MÉTODO DE CLASSIFICAÇÃO NÃO SUPERVISIONADA BASEADO EM CURVAS PRINCIPAIS

ELSON C. C. MORAES, DANTON D. FERREIRA

Departamento de Engenharia, Universidade Federal d Lavras
Caixa Postal 3037,37200-000, Lavras, MG, Brasil
E-mails: elsonclaudio@gmail.com, danton@deg.ufla.br

Abstract - The manuscript presents a new grouping method and pattern classification based on Principal Curves (CP). The main curves consist of a nonlinear generalization of the Principal Component Analysis (PCA) and are defined as one-dimensional smooth curves that pass in the middle of a multidimensional dataset, providing a one-dimensional signature. The CP extraction algorithm used was the non-smooth k-segment (k-seg). The proposed grouping method splits the CP group, originally obtained by the k-seg algorithm, in two or more curves, eliminating the greater-length interconnections among the segments that make up the CP. The number of eliminated interconnections is made according to user-defined parameters. It is then calculated the distance of the data to the new CPs and the lettering of the data is made, subsequently, according to the criterion of the smallest distance (squared Euclidean distance) of the data to curves obtained by the method.) The method was applied to the following bases: Double Spiral, Half-rings, Iris, Diabetes and Wine. These bases have varied dimensionality and characteristics. The results were compared with the algorithms K-means and Particle Swarm optimization (PSO). The method achieved superior results in four of the five bases. To the Diabetes base the result was below the PSO and the k-means algorithm. However, the difference between the best results and our method had a 6.96% error.

Keywords - Pattern recognition, clustering, k-segments, principal curve, k-means, particle swarm optimization.

Resumo - O trabalho apresenta um novo método de agrupamento e classificação de padrões baseado em Curvas Principais (CP). As curvas principais consistem numa generalização não linear da Análise de Componentes Principais (PCA) e são definidas como curvas suaves, unidimensionais, que passam no meio de um conjunto multidimensional de dados, fornecendo uma assinatura unidimensional do mesmo. O algoritmo de extração de CP utilizado foi o k-segmentos não suave (k-seg). O método de agrupamento proposto divide a CP originalmente obtida pelo algoritmo k-seg em duas ou mais curvas, eliminando-se as interligações de maior comprimento entre os segmentos que compõe a CP. O número de interligações eliminadas é feito de acordo com os parâmetros definidos pelo usuário. Em seguida é calculada a distância dos dados às novas CPs e, posteriormente, é feita a rotulação dos dados de acordo com o critério da menor distância (quadrado da distância euclidiana) dos dados às curvas obtidas pelo método. O método foi aplicado às seguintes bases: Espiral dupla, Half-rings, Iris, Diabetes e Wine. Estas bases possuem dimensionalidade e características variadas. Os resultados foram comparados com os algoritmos k-means e o Particle Swarm Optimization (PSO). O método obteve resultados superiores em quatro das cinco bases. Na base Diabetes o resultado foi inferior ao PSO e ao k-means, entretanto, a diferença entre o melhor resultado e o nosso método foi de 6,96% de erro.

Palavras-chave - Reconhecimento de padrões, agrupamento, K-segmentos, curvas principais, k-means, particle swarm optimization.

1 Introdução

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 (Theodoridis&Koutroumbas, 2009).

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.

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 supervisionado são limitados a algumas aplicações que envolvem dados com distribuições específicas. Alem disso, é uma realidade conhecida que as bases de dados podem ser de diferentes naturezas e complexidades.

O presente trabalho teve como principal objetivo explorar a capacidade de representação de dados da técnica de curvas principais (Hastie&Stuetzle, 1989) 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 (Verbeek, Vlassis,

&Krose, 2002) 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 ParticleSwarmOptimization (PSO) (Esmin, Coelho, &Matwin, 2013) foram também aplicados às bases de dados para fins de comparação.

2 Material e Métodos

2.1 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 não linear e compacta do mesmo.

