UNIVERSIDADE FEDERAL DE UBERLÂNDIA - UFU

PLANO DE TRABALHO EM INICIAÇÃO CIENTÍFICA

Desenvolvimento de toolbox de análise multivariada para o matlab.

Uberlândia

28/09/10

IDENTIFICAÇÃO DO TRABALHO

Título

Desenvolvimento de toolbox de análise multivariada para o matlab.

Resumo

O Matlab é um sistema interativo cujo elemento básico de informação é uma matriz que não requer dimensionamento. Trata-se de um software interativo de alto desempenho voltado para o cálculo numérico, uma ferramenta bastante utilizada em ambientes de pesquisas acadêmicas. Tal utilização se deve principalmente as inúmeras facilidades que ele fornece para o desenvolvimento e testes de novos métodos. Apesar dos benefícios existentes durante o desenvolvimento, se nota uma ausência de ferramentas quanto a realização de testes estatísticos dos métodos desenvolvidos. Essa carência estatística faz com que seja preciso a utilização de outros programas para realizar a etapa de análise dos dados, como o R Project ou o Weka (ambos softwares livre, de origem neozelandesa). Apesar de estes conseguirem resolver o problema existe, uma perda de desempenho pode ocorrer nessa etapa, visto que é necessário trocar de software para realização das análises. Além disso, para a utilização dos dados provenientes do Matlab nos demais softwares, faz-se necessário que os mesmo sejam convertidos para um formato que os mesmos sejam capazes de reconhecer e salvos em discos. Esse processo pode levar a erros de arredondamento e outros tipos de alterações nos dados, ainda que sutis, mas que podem comprometer a etapa de análise, levando a resultados diferentes dos esperados.

INTRODUÇÃO

Matlab é um software de alto desempenho criado no fim dos anos 70 por Cleve Moler, da Universidade do novo México, altamente voltado ao cálculo numérico. É um sistema interativo cujo elemento básico de informação é uma matriz que não requer dimensionamento. Esse sistema permite a resolução de muitos problemas numéricos em apenas uma fração do tempo que se gastaria para escrever um programa semelhante em linguagens como Fortran, Basic ou C, por exemplo. O Matlab utiliza uma linguagem própria e integra análise numérica, cálculo com matrizes, processamento de sinais e construção de gráficos em um ambiente fácil de usar, onde problemas e soluções são expressos como eles são escritos matematicamente, ao contrário da programação tradicional. Foi inicialmente adotado por engenheiros de projeto de controle, e rapidamente se espalhou para outros campos. E é inclusive muito popular no ramo de processamento de imagens e utilizado no ensino de álgebra linear e análise numérica [1].

O Matlab é muito usado em ambientes acadêmicos, mais especificamente na área de pesquisas. Por apresentar um fácil manuseio e grande potencial de funcionalidades devido a vasta quantidade de métodos implementados disponíveis, esses métodos são implementados através dos toolboxes, bibliotecas com metodologias utilizáveis para fins específicos à áreas desejadas. Porém, apesar de tantos benefícios, ele ainda apresenta uma carência no que diz respeito a métodos estatísticos, fazendo com que seja preciso utilizar softwares alternativos para auxiliar na etapa de análise dos dados, como o R [2,13] e o Weka [10,14].

R é o nome de um popular programa que está em uso por crescente número de analistas de dados, em empresas e no mundo acadêmico. Seu uso tem se tornado padrão porque os processos de mineração de dados vivem uma era dourada, quer estejam em uso para determinar preços de publicidade, descobrir novos medicamentos mais rápido ou fazer a sintonia fina de modelos financeiros. Empresas as mais diversas, como por exemplo, Google, Pfizer, Merck, Bank of America e Shell, usam o R. Mas a R também encontrou rápida aceitação entre os estatísticos, engenheiros e cientistas que não conhecem bem a programação de computadores e o consideram fácil de usar. Isso por que ele permite que os estatísticos realizem análises muito intricadas e complicadas sem que precisem conhecer em detalhe o funcionamento dos sistemas de computação [2,13].

Já o software Weka (*Waikato Environment for Knowledge Analysis*) começou a ser escrito em 1993, usando Java, na Universidade de Wakato, Nova Zelândia, sendo adquirido posteriormente por uma empresa no final de 2006. O Weka encontra-se licenciado ao abrigo da General Public License, sendo portanto possível estudar e alterar o respectivo código fonte. O Weka tem como objetivo agregar algoritmos provenientes de diferentes abordagens/paradigmas na sub-área da inteligência artificial dedicada ao estudo da aprendizagem por parte de máquinas. Essa sub-área pretende desenvolver algoritmos e técnicas que permitam a um computador "aprender" (no sentido de obter novo conhecimento), quer indutiva, quer dedutivamente.O Weka procede à análise computacional e estatística dos dados fornecidos recorrendo a técnicas de *data-minning* tentando, indutivamente, a partir dos padrões encontrados, gerar hipóteses para soluções e no extremos inclusive teorias sobre os dados em questão [10,14].

Entretanto a artimanha de utilizar esses Softwares não é eficiente, porque para utilização de dados providos do Matlab em outro programa, é necessário, primeiro, a conversão da linguagem para uma correspondente ao pretendido. E essas conversões geram eventuais transtornos aos dados como erros de arredondamento e de outros tipos . E o plano de trabalho tem como intuito fazer a implementação no toolbox do matlab, essa implementação o aperfeiçoará tornando-o mais eficiente na resolução de problemas estatísticos e promovendo alto desempenho na etapa de análise de dados.

OBJETIVO

O objetivo deste projeto é a implementação de toolbox de análise multivariada estatística, ciência que se dedica à coleta, análise e interpretação de dados, a qual se utiliza das teorias probabilísticas para explicar a freqüência de fenômenos e para possibilitar a previsão destes no futuro.

Algumas práticas estatísticas incluem, por exemplo, o planejamento, a sumarização e a interpretação de observações. Dado que o objetivo da estatística é a produção da melhor informação possível a partir dos dados disponíveis, alguns autores sugerem que a estatística é um ramo da teoria da decisão.

METODOLOGIA

Levantamento das técnicas estatísticas a serem implementadas

Uma parte importante do trabalho diz respeito ