CLASSIFICAÇÃO DE DADOS AMOSTRAIS BASEADO NO ALGORITMO K-SEGMENTOS

Zaudir Dal Cortivo
Programa de Pós-Graduação em Métodos Numéricos – PPGMNE/UFPR
Secretaria de Estado da Educação – SEED/PR
81531-980 – Curitiba – Pr
E-mail: zdalcortivo@gmail.com

Jair Mendes Marques
Programa de Pós-Graduação em Métodos Numéricos – PPGMNE/UFPR
81531-980 – Curitiba – Pr
E-mail: jair.marques@utp.br

RESUMO - Neste artigo, é proposto um classificador para dados amostrais usando o algoritmo k-segmentos de Verbeek. Este algoritmo gera as curvas principais que são uma generalização das componentes principais lineares. Para avaliação do algoritmo, foram utilizados três conjuntos de dados amostrais. O método consiste na substituição das distâncias das médias (centroides) das classes pelas distâncias aos segmentos gerados pelo algoritmo, como medida de classificação. Experimentalmente, os resultados apresentaram desempenho igual ou melhor para esta nova abordagem.

Palavras-chave: Análise Discriminante. k-segmentos. Curvas Principais.

1. Introdução

A análise discriminante é uma técnica da estatística multivariada que tem como um de seus objetivos descobrir as características que distinguem os membros de uma classe, definidos *a priori*, de modo que sendo conhecidas as características de um novo indivíduo se possa prever a que classe ele pertence. Uma observação é classificada na classe cuja distância euclidiana está mais próxima de um valor central (centroide). Esta forma de classificação é denominada método de Fisher [10].

Aplicações da análise discriminantes são encontradas em diversas áreas: [1] aplicaram na análise de insolvência de empresas de grande porte (sociedades anônimas); [12] aborda o estudo das imagens de face como um problema de reconhecimento de padrões; [4] apresentaram uma nova abordagem para classificação. Para avaliação do método, utilizou-se um conjunto de dados da área médica. Outras aplicações podem ser encontradas em [3], [11], [13], [14], [17], [21] e [21]. Novas técnicas têm sido desenvolvidas, como [15], que comparou diversos algoritmos de classificação e

concluiu que o valor da razão (R) dos autovalores influenciam na acurácia do classificador, $R = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$ (1) com $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

As curvas principais, primeiramente definidas por [8], são curvas unidimensionais que passam no "meio" de um conjunto de dados no espaço p-dimensional, fornecendo uma descrição não linear dos dados, e sua forma é sugerida pelo conjunto de dados. Posteriormente, surgiram outras definições, como a de k-segmentos [21].

Trabalhos utilizando a curva principal para classificação foram desenvolvidos por [6], que usaram a definição de [8, 9]. [3] desenvolveram um algoritmo para extração e classificação de dados utilizando duas definições de curvas principais, a de [8] e a de [2]. [5] propôs a classificação de navios em curvas principais baseado no algoritmo k-segmentos. [20] propuseram um novo classificador para dados *microarrays* usando curvas principais. Uma curva principal é calculada para cada classe, e uma nova observação amostral é classificada para a classe da curva com menor distância, segundo a esperança do erro quadrado (*Expected Squared Error - ESE*). Resultados experimentais mostram que a PC tem melhor desempenho quando o tamanho da amostra é pequeno. [12] investigaram a eficiência de classificação, utilizando MP's (*Morphological profiles*) construídas a partir das características de NLPCA (*Nonlinear Principal Componentes Analysis*).

A proposta deste artigo é um classificador de dados amostrais multivariados, denominado método k-segmentos, baseados em componentes principais não lineares. A técnica consiste em, primeiramente, calcular os segmentos para cada classe e, em seguida, uma nova observação amostral é rotulada para a classe cuja distância euclidiana é a mais próxima a ela. Experimentos são realizados para mostrar o seu desempenho, o qual é medido pela taxa aparente de erro.

