USO DE ANÁLISE DE COMPONENTES PRINCIPAIS NA SELEÇÃO DE VARIÁVEIS PARA CLASSIFICAÇÃO EM BASES DE DADOS CONTAMINADAS POR RUÍDO BRANCO

Karine S. da Silva[1], Juscelino I. de Oliveira Jr[1], José C. F. da Rocha[1], Alaine M. Guimarães[1]

[1]Mestrado de Computação Aplicada - Universidade Estadual de Ponta Grossa, Setor de Ciências Agrárias e de Tecnologia, Ponta Grossa, PR, Brasil

{karine.sato.silva}@gmail.com, {jrocha, alainemg}@uepg.br

Abstract. *Induced models techniques can be used in the trial to discover knowledge in databases, however, the requirement related to the complexity of the sample might make impracticable the achievement of reliable results. A way to reduce the demands of the sample complexity is to select a subset of variables. This work evaluates how assertions of independence on the variables of the application domain affect the performance of B2 and B4 methods, based on the Principals Components Analysis, in the variable selection for the induction of Artificial Neural Networks. The difference in the performance of the classifiers given the presence or absence of information of independence was determined by experiments performed on synthetic and agricultural data.*

Keywords: *Principals Components Analysis, Artificial Neural Networks, Independence*

Resumo. *Técnicas de indução de modelos podem ser usadas na tentativa de descobrir conhecimento em bases de dados, contudo, o requerimento relativo à complexidade da amostra pode inviabilizar a obtenção de resultados confiáveis. Uma forma de reduzir as exigências da complexidade da amostra é selecionar um subconjunto de variáveis. Este trabalho avalia como asserções de independência sobre as variáveis do domínio de aplicação afetam o desempenho dos métodos B2 e B4, baseados na Análise de Componentes Principais, na seleção de variáveis para indução de Redes Neurais Artificiais. A diferença no desempenho dos classificadores dada a presença ou ausência de informações de independências foi determinada em experimentos realizados sobre bases dados sintéticas e agrícolas.*

Palavras-chave: *Análise de Componentes Principais, Redes Neurais Artificiais, independência*

1. Introdução

A disponibilidade de bases de dados multivariadas sobre agricultura tem motivado o uso de técnicas de Mineração de Dados (MD) na geração de modelos que relacionam as variáveis que influenciam o rendimento da atividade agrícola [Alvarenga e Davide, 1999], [Santos *et al*. 2010]. Contudo, em algumas situações, as bases de dados agrícolas

têm um grande número de atributos (alta dimensionalidade) e um pequeno número de registros, o que dificulta o emprego de métodos multivariados [Lattin *et al*. 2011]. Em bases de dados com alta dimensionalidade pode ocorrer de algumas das variáveis observadas serem redundantes ou terem pouca influência no comportamento dos objetos de interesse.

Um dos problemas relacionados à alta dimensionalidade dos dados é que esta impõe exigências sobre o tamanho da amostra (número de observações) que devem ser usadas na indução dos modelos. Esta questão é capturada pelo conceito de complexidade da amostra na teoria da aprendizagem computacional [Mitchel, 1997]. Este conceito estabelece uma forma de determinar o número de observações necessárias para induzir/regredir um determinado modelo com um grau de precisão e significância arbitrário.

Outro elemento que dificulta a detecção de padrões em base de dados é a presença de ruído [Tan *et al*. 2009]. Dados ruidosos são àqueles que contem erros ou estão fora da faixa aceitável de valores (*outliers*), isso pode ocorrer na coleta dos dados, falhas nos aparelhos de medição ou devido ao erro humano. Em bases de dados de alta dimensionalidade a presença de ruído também pode dar origem a relações espúrias entre as variáveis de entrada bem como mascarar a influência das variáveis preditoras sobre a variável estudada [Ting *et al.* 2006].