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$\mathbf{f}(t)$uma curva suave (infinitamente diferenciável) em $\mathfrak{R}^d$padronizado em$t$ e, para qualquer $\mathbf{x}$ $\mathfrak{R}^d$, $t_f(\mathbf{x})$ denota o maior valor do parâmetro t para o qual a distância entre $\mathbf{x}$e$\mathbf{f}(t)$é minimizada:

$$f(t) = [\, f_1(t)\, f_2(t) \ldots f_d(t)]^T \qquad (1)$$

onde$T$ denota a matriz transposta.

O índice de projeção é definido como:

$$t_f(\mathbf{x}) = sup_t\{t: \|\mathbf{x} \quad \mathbf{f}(T)\| = inf_\mu\|\mathbf{x} \quad \mathbf{f}(\mu)\|\} \qquad (2)$$

em que x é um evento arbitrário pertencente a **X** e μ é uma variável auxiliar definida em . 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 **x**. Se houver mais de um valor possível, o maior deles é selecionado.

Diversos algoritmos foram propostos para extrair curvas principais. Este trabalho utiliza o algoritmo de (Verbeeket al, 2001), conhecido por k-segmentos não suave (k-seg), que além de ser robusto e ter convergência garantida, é menos susceptível a mínimos locais. O mesmo constrói a curva adicionando segmentos através da adição de um segmento a cada passo de interação a partir de um segmento inicial até que ocorra a convergência. O diagrama de blocos da Figura 1 ilustra o algoritmo do k-segmentos não suave.

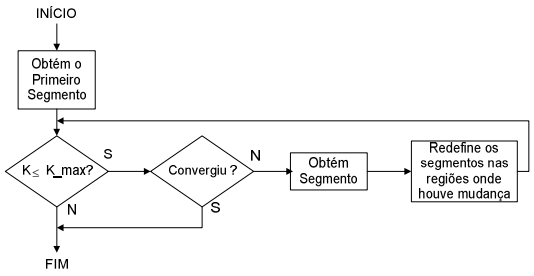

Figura 1: Diagrama de blocos do algoritmo k-segmentos.

2.2 *Método proposto*

O método de agrupamento de dados proposto nesse 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.

As etapas do método proposto para encontrar os agrupamentos em um conjunto de dados qualquer são descritas abaixo e ilustradas no diagrama de blocos da Figura 2:

1. Passo 1: Construção da curva principal de acordo com o algoritmo k-seg;
2. Passo 2: Cálculo dadistância entre os segmentos da curva principal. O comprimento das interligações é obtido utilizando o quadrado da distância euclidiana entre os vértices das interligações;
3. 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;
4. 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.

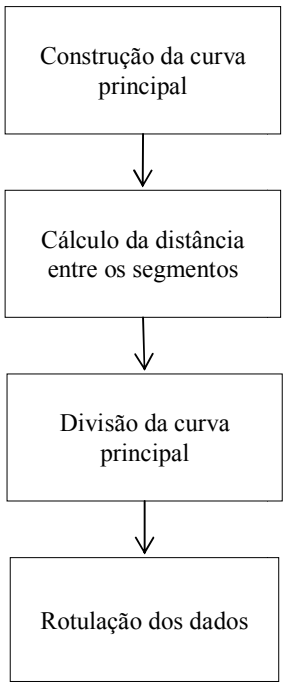

Figura 2: Diagrama de blocos do método proposto.

2.3 *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 de dimensão variada, dados reais e simulados, envolvendo problemas clássicos de agrupamento e classificação.

Foram utilizados cinco conjuntos de dados da UCI, duas bases de dados sintéticas, Espiral Dupla e HalfRings e três bases reais, Iris, Diabetes e Wine.

A Tabela 1 resume em detalhes as bases de dados utilizadas.

Tabela 1: Descrição da base de dados

Dados	Nº de atributos	Nº de classes	Nº total de dados
Espiral Dupla	2	2	206
HalfRings	2	2	373
Iris	4	3	150
Diabetes	8	2	768
Wine	14	3	178

3 Resultados e discussão

Os resultados e discussões estão divididos por bases de dados nesta seção.

3.1 *Espiral dupla*

A base de dados modelada pela curva principal é mostrada na Figura 3.

