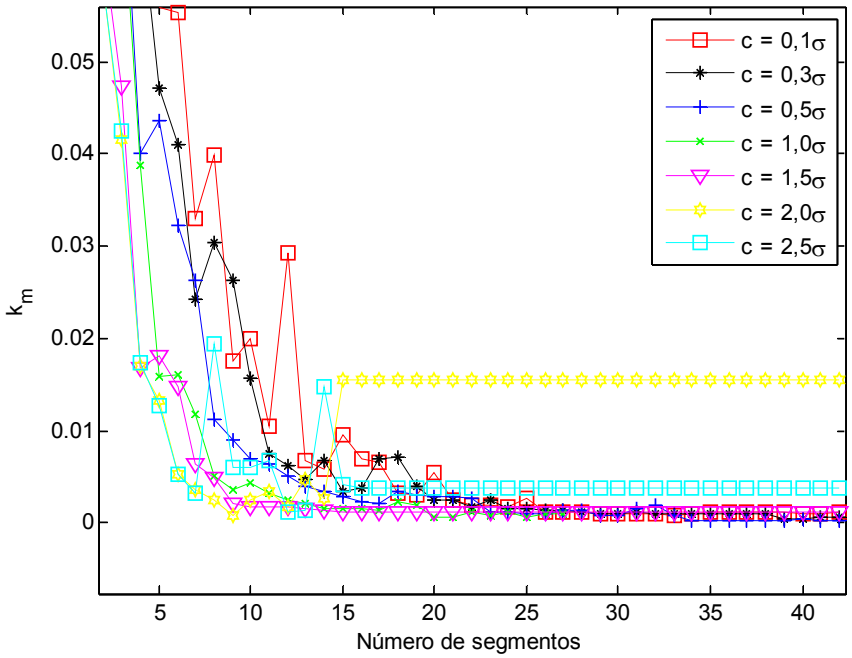


Figura 57 Valor de K_m em função do número de segmentos para a base de dados *clusters* circulares (Zoom da Figura 56)

6 CONCLUSÃO E CONSIDERAÇÕES FINAIS

Este trabalho apresentou um método baseado em Curvas Principais. O objetivo foi realizar o reconhecimento de padrões de forma não supervisionada. O método proposto foi aplicado às bases bidimensionais e multidimensionais de características e dimensionalidade variada.

O método proposto explora a boa capacidade de representação de dados das Curvas Principais e, como vantagem, mantém os segmentos originalmente construídos pelo algoritmo k-seg, não alterando, portanto, a forma de representação da curva construída pelo algoritmo, além de ser bastante simples em comparação com outros métodos de agrupamento que também utilizam curvas principais.

Os resultados obtidos foram comparados a quatro algoritmos, dois lineares (KM-M, KM-E) e um não linear, a rede SOM. O método proposto obteve bons resultados para as bases de dados, com destaque para as bases bidimensionais, em que o resultado foi superior aos demais métodos.

Para as bases de dados multidimensionais o método proposto foi superior aos métodos KM-E e KM-M em termos de desempenho e inferior à rede SOM. O bom resultado alcançado pela rede SOM é justificado pela sua alta capacidade de mapeamento não linear de dados.

Os resultados mostraram que o método proposto, apesar de bastante sensível aos parâmetros de entrada, pode levar a bons resultados para bases de dados com distribuições alongadas ou esféricas.

Do ponto de vista de capacidade de generalização o método proposto obteve resultados próximos aos demais métodos para a base *Wine*, nas bases Iris e espiral dupla obteve resultados próximos ao *K-means*, nas bases Diabetes e *half rings* o método proposto obteve uma taxa de erro elevada frente aos demais métodos. O fato de os dados serem sorteados aleatoriamente durante a análise da

capacidade de generalização dos métodos influi diretamente na taxa de erro, devido à distribuição aleatória dos dados e consequentemente a dificuldade em classificar os dados corretamente. Isso explica as elevadas taxas de erro obtidas nesta etapa.

Em relação aos parâmetros de entrada, no geral, comprimentos de segmentos menores, entre 0,3σ e 1,5σ, levam a melhores resultados. Por outro lado, a escolha do número de segmentos adequado ainda é um desafio e merece ser mais bem investigado em trabalhos futuros.

O índice proposto (K_m) para avaliar a qualidade dos agrupamentos mostrou algumas incoerências em relação aos resultados de desempenho. Foi verificado que o cálculo do mesmo precisa ser modificado em função do cálculo da distância do evento ao segmento, de tal forma que as interligações entre os segmentos sejam levadas em consideração, que pode ser implementado em trabalhos futuros.

Como vantagem, o método proposto disponibiliza as distâncias entre os segmentos da CP, que podem ser utilizadas para se ter uma ideia da quantidade de agrupamentos presentes no conjunto de dados.

Em trabalhos futuros, pretende-se investigar o uso de métodos para refinar os parâmetros de entrada do método proposto, melhorando assim tal desempenho em determinadas distribuições de dados.