Uma possível abordagem para reduzir as exigências referentes à complexidade da amostra e diminuir a influência do ruído nos resultados da análise é o uso de técnicas seleção de variáveis para reduzir a dimensionalidade dos dados [Foster *et al.* 2009] [Staab 2005]. Basicamente, técnicas de seleção de variáveis são empregadas antes da geração de modelos e descoberta de padrões com o objetivo de determinar quais atributos são relevantes para a descrição do fenômeno estudado [Piramuthu, 1998]. Uma possível abordagem para seleção de variáveis é empregar os métodos B2 e B4, baseados na Análise de Componentes Principais (ACP), que foram propostos por Jolliffe (1972).

A abordagem apresentada por Joliffe (1972) não explora o conhecimento sobre o domínio da aplicação quando da seleção de variáveis. No entanto, a incorporação de conhecimento a priori pode ter um grande impacto sobre os resultados da análise e da MD [Sarabjot *et al*. 1995] [Russell & Norvig, 2004]. Diversos trabalhos têm abordado o uso de informação do domínio no cômputo da ACP, por exemplo, pela imposição pela suposição de restrições de independência, expressas em um modelo gráfico gaussiano, sobre a inversa da matriz de covariância [Wiesel e Hero, 2009] [Lauritzen, 2004]. Contudo, poucos trabalhos têm explorado o uso de independência marginal na seleção de variáveis usando ACP, uma forma de conhecimento que é importante para a simplificação de modelos [Silva e Ghahramani, 2009].

Considerando isto, este trabalho propõe o uso de informações a priori durante o processo de seleção de variáveis usando os métodos B2 e B4. A informação a priori é inserida no processo de seleção pela imposição de restrições, baseadas em asserções de independência marginal, que devem ser observadas no cálculo da matriz de covariância amostral, a qual é usada para computar os componentes principais. A abordagem proposta foi avaliada na seleção de variáveis para indução de classificadores baseados em Redes Neurais Artificiais Multicamadas (RNA). Os testes foram realizados sobre bases de dados sintéticas e agrícolas permeadas por diferentes níveis de ruído branco do tipo gaussiano e uniforme. Os resultados mostraram que o uso do conhecimento *a priori*

pode ser útil ao processo de seleção de variáveis permitindo a escolha dos atributos de entrada do classificador mesmo na presença de altos níveis de ruído gaussiano. Para dados permeados por ruído uniforme os resultados não indicam que a disponibilidade de informações *a priori* melhoram o processo de seleção de variáveis.

Este trabalho está organizado da seguinte maneira, a seção 2 apresenta uma breve revisão bibliográfica, com subseções 2.1 referente aos classificadores baseados em RNA, subseção 2.2 sobre a ACP e a subseção 2.3 referente aos métodos de seleção de variáveis. A seção 3 apresenta a metodologia utilizada no trabalho. Os resultados são apresentados na seção 4, em que os resultados do classificador RNA sobre a base de dados sintética são apresentados na subseção 4.1 e os resultados referentes à base de dados agrícolas estão na subseção 4.2. Finalizando o artigo apresentamos a conclusão na seção 5.

2. Revisão bibliográfica

2.1 Classificadores baseados em RNA

Uma das principais tarefas de MD é a classificação, que consiste em descobrir um modelo que mapeie os objetos de um domínio em um conjunto de categorias pré-definidos a partir da análise dos atributos descritores [Tan *et al*. 2009]. Um classificador do tipo Rede Neural de Múltiplas Camadas é um grafo acíclico e direcionado em que os nós são agrupados em camadas. Existe uma camada de entrada que está relacionada às variáveis observadas, uma ou mais camadas de nós intermediários que respondem pelo processamento da rede e uma camada de saída que informa a classificação do objeto sob análise. Os nós da camada de entrada são conectados à primeira camada de nós intermediários e estes sucessivamente à camada de saída. Cada nó implementa uma função sobre suas entradas e comunica o resultado para a próxima camada. Este trabalho utiliza RNA cujos nós implementam função de ativação do tipo sigmóide. Os arcos entre os nós são associados a pesos que ponderam as entradas que um nó recebe de seus antecessores.