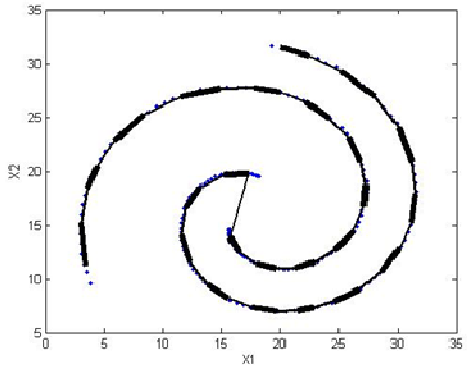

Figura 3. Curva principal original construída sobre um conjunto de dados em duas dimensões.

A Figura 4 mostra as distâncias entre os vértices das interligações da curva principal mostrada na Figura 3. Pode-se concluir que existem apenas dois agrupamentos, uma vez que apenas um segmento de interligação muito longo em relação aos outros segmentos foi encontrado.

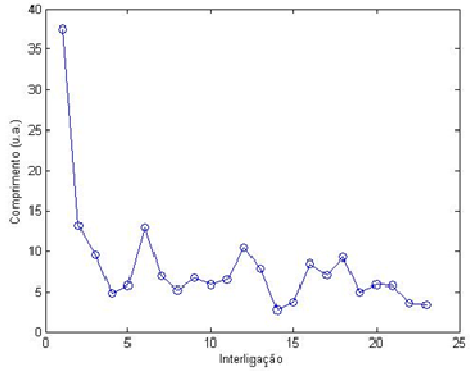

Figura 4. Distâncias entre os vértices das interligações.

Os agrupamentos obtidos pelo método são mostrados na Figura 5. Observe que o método obteve 100% de acerto para esta base de dados. A Figura 6 apresenta os agrupamentos obtidos pelo algoritmo k-means com um acerto de 51,13% .

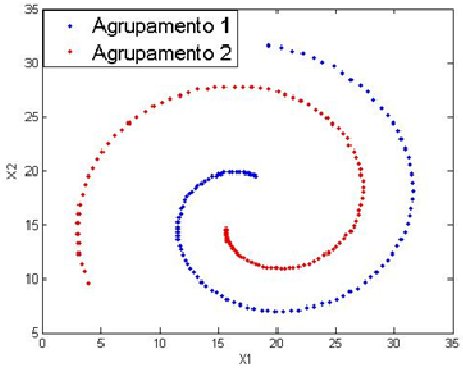

Figura 5. Agrupamento obtido pelo método proposto

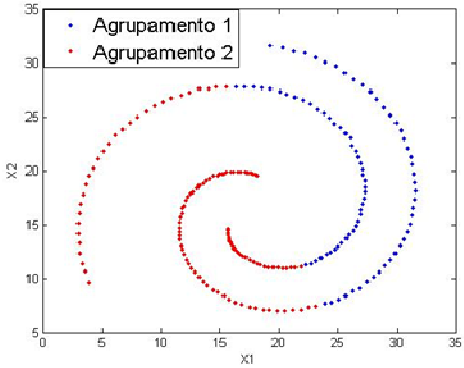

Figura 6. Agrupamento obtido pelo método kmeans

3.2 *HalfRings*

A base de dados modelada pela curva principal é mostrada na Figura 7.

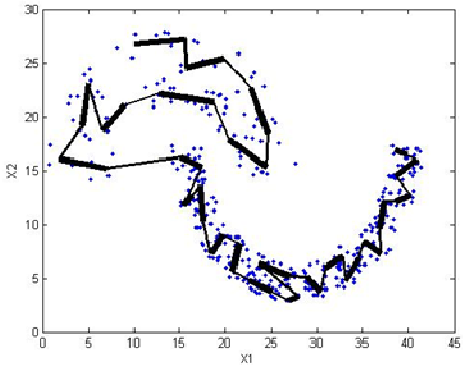

Figura 7. Curva principal original construída sobre um conjunto de dados em duas dimensões.

A Figura 8 mostra as distâncias entre os vértices das interligações da curva principal mostrada na Figura 7. Pode-se concluir a presença de apenas dois agrupamentos, uma vez que apenas um segmento de interligação muito longo em relação aos outros segmentos foi encontrado.