2. Análise Discriminante de Fischer

A análise discriminante de Fisher para diversas populações (ADF) está preocupada com a análise de g classes de amostras, todas descritas pelas mesmas p variáveis, com a análise das matrizes $X_{n1\,x\,p}, X_{n2\,x\,p}, \ldots, X_{nj\,x\,p}$. O alvo é procurar combinações lineares das variáveis que maximizem a razão entre a variância entre as classes e a variância dentro das classes. Convém salientar que o termo "classe" se refere a um grupo de amostras, e o termo "grupo" refere-se a um conjunto de variáveis, [16].

Dada a matriz de dados $X_{n\,x\,p}$, onde n é o número de observações e p é o número de variáveis. A matriz X é dividida em g classes amostrais diferentes com todas as p variáveis, de forma que $n = \sum_{i=1}^{g} n_i$ (2). A matriz de covariância dentro de cada classe é definida como:

DOI: 10.5540/03.2014.002.01.0057

$$S_W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(\underline{x}_{ij} - \underline{\bar{x}}_i\right)\left(\underline{x}_{ij} - \underline{\bar{x}}\right)^t \quad (3)$$

e a matriz de covariância entre classes é definida como:

$$S_B = \sum_{i=1}^{g} \left(\underline{\bar{x}}_i - \underline{\bar{x}}\right)\left(\underline{\bar{x}}_i - \underline{\bar{x}}\right)^t \quad (4)$$

Onde x_{ij} é a j-ésima amostra da i-ésima classe (representada com um vetor coluna), $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ (5) é o vetor médio da classe i, e $\bar{x} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij}$ (6) é o vetor médio. Seja $\lambda_1, \lambda_2, \ldots, \lambda_s > 0$, onde s ≤ min(g -1, p), autovetores não nulos de $S_w^{-1}S_B$ (7) e $\hat{e}_1$, $\hat{e}_2$, ... , $\hat{e}_s$ os autovetores correspondentes normalizados. Então, o vetor dos coeficientes w maximiza a razão:

$$J(w) = \frac{w' S_B w}{w' S_w w} \quad (8)$$

A matriz $B_{LDA} = [\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_s]$ é a matriz dos s autovetores normalizados e $U = XB_{LDA}$ é a matriz dos espaços discriminantes (ou Matriz de transformação canônica). A região de classificação para classe i é definida por $C_i = \{\underline{u} \in U / \|\underline{u} - \underline{\bar{u}}_i\| < \|\underline{u} - \underline{\bar{u}}_h\| \text{ para todo } h \neq i\}$ (9), onde i = 1, 2, ..., g, e $\underline{\bar{u}}_i$ é a i-ésima média canônica. O espaço U é divido em g regiões de classificação, C_1, C_2 , ..., C_g tal que $C_h \cap C_i = \emptyset$ se $h \neq i$ e $\bigcup_{i=1}^{g} C_i = U$. Uma nova observação é classificada na classe com menor distância do centroide $\underline{\bar{u}}_i$.

3. Curvas principais: método dos k-segmentos

A curva principal é uma generalização não linear de componentes principais e fornece uma curva suave unidimensional aproximada para um conjunto de dados em $\mathbb{R}^p$ [9]. [18] desenvolveram o algoritmo k-segmentos para construção das curvas principais não lineares. A seguir, é dada uma breve descrição deste algoritmo.

Este método tem convergência garantida e usa uma definição probabilística para encontrar as componentes principais, por meio da maximização de uma função de log-verossimilhança. O método assume que os dados são corrompidos por algum ruído e, por esse motivo, não requer que a curva seja um ajuste exato dos dados. A construção de uma curva principal é de forma incremental, isto é, inicia-se com apenas um segmento, sendo este número aumentado progressivamente. Depois de localizados os k-segmentos diferentes no conjunto de dados, estes são ligados formando uma linha poligonal, a qual pode ser usada como uma primeira aproximação para a curva principal, e finalmente faz-se o

DOI: 10.5540/03.2014.002.01.0057

alisamento da linha para obtenção da curva principal suave. O método tem uma sequência de vários algoritmos: do k-médias é extendido para o algoritmo k-lines, e deste se estende para o algoritmo k-segmentos.