REFERÊNCIAS

ALBUS, J. E. et al. **Syntactic pattern recognition, applications**. New York: Springer Science & Business Media, 2012. 270 p. (Communication and Cybernetics, 14).

ANTONIO, A. et al. Algoritmos para reconhecimento de padrões. **Revista Ciências Exatas**, Taubaté, v. 8, p. 129-145, 2002.

ASSUNÇÃO, R. M.; LAGE, J. P.; REIS, E. A. Análise de conglomerados espaciais via árvore geradora mínima. **Revista Brasileira de Estatistica**, Rio de Janeiro, v. 63, n. 220, p. 7–24, 2002.

BANFIELD, D.; RAFTERY, A. E. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. **Journal of the American Statistical Association**, New York, v. 87, n. 417, p. 7–16, 1992.

BATCHELOR, B. G. **Pattern recognition**: ideas in practice. New York: Springer Science & Business Media, 2012.

BELLMAN, R. et al. **Adaptive control processes**: a guided tour. Princeton: Princeton university, 1961. v. 4.

CASCAO, L. V. C. **Modelos de inteligência computacional para apoioa triagem de pacientes e diagnostico clinico de tuberculose pulmonar**. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2011.

CHANG, K.; GHOSH, J. **Principal curve classifier-a nonlinear approach to pattern classification**. 1998a. Disponível em: <https://www.researchgate.net/profile/Joydeep_ Ghosh6/publication/3753564_Principal_curve_classifier-a_nonlinear_approach_to_ patternclassification/links/0c9605225f0cce7498000000.pdf>. Acesso em: 25 out. 2015.

CHANG, K.; GHOSH, J. **Principal curves for nonlinear feature extraction and classification**. 1998b. Disponível em: <http://proceedings.spiedigitallibrary.org/ proceeding.aspx?articleid=937602>. Acesso em: 23 set. 2015.

CLEJU, P.; FRÄNTI, P.; WU, X. "Clustering by principal curve with tree structure". In: INTERNATIONAL SYMPOSIUM ON SIGNALS, CIRCUITS & SYSTEMS, 2., 2005, Iasi. **Proceedings…** Iasi: ISSCS, 2005. p. 617-620.

DELICADO, P. Another look at principal curves and surfaces. **Journal of Multivariate Analysis**, New York, v. 77, n. 1, p. 84–116, 2001.

DELICATO, P.; HUETRA, M. Principal curves of oriented points. **Computational Statistics,** Heidelberg**,** v. 18, n. 2, p. 293-315, 2003.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society. Series B (Methodological)**, London, v. 1, p. 38, 1977.

DEVIJVER, P. A.; KITTLER, J. **Pattern recognition**: a statistical approach. London: Prentice-Hall, 1982. 448 p.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2nd ed. New York: J. Wiley, 2012. 680 p.

EINBECK, J.; TUTZ, G.; EVERS, L. Local principal curves. **Statistics and Computing**, London, v. 15, n. 4, p. 301-313, 2005.

FAIER, J. M. **Curvas principais aplicadas na identificação de descargas parciais em equipamentos de potência**. 2006. 102 p. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal o Rio de Janeiro, Rio de Janeiro, 2006.

FERREIRA, D. D. et al. A new power quality deviation index based on principal curves. **Electric Power Systems Research**, Lausanne, v. 125, p. 8–14, 2015.

FERREIRA, D. D. et al. Exploiting principal curves for power quality monitoring. **Electric Power Systems Research**, Lausanne, v. 100, p. 1–6, 2013. Disponível em: <http://doi.org/ 10.1016/ j.epsr.2013.02.006>. Acesso em: 23 jan. 2015.

FERREIRA, D. D. et al. Sistema de detecção de distúrbios elétricos baseado em curvas principais e redes neurais. In: SIMPÓSIO BRASILEIRO DE SISTEMAS ELÉTRICOS, 3., 2010, Belém. **Anais...** Belém: [s. n.], 2010. 1 CD ROM.

FRÄNTI, P.; KIVIJÄRVI, J. Randomised local search algorithm for the clustering problem. **Pattern Analysis & Applications**, Heidelberg, v. 3, n. 4, p. 358–369, 2000.

GONÇALVES, M. L.; ANDRADE NETTO, M. L.; ZULLO JÚNIOR, J. Um sistema neural modular para classificação de imagens utilizando mapas de kohonen. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 8., 1996, Salvador. **Anais…** Salvador: INPE, 1996. p. 845–849.

GU, I. Y. H. et al. A statistical-based sequential method for fast online detection of fault-induced voltage dips. **IEEE Transactions on Power Delivery**, New York, v. 19, n. 2, p. 497–504, 2004.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, Boston, v. 17, n. 2/3, p. 107–145, 2001.