Desta forma a RNA relaciona os atributos que descrevem um objeto sob análise com a categoria daquele objeto. A geração deste tipo de modelo pode ser feita com o uso de algoritmos de treinamento como o procedimento de retro-propagação [Haykin, 2001]. Dada a topologia da RNA este algoritmo ajusta os pesos dos arcos da rede a partir de um conjunto de observações armazenadas em uma base de dados. Usualmente, os pesos são ajustados de forma que o erro quadrático médio seja minimizado.

A confiabilidade de um modelo induzido pelo algoritmo de retro-propagação depende da disponibilidade de um número mínimo de observações durante o treinamento [Mitchell, 1997]. No contexto da teoria da aprendizagem computacional aproximadamente correta, a complexidade da amostra para aprender uma RNA com erro ε e significância δ é dada pela Equação 1, em que *m* representa o número de amostras. Nesta expressão, VC indica a dimensão de Vapnik-Chervonenkis que indica o tamanho do espaço de hipóteses de modelos possíveis. A dimensão VC para RNA depende do número de atributos de entrada e é dada pela Equação 2, em que *s* é o número de nós da camada oculta, *r*, é o número de nós da camada de entrada e a constante *e* é a base do logaritmo natural.

$$m \geq \frac{1}{\varepsilon}\left(4log\left(\frac{2}{\delta}\right) + 8VC(H)log\left(\frac{13}{\varepsilon}\right)\right)$$

Equação 1 – Estimativa do número de amostras *m*

$$VC(C) = 2\,s\,(r+1)\log(e\,s)$$

Equação 2 – Dimensão VC para RNA proposta por Mitchell (1997)

Outro fator importante a ser considerado durante o treinamento de RNA é a presença de ruído nos dados [Tan *et al*. 2009]. Neste trabalho, considera-se a presença de ruído branco do tipo gaussiano e uniforme. O ruído branco do tipo gaussiano é um ruído com distribuição normal caracterizado por ter média zero. Já o ruído uniforme é sabidamente não gaussiano e tem sua distribuição uniforme dado um intervalo Δ.

2.2 Análise de Componentes Principais

A ACP é uma técnica multivariada que transforma o conjunto de dados originais, tal que os dados transformados são projetados em um espaço que maximiza a variância da amostra [Jolliffe 2002]. Seja $\mathbf{X}_{nxp}$ uma base de dados com n amostras e p variáveis, e seja $\{X_1 ... X_p\}$ o conjunto de variáveis originais. A ACP determina um novo conjunto de variáveis $\{Y_1 ... Y_p\}$ em que cada Y_i é chamado de componente principal e é uma combinação linear das variáveis originais. As combinações lineares são ortogonais entre si e mantendo o máximo possível da variação dos dados. Os componentes principais podem ser calculados a partir da covariância populacional Ω, contudo, em casos práticos emprega-se a matriz de covariância amostral S usando a fórmula:

$$S = \begin{pmatrix} cov(\boldsymbol{X_1},\boldsymbol{X_1}) & \dots & cov(\boldsymbol{X_1},\boldsymbol{X_p}) \\ \dots & \dots & \dots \\ cov(\boldsymbol{X_p},\boldsymbol{X_1}) & \dots & cov(\boldsymbol{X_p},\boldsymbol{X_p}) \end{pmatrix}$$, onde cov($\mathbf{X}_i$,$\mathbf{X}_j$) é a covariância de $\mathbf{X}_i$ e $\mathbf{X}_j$

Equação 3 – Matriz de covariância, [Ferreira, 2008]

A Equação 3 não considera a existência de conhecimento *a priori* sobre relações de independência entre as variáveis de $\{X_1 ... X_p\}$. Cramer (1999) propõe uma forma de estimar a matriz de covariância quando existe uma hipótese de independência marginal entre determinadas variáveis de **X**. Sejam **L**,**K** e **R** três subconjuntos disjuntos de **X** tal que **K**∪**L**∪**R** ≡ **X** e **L**, **K** ≠ {}. Se as variáveis em **L** e **K** são independentes entre si, então a matriz de covariância $\Omega_{K,L}$ = cov(x_K, x_L) = **0**. Segundo Cramer (1999) a estimativa Ω'' da matriz de covariância, assumindo-s relações de independência entre as variáveis dos subconjuntos **L** e **K** é dada pela Equação 4.

$$\Omega'' = \begin{bmatrix} \Omega' & \Omega' S'_{R*(L,K)} \\ S_{R*(L,K)}\Omega' & S_{R:(L,K)} + S_{R*(L,K)}\Omega S'_{R*(L,K)} \end{bmatrix}$$, onde $\Omega' = \begin{bmatrix} S_{L,L} & 0 \\ 0 & S_{K,K} \end{bmatrix}$

Equação 4 – Teorema proposto por Cramer (1999) assumindo independência entre K e L

2.3 Seleção de variáveis com ACP, os métodos B2 e B4

Joliffe (1972) apresenta os métodos B2 e B4 para seleção de variáveis usando a ACP. O método B2 seleciona um conjunto composto por q variáveis que pertencentes a $\{X_1 ... X_p\}$ tal que as (p-q) variáveis removidas são aquelas associadas com os últimos (p-q) componentes principais. O método B4 tem o mesmo princípio que o método B2, porém as (p-q) variáveis removidas são associadas com os primeiros (p-q) componentes.

3.Material e Métodos

Esta seção apresenta um conjunto de experimentos cujo objetivo foi avaliar o uso de conhecimento a priori, na forma de asserções de independência marginal no processo de escolha das variáveis usando os métodos B2 e B4. Os experimentos foram realizados sobre duas bases de dados, uma sintética e outra agrícola, permeadas por diferentes níveis de ruído.

A base de base de dados sintética contem dez atributos preditores {X1, ...,X10} (variáveis independentes) e um atributo meta W (variável dependente). As variáveis X1, X2, X4 e X7 são independentes entre si e são definidas a partir de quatro variáveis aleatórias latentes, Z1, Z2, Z4 e Z7, com distribuição normal padronizada. As variáveis X3, X5, X6, X8, X9 e X10 são combinações lineares das variáveis latentes Z1..., Z10 definidas como anteriormente. A variável W é dada pela expressão X1 + 0, 8X2 + 0,75X7. Foram geradas 1500 observações deste modelo.

A base de dados agrícola é referente a dados agronômicos com atributos físico-químicos do solo e de produtividade do milho obtidos por meio de equipamentos de agricultura de precisão na região de Campos Novos Paulista – SP, cedida pelo Profº Dr. José Paulo Molin. A base tem vinte atributos preditores e um atributo meta que é a produtividade da cultura em questão, com um total de 2138 observações. Os atributos químicos presentes na base de dados são acidez ativa (pH), acidez total (H_Al), teor de matéria orgânica total (MO), teor de fósforo disponível (P), teor de cálcio trocável (Ca), teor de magnésio trocável (Mg), teor de potássio trocável (K), soma de bases (SB), capacidade de troca de cátions em pH 7 (CTC), saturação por base (V) e saturação por alumínio (m). Já os atributos físicos do solo envolvem a condutividade elétrica superficial (camada de 0 – 20 cm), condutividade elétrica subsuperficial (camada de 20 – 40 cm), e os Índices de Cone (IC) de cada camada do solo (A: 5 – 10 cm, B: 10 – 15 cm, C: 15 – 20 cm, D: 20 – 25 cm, E: 25 – 30 cm, F: 30 – 35cm, G: 35 – 40 cm) [Guimarães, 2005].

A partir da base de dados original foram geradas bases de dados em que os valores originais dos atributos foram adicionados com ruído branco gaussiano com média 0 e desvio padrão 0,25; 0,50; 0,75 ... 5. Para cada valor do desvio padrão foi geradas 20 conjunto de dados. Também foram geradas 20 bases de dados com ruído uniforme com os seguintes intervalos, Δ: [-0,125; 0,125], [-0,25; 0,25], [-0,325; 0,325] [-0,5; 0,5] ... , [-2,5; 2,5]. Dessa forma, ao todo, foram analisadas 400 bases de dados para o ruído branco gaussiano e 420 bases para o ruído uniforme.