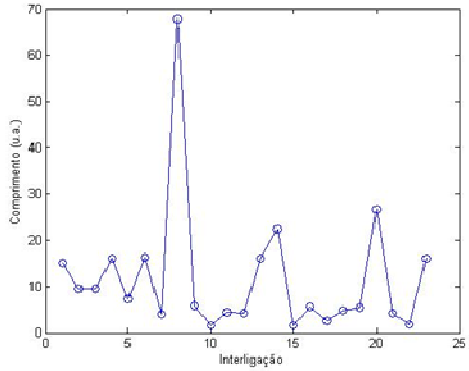

Figura 8– Exemplo de agrupamento gerado pelo método proposto

Os agrupamentos obtidos pelo método proposto são mostrados na Figura 9, em que um acerto na classificação dos dados de 100% foi alcançado. A Figura 10 apresenta o resultado do agrupamento de dados alcançado pelo algoritmo k-means, que obteve uma taxa de acerto de 52.85%.

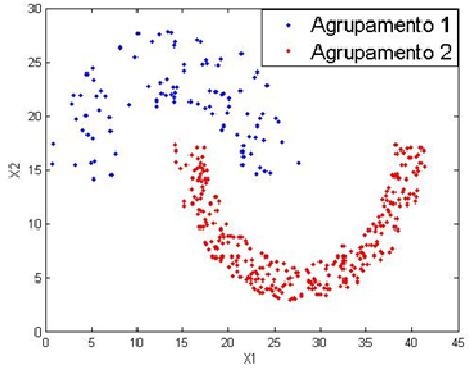

Figura 9. Agrupamento obtido pelo método proposto

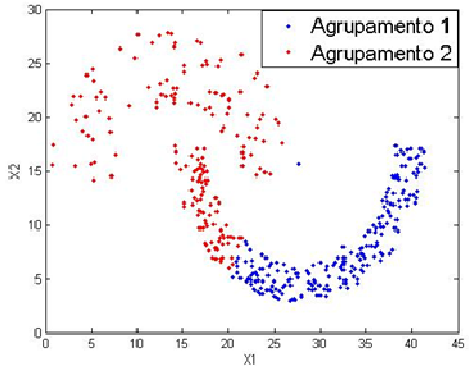

Figura 10. Agrupamento obtido pelo método k-means

A seguir são apresentados os resultados referentes às bases de dados reais: Iris, Diabetes e Wine.

Observa-se na Tabela 2 que os resultados obtidos pelo método proposto foram superiores para

as bases Wine e Iris em comparação com os métodos k-means e PSO.

Para a base Diabetes o método proposto obteve um resultado um pouco inferior ao PSO e semelhante ao k-means. Cumpre ressaltar que os resultados mostrados referentes ao algoritmo PSO foram extraídos de (Coelho, 2014).

Tabela 2: Resultado do método proposto em comparação com o k-means e o PSO em termos de taxa de erro (%)

Base de dados	k-means	PSO	Método proposto
Iris	10,66	4,00	**1.33**
Diabetes	33,98	31,25	33,98
Wine	29,77	26,97	**7,85**

4 Conclusões

O trabalho apresentou um novo método de agrupamento baseado em curvas principais. Tendo em vista os bons resultados obtidos com bases bidimensionais e multidimensionais, conclui-se que essa técnica é uma ferramenta com grande potencial a ser aplicado na área de reconhecimento de padrões.

Como vantagens, vale ressaltar que o método proposto mantém os segmentos originalmente construídos pelo algoritmo k-seg, não alterando, portanto, a forma de representação da curva principal construída pelo algoritmo, além de ser bastante simples em comparação com outros métodos de agrupamento que também utilizam curvas principais ((Banfield e Raftery(1992); (Chang e Chosh, 1998); (Cleju, Fränti, e Wu,2005); e (Liu et al, 2013); (Stanford e Raftery, 2000)).

Agradecimentos

Os autores agradecem a Capes, CNPq e FAPEMIG, pelo suporte financeiro.