HASTIE, T.; STUETZLE, W. Principal curves. **Journal of the American Statistical Association**, New York, v. 84, n. 406, p. 502–516, 1989.
JOLLIFFE, I. **Principal component analysis**. New York: J. Wiley, 2002.

JOLLIFFE, I. T.; HOPE, P. B. Representation of daily rainfall distributions using normalized rainfall curves. **International Journal of Climatology**, Chichester, v. 16, p. 1157–1163, 1996.

KÉGL, Â. et al. Learning and design of principal curves. pattern analysis and machine intelligence. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, New York, v. 22, n. 3, p. 281–297, 2000.

KOHONEN, T. **Self-organization and associative memory**. New York: Springer, 2012. v. 8.

KOHONEN, T. The self-organizing map. **Neurocomputing**, Oxford, v. 21, n. 1, p. 1–6, 1998.

LANGE, N.; BISHOP, C. M.; RIPLEY, B. D. Pattern Recognition and Neural Networks. **Journal of the American Statistical Association**, New York, v. 92, n. 440, p. 1642-1645, 1996. Disponível em: <http://doi.org/10.2307/2965437>. Acesso em: 15 dez. 2016.

LICHMAN, M.; BACHE, K. **UCI Machine learning repository**. Irvine: University of California, 2013. Disponível em: <http://archive.ics.uci.edu/ml>. Acesso em: 23 set. 2015.

LIDEN, R. Técnicas de Agrupamento -Tutorial. **Revista de Sistemas de Informação da FSMA**, Rio de Janeiro, v. 4, p. 18–36, 2009.

LIU, X. et al. Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network. **Frontiers of Earth Science**, Lausanne, v. 7, n. 2, p. 206–216, 2013. Disponível em: <http://link.springer.com/10.1007/s11707-012-0350-y>. Acesso em: 22 set. 2015.

MURTAGH, F. **Multidimensional clustering algorithms**. Vienna: Physika Verlag, 1985. v. 1.

PIMENTEL, E. P.; FRANÇA, V. F.; OMAR, N. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 14., 2003, Rio de Janeiro. **Anais...** Rio de Janeiro: UFRJ, 2003. 1 CD ROM.

RAMIREZ, I.; SPRECHMANN, P.; SAPIRO, G. Classification and clustering via dictionary learning with structured incoherence and shared features. In: EEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 1., 2010, San Francisco. **Proceedings…** San Francisco: CVPR, 2010. p. 3501–3508.

RIBEIRO, M. V. et al. Power quality disturbances detection using HOS. **IEEE Power Engineering Society**, New York, v. 1, p. 6, 2006.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, Antwerpen, v. 20, p. 53–65, 1987.

SILVA, R. N. et al. Non-invasive method to analyse the risk of developing diabetic foot. **Healthcare Technology Letters**, London, v. 1, n. 4, p. 109–113, 2014.

SILVA, R. N. **Identificação de pacientes com potencial para desenvolver o pé diabético baseada em técnicas de reconhecimento de padrões e ações de auto cuidado**. 2014. 84 p. Dissertação (Mestrado em Modelagem de Sistemas Biológicos) – Universidade Federal de Lavras, Lavras, 2014.

SONG, Y. et al. **Iknn**: informative k-nearest neighbor pattern classification, in Knowledge discovery in databases: PKDD. Berlin: Springer, 2007. p. 248–264.

STANFORD, D. C.; RAFTERY, A. E. Finding curvilinear features in spatial point patterns : principal curve clustering with noise. **IEEE transactions on pattern analysis and machine intelligence**, New York, v. 22, n. 6, p. 601–609, 2000.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. Burlington: Academic, 2009.

TIBSHIRANI, R. Principal curves revisited. **Statistics and Computing,** London, v. 2, n. 4, p. 8, 1992.

VERBEEK, J. J.; VLASSIS, N.; KROSE, B. A k -segments algorithm for finding principal curves. **Pattern Recognition Letters**, Amsterdam, v. 23, n. 8, p. 1009–1017, 2002.

WEST, D. Neural network credit scoring models. **Computers & Operations Research**, New York, v. 27, n. 11, p. 1131–1152, 2000.

APÊNDICE

Nesta seção, a lista de trabalhos publicados e que são frutos dessa dissertação são apresentados. Estes estão organizados em ordem cronológica.

A.1 Artigo Publicados em Anais de Congresso

1. MORAES, C. E.; FERREIRA, D. D. "Método de classificação não supervisionada baseado em curvas principais". In: CONGRESSO MINEIRO DE ENGENHARIA E TECNOLOGIA, 1., 2015, Lavras. **Anais...** Lavras: UFLA, 2015. 1 CD ROM.

2. MORAES, C. E.; FERREIRA, D. D. " Classificação não supervisionada com curvas principais". In: CONGRESSO DE PÓS-GRADUAÇÃO, 24., 2015, Lavras. **Anais...** Lavras: UFLA, 2015. 1 CD ROM.