Em seguida os métodos B2 e B4 foram usados para selecionar cinco variáveis em cada uma das bases de dados usando o método DVS calculado a partir da matriz de covariância sem independência (S). As variáveis selecionadas foram usadas para treinar uma RNA com cinco nós de entrada, uma camada oculta com cinco neurônios artificiais usando a função sigmóide como função de ativação, e um nó de saída. O desempenho dos classificadores foi determinado pela soma dos quadrados dos erros na classificação dos casos que da base de treinamento.

Na sequência os métodos B2 e B4 foram empregados para selecionar cinco variáveis em cada uma das bases usando a matriz Ω''. Como no caso anterior, as variáveis selecionadas foram usadas no treinamento de RNA e o desempenho de cada modelo foi medido pela soma dos quadrados dos erros na determinação do nível de

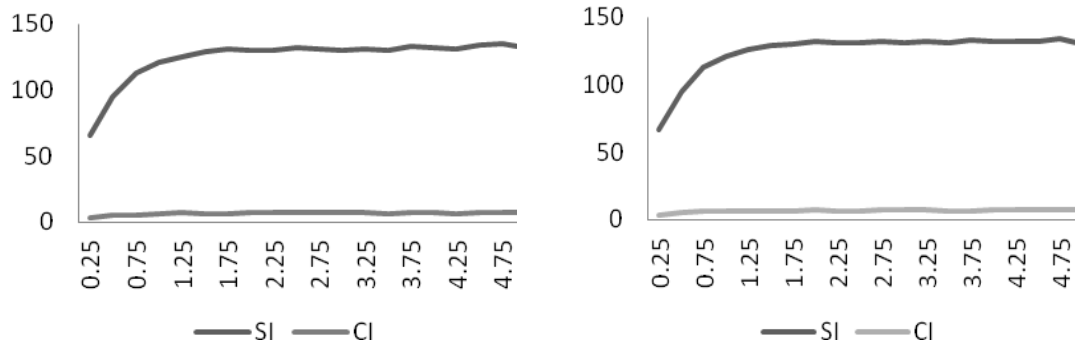
produção de milho. Todos os experimentos foram realizados utilizando a ferramenta estatística R, versão 2.14.0.

4. Resultados e Discussão

4.1 Base de dados sintética

A Figura 1 mostra os gráficos da média dos desempenhos dos classificadores gerados com as variáveis selecionadas pelos métodos B2 e B4 sobre a base de dados sintética na presença de ruído branco do tipo gaussiano. A Figura 1(a) apresenta o resultado obtido com o uso do método B2 e a Figura 1(b) mostra o resultado do método B4. Em ambos os gráficos são mostrados os resultados da ACP considerando-se a matriz de covariância amostral S (SI) e a matriz de covariância amostral com asserções de independência Ω'' (CI). Como pode ser observado nos gráficos o erro sofreu uma redução de uma ordem de magnitude na soma dos quadrados dos erros quando a RNA foi gerada a partir dos dados selecionados com o uso da matriz Ω'' no cálculo da ACP.

A queda na soma do quadrado dos erros foi de uma ordem de magnitude e pode estar relacionada ao fato de que ao fixar-se as relações de independência entre as variáveis, alguns elementos da matriz de covariância não podem ser afetados por relações espúrias presentes nos dados mesmo na presença de ruído. Além disto, sejam X1 e X2 duas variáveis independentes, se ambas estiverem em um mesmo componente principal *P* elas não contribuem para o autovalor fazendo mais provável que o mesmo seja rejeitado pelos métodos B2 e B4, isto pode aumentar a possibilidade de *P* ser rejeitado pelos métodos B2 e B4.



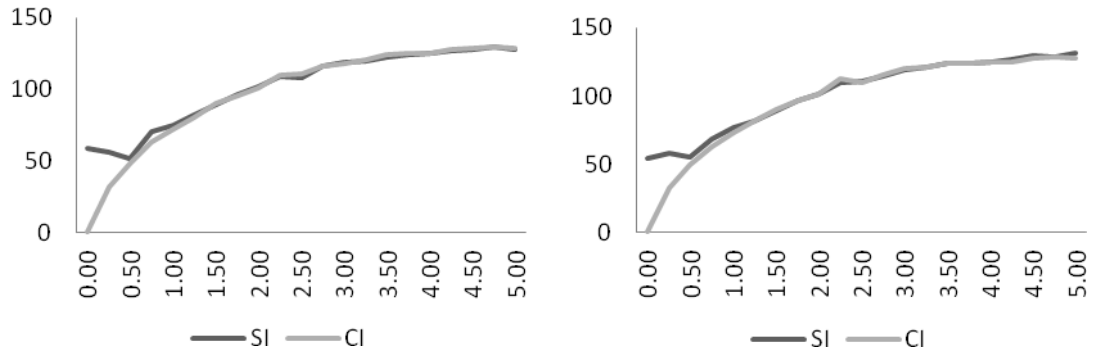
(a) (b)

Figura 1(a) e (b) – Resultados obtidos a partir dos dados sintéticos com ruído gaussiano

A Figura 2 mostra os gráficos do desempenho da RNA utilizando os métodos B2 e B4 para a seleção das variáveis sobre a base de dados sintética na presença de ruído uniforme. A Figura 2(a) apresenta o resultado obtido com o uso do método B2 e a Figura 2(b) mostra o resultado do método B4. Assim como na Figura 1, em ambos os gráficos são mostrados os resultados da ACP considerando-se a matriz de covariância amostral S (SI) e a matriz de covariância amostral com asserções de independência Ω'' (CI).

Pode ser observado que o impacto do ruído branco uniforme sobre o desempenho do processo de seleção e classificação é maior do que aquele observado no caso do ruído gaussiano. Na Figura 2, é possível observar que a seleção de variável

usando relações independência tem um desempenho superior ao da abordagem que não explora este tipo de informação para valores de experimentos em que $(\Delta \subseteq [-1;1])$.



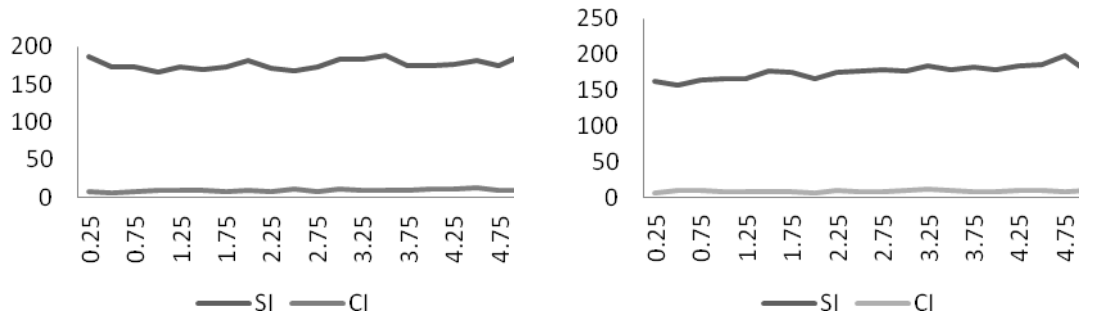
(a) (b)

Figura 2 (a) e (b) – Resultados obtidos a partir dos dados sintéticos com ruído uniforme

4.2 Base de dados real

Apesar de muitos estudos analisarem bases de dados em que constam atributos referentes às propriedades físicas e químicas do solo [Carneiro *et al*. 2009], [Gomes *et al*. 2004] estes trabalhos não consideram o fato de que alguns atributos podem ser independentes entre si. Considerando isto, este experimento explorou o fato de que o Índice de Cone (IC) independe dos valores do P, K, Mg, Ca, H_Al, pH e V em todas as camadas no cômputo da matriz de covariância.

Os resultados do classificador RNA para a base de dados agrícola com ruído branco gaussiano são apresentados na Figura 3 (a) e (b). A Figura 3(a) mostra o resultados de SI e CI a partir da seleção de variáveis utilizando o método B2, e a Figura 3(b) apresenta os resultados SI e CI a partir do método B4. Assim como na Figura 1(a) e (b), os gráficos apresentados na Figura 3(a) e 3(b) apresentam o mesmo comportamento, em que utilizando a matriz de covariância com relações de independência Ω'' tem-se a soma dos quadrados dos erro reduzida.