4. Classificador baseado em curvas principais.

O algoritmo classificador foi construído semelhantemente ao modelo de Fischer.

1) A análise discriminante para os dados amostrais é aplicada. A transformação das variáveis originais X em uma nova matriz $Y = B_{LDA}^{t}X$, onde B_{LDA} é a matriz dos autovetores, que contêm os coeficientes discriminantes que maximizam a razão das variâncias entre as classes e dentro das classes.

2) Cálculo da matriz dos espaços discriminantes $U = XB_{LDA.}$

3) Determinar os segmentos para cada classe (algoritmo k-segmentos). Na extração das curvas são empregadas as matrizes dos espaços discriminantes U de cada classe, das quais são obtidas as matrizes de nós (*edges*) e de suas coordenadas (*vertices*).

4) Cálculo das distâncias euclidianas. Uma nova observação **x** é classificada na classe que contenha o segmento com a menor distância, isto é, seja f_i um segmento pertencente a classe i (i=1, 2, ..., g). Para uma observação amostral x, seja d_i a distância entre x e f_i. Se $d_i = \min_{k = 1,2, \ldots,g} \|x - f_k\|^2$, então x é classificado na classe k.

5. Resultados experimentais

Para avaliar o algoritmo proposto, foram utilizados três conjuntos de dados amostrais: Iris, Tiroide e *Wine*, ambos com 3 classes, todos obtidos na *UCI – Machines Respository* [19]. Cada conjunto tem a seguinte composição: Iris contém 150 observações e 4 variáveis; Tiroide contém 215 observações e 5 variáveis; e *Wine* contém 178 e 13 variáveis. Para cada conjunto, retirou-se uma observação x_i para classificação, o algoritmo executou a matriz $X_{n-1 \times p}$.

Conjunto	**Número de classificações Incorretas**		**% de erro**	
	Fisher	k-segmentos	Fisher	k-segmentos
Iris	3	3	2	2
Wine	2	0	1,12	0
Tiroide	13	5	6,04	2,32

Tabela 1: Taxa de Erro Aparente (proporção de itens mal classificados)

DOI: 10.5540/03.2014.002.01.0057

O desempenho da classificação foi medido pela taxa aparente de erro (APER). Para o conjunto Iris, o algoritmo teve desempenho igual ao de Fisher e ao classificador desenvolvido por [3]; porém nos demais, o desempenho foi superior ao de Fisher. No conjunto *Wine* a taxa de erro foi de 1,12% para o método de Fisher contra 0% do k-segmentos e para o conjunto Tiroide, o desempenho do algoritmo k-segmento foi superior em 3,72%, conforme é mostrado na Tabela 1.

6. Conclusão

Este artigo propõe um classificador baseado no algoritmo k-segmentos (curvas principais não lineares). A troca das médias pelas curvas principais pode resultar em uma melhor classificação dos dados. Na continuidade deste trabalho, entende-se a importância de aplicar este classificador a outros conjuntos de dados e na comparação com outros métodos de classificação.

Referências

[1] R. S. Mateus, R. O. Lacerda de Melo, T. A. Faria, "Análise de insolvência empresarial: uma abordagem financeira fundamentalista com aplicação do método estatístico multivariado e da técnica discriminante". ReCont: Registro Contábil. Vol. 2, nº. 1, (2011).

[2] J. D. Banfield, A. E. Raftery, "Ice floe identification in satellite imagens using mathematical morphology and clustering about principal curves." American Statistical Assoc., v. 87, pp. 7-16, (1992).