Referências Bibliográficas

Banfield, D., & Raftery, A. E. (1992). Ice Floe Identification in Satellite Images Using Mathematical Morphology and Clustering About Principal Curves. Journal of the American Statistical Association, 87(417), 7–16.

Chang, K., & Ghosh, J. (1998). Principal curves for nonlinear feature extraction and classification, 120–129.

Cleju, I., Fränti, P., & Wu, X. (2005). Clustering Based on Principal Curve. In H. Kalviainen, J. Parkkinen, & A. Kaarna (Eds.), Image Analysis SE - 88 (Vol. 3540). Springer Berlin Heidelberg. http://doi.org/10.1007/11499145_88

Coelho, R. A. (2014). Algoritmo de enxame de part{í}culas ensemble para clusteriza{ç}{ã}o de dados.

Esmin, A. a a, Coelho, R. a., & Matwin, S. (2013). A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artificial Intelligence Review, 2014, 1–23. http://doi.org/10.1007/s10462-013-9400-4

Hastie, T., & Stuetzle, W. (1989). Principal Curves. Journal of the American Statistical Association, 84(406), 502–516.

Lichman, M., & Bache, K. (2013). {UCI} Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2013. Retrieved from http://archive.ics.uci.edu/ml

Liu, X., Lu, F., Zhang, H., & Qiu, P. (2013). Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network. Frontiers of Earth Science, 7(2), 206–216. Retrieved from http://link.springer.com/10.1007/s11707-012-0350-y

Stanford, D. C., & Raftery, A. E. (2000). Finding Curvilinear Features in Spatial Point Patterns : Principal Curve Clustering with Noise. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(6), 601–609.

Theodoridis, S., & Koutroumbas, K. (2009). Pattern Recognition. IEEE TRANSACTIONS ON NEURAL NETWORKS, 19(2), 376. http://doi.org/10.1016/B978-1-59749-272-0.50001-3

Verbeek, J. J., Vlassis, N., & Krose, B. (2002). A k - segments algorithm for finding principal curves. Pattern Recognition Letters, 23(8), 1009–1017.

Banfield, D., & Raftery, A. E. (1992). Ice Floe Identification in Satellite Images Using Mathematical Morphology and Clustering About Principal Curves. Journal of the American Statistical Association, 87(417), 7–16.

Chang, K., & Ghosh, J. (1998). Principal curves for nonlinear feature extraction and classification, 120–129.

Cleju, I., Fränti, P., & Wu, X. (2005). Clustering Based on Principal Curve. In H. Kalviainen, J. Parkkinen, & A. Kaarna (Eds.), Image Analysis SE - 88 (Vol. 3540). Springer Berlin Heidelberg. http://doi.org/10.1007/11499145_88

Coelho, R. A. (2014). Algoritmo de enxame de part{í}culas ensemble para clusteriza{ç}{ã}o de dados.

Esmin, A. a a, Coelho, R. a., & Matwin, S. (2013). A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artificial Intelligence Review, 2014, 1–23. http://doi.org/10.1007/s10462-013-9400-4

Hastie, T., & Stuetzle, W. (1989). Principal Curves. Journal of the American Statistical Association, 84(406), 502–516.

Lichman, M., & Bache, K. (2013). {UCI} Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2013. Retrieved from http://archive.ics.uci.edu/ml

Liu, X., Lu, F., Zhang, H., & Qiu, P. (2013). Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network. Frontiers of Earth Science, 7(2), 206–216. Retrieved from http://link.springer.com/10.1007/s11707-012-0350-y

Stanford, D. C., & Raftery, A. E. (2000). Finding Curvilinear Features in Spatial Point Patterns : Principal Curve Clustering with Noise. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(6), 601–609.

Theodoridis, S., & Koutroumbas, K. (2009). Pattern Recognition. IEEE TRANSACTIONS ON NEURAL NETWORKS, 19(2), 376. http://doi.org/10.1016/B978-1-59749-272-0.50001-3

Verbeek, J. J., Vlassis, N., & Krose, B. (2002). A k - segments algorithm for finding principal curves. Pattern Recognition Letters, 23(8), 1009–1017.