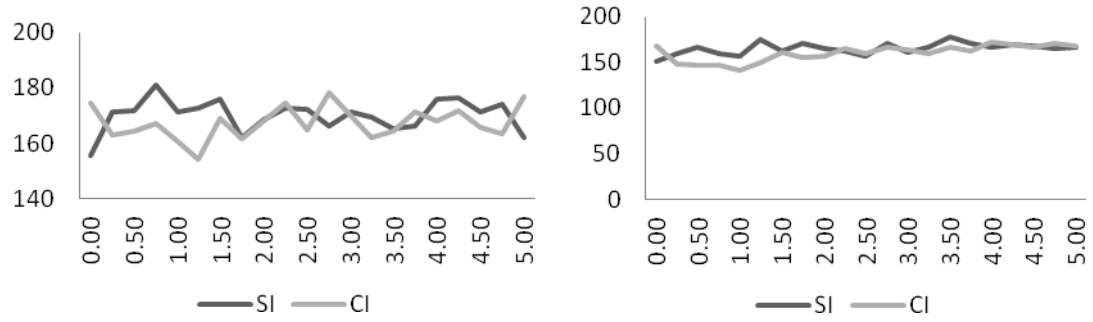


(a) (b)

Figura 3(a) e 3(b) – Resultados obtidos a partir dos dados reais com ruído gaussiano

A Figura 4(a) mostra o resultados dos processos de seleção SI e CI com o método B2 utilizando o ruído uniforme, e a Figura 4(b) apresenta os resultados de SI e CI com o método B4. Nestes testes não se observou um padrão que indicasse que o uso de informações de independência durante a seleção de variáveis melhorasse o

desempenho da etapa de classificador na presença de ruído uniforme para os dados agrícolas.



(a) (b)

Figura 4(a) e 4(b) – Resultados obtidos a partir dos dados reais com ruído uniforme

5. Conclusão

Este trabalho avaliou o uso de informações de independência no cômputo da matriz de covariância de uma amostra de dados para seleção de variáveis com Análise de Componentes Principais. A seleção de variáveis foi realizada por meio dos métodos B2 e B4, propostos por Joliffe (1972). Os experimentos foram realizados sobre bases de dados sintéticas e agrícolas permeadas por ruído branco gaussiano e uniforme. Os resultados mostram que quando as amostras são permeadas por ruído gaussiano o uso de informações de independência marginal durante a seleção de variáveis tem um efeito positivo na indução do modelo de classificação. Isto reflete no fato de que os modelos gerados apresentam um melhor desempenho quando da classificação de novas amostras. No caso do ruído uniforme a melhora de desempenho foi observada apenas para pequenos níveis de ruído na base de dados sintéticas, cuja relação entre as variáveis era linear.

A diferença no desempenho dos procedimentos de seleção de variáveis e de classificação observados nas Figuras 1 e 3, e Figuras 2 e 4 parece estar diretamente ligada ao tipo ruído que permeia os dados. Segundo Tiwari *et al.* (2011) a presença de ruído uniforme tem um efeito adverso sobre procedimentos de análise de dados que se baseia em modelos lineares, como é o caso da seleção de variáveis baseada ACP.

Como trabalho futuro sugere-se a utilização de outros métodos de seleção de variáveis baseadas em Análise de Componentes Principais. Finalmente, é sugerida a realização de experimentos que usam Análise de Componentes Principais Supervisionada como método de seleção de variáveis.

6. Agradecimentos

A CAPES pela bolsa de estudos e ao Profº Dr. José Paulo Molin pelo fornecimento da base de dados agrícola para este estudo.

7. Referências Bibliográficas

Alvarenga, M.I.N. e Davide, A.C. (1999) "Características físicas e químicas de um latossolo vermelho-escuro e a sustentabilidade de agroecossistemas", Revista Brasileira de Ciência do Solo, v.23, p. 933-942.