[3] K. Chang, J. Ghosh, "Principal curve classifer – A nonlinear approach to pattern classification." IEEE-Int. Joint Conf. Neural Networks, p. 695-700, (1998).

[4] S. Datta, "Classification of Breast cancer versus normal samples from mass spectrometry profiles using linear discriminant analysis of important features selected by random forests." Statistical Applications in Genetics and Molecular Biology, vol. 7, nº. 02, (2008).

[5] H. L. Fernandez, "Classificação de navios baseado em curvas principais." Dissertação de Mestrado – Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, (2005).

[6] S. Gardner, N. J. Le Roux, "Biplot methodology for discriminate analysis based upon robust methods and principal curves." Proceedings of the 8th Conference of the International Federation of Classification Societies, Springer-Verlag, Berlin, pp. 169-176, (2002).

[7] T. Hastie, "Principal curves and Surfaces." California, Tese Pós-doutorado, Stanford Linear Accelerator Center, Stanford University, 1984.

[8] T. Hastie, W. Stuetzle, "Principal curves", JASA Journal. American. Statistic. assoc., 84, pp. 502–516, (1989).

[9] T. Hastie., R. Tibshiriani, J. Friedman, "The elements of statistical learning." Data mining, inference, and prediction, Springer Science and Business Media, 2ª ed., New York, (2009).

[10] R. A. Johnson, D. W. Wichern, "Applied multivariate statistical analysis", 4ª ed., Upper Saddle River: Prentice Hall, (1998).

[11] B. Kégl, A. Krzyzak, T. Linder, Z. Kenneth, "Learning and design of principal curves." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, nº. 3, pp. 281-297, (2000).

[12] E. C. Kitani, "Análise discriminantes lineares para modelagem e reconstrução de imagens de faces" Dissertação de Mestrado, FEI, São Bernardo do Campo/SP, (2007).

[13] G. Licciardi, P. R. Marpu, J. Chanussot, J. A. Benediktsson, "Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles." IEEE-Geoscience and Remote Sensing Letters, vol. 9, nº. 3, (2012).

[14] N. J. Le Roux, S. Gardner, P. Olivier, "biplots for displaying financial performance graphically." SA Journal of accounting Research, vol. 17, nº 01, pp. 41-64, (2003).

[15] A. H. Monahan, "Nonlinear principal component analysis by neural networks: Theory and application to the Lorenz system," Journal of Climate, American Meteorological Society. vol. 13, pp. 821–835, (2000).

[16] L. Norgaard, B. Rasmus, F. Westad, S. B. Engelsen, "A modification of canonical variates analysis to handle highly collinear multivariate data." Journal of Chemometrics, v. 20, pp.425-435, (2007).

[17] J. Novakovic, S. Rankov, "Classification performance using principal component analysis and different value of the ratio R." Int. J. of Computers, Comunnications e control, vol. 4, no. 02, pp. 317-327, (2011).

[18] R. C. Torres, J. M. Seixas, W. Soares Filho, Classificação de sinais de sonar passivo utilizando componentes principais não lineares. Learning and Nonlinear Models – Revista da Sociedade Brasileira de Redes Neurais, Vol. 2, No. 2, pp. 60-72, (2004).

[19] UCI MACHINES LEARNING REPOSITORY. Universidade da California de Irvine. Disponível em: http://archive.ics.uci.edu/ml/index.html. acesso em: 10/08/2013.

[20] J. J. Verbeek, N. Vlassis, B. Kröse, "A k-segments algorithm for finding principal curves." Elsevier: Pattern Recognition Letters, vol. 23, pp. 1009–1017, (2002).

[21] L. P. Yunsong Qi, S. Huaijiang, "Microarrays data classification basead on principal curves." Seventh International Conference on Fuzzy Systems and Knowledge Discovery, pp. 2199-2202, (2010).

DOI: 10.5540/03.2014.002.01.0057