Carneiro, M.A.C., Souza, E.D., Reis, E.F., Pereira, H.S. e Azevedo, W.R. (2009) "Atributos físicos, químicos e biológicos de solo de cerrado sob diferentes sistemas de uso e manejo", Revista Brasileira de Ciência do Solo, v.33, p. 147-157.

Cramer, E. (1999) "Estimation of the mean and the covariance matrix under a marginal independence assumption – an application of matrix differential calculus", Linear Algebra and its Applications, v. 288, p. 219-228.

Gomes, J. B. V., Curi, N., Motta, P. E. F., Ker, J. C., Marques, J. J. G. S. M. e Schulze, D. G. (2004) "Análise de componentes principais de atributos físicos, químicos e mineralógicos de solos do bioma cerrado" Revista Brasileira de Ciência do Solo, v.28, 137-153.

Ferreira, D.F. (2208). Estatística Multivariada. UFLA, 1º edição.

Foster, D.P., Kakade, S.M. e Zhang, T. (2009) "Multi-view dimensionality reduction via canonical correlation analysis", Relatório Técnico – Chicago.

Guimarães, A.M. (2005) "Aplicação de computação evolucionária na mineração de dados físico-químicos da água e do solo", Tese – Faculdade de Ciências Agronômicas da UNESP.

Haykin, S. S. (2001) Redes Neurais - Principios e Prática, Bookman Companhia.

Kantardzic, M. (2002) Data mining : concepts, models, methods and algorithms. New York : IEEE.

Jolliffe, I. T. (1972) "Discarding Variables in a Principal Component Analysis. I: Artificial Data", Journal of the Royal Statistical Society, v. 21, p. 160-173.

Jolliffe, I. T. (2002) Principal Component Analysis. Springer, 2º edição.

Lattin, J., Carroll, J. D. e Green, P. E. (2011) Análise de Dados Multivariados. Cengage Learning, 1º edição.

Lauritzen, S.L. (2004) Graphical Models. Oxford University Press.

Libralon, G. L. (2007) "Investigação de combinações de técnicas de detecção de ruído para dados de expressão gênica", Dissertação. Instituto de Ciências Matemáticas e de Computação - USP.

Mitchell T. (1997) Machine Learning. New York: McGraw-Hill.

Nettleton, D.F., Puig, A. O. e Fornells, A. (2010) "A study of the effect of different types of noise on the precision of supervised learning techniques", Journal Artificial Intelligence, v. 33, 275-306.

Piramuthu, S. (1998) "Evaluating Feature Selection Methods for Learning in Data Mining Applications", European Journal of Operational Research, v. 5, 483-494.

Russel, S. J. e Norvig, P. (2004). Artificial intelligence: a modern approach. Prentice Hall. (Prentice Hall series in artificial intelligence).

Santos, J.S., Santos, M.L.P. e Conti, M.M. (2010). "Comparative Study of Metal Contents in Brazilian Coffees Cultivated by Conventional and Organic Agriculture Applying Principal Component Analysis", Journal of Brazilian Chemical Society, v. 21, p. 1468-1476.

Sarabjot, S.A., Bell, D.A. e Hughes, J.G. (1995) "The role of domain knowledge in data mining" Proceedings of the 4th international conference on information and knowledge management, p. 37-43.

Silva, R. e Ghahramani, Z. (2009) "Factorial mixture of gaussians and the marginal independence model", Proceedings of the 12th international conference on artificial intelligence and statistics, p. 520-527.

Staab, B. (2005) "Investigation of noise and dimensionality reduction transforms on hyperspectral data as applied to target detection", Chester F. Carlson Center og Imaging Science, Rochester Institute of Technology.

Tan, P. N., Steinbach, M. e Kumar, V. (2009) Introdução ao Data Mining. Ciência Moderna, 1º edição.

Tiwari, S., Singh, A.K. e Shukla, V.P. (2011) "Statistical moments based noise classification using feed forward back propagation neural network", International journal of computer applications, v. 18, n. 2, p. 36-40.

Wiesel, A. e Hero, A.O.III (2009) "Decomposable principal component analysis", IEEE Transactions on signal processing, v. 57, n. 11, p. 4369-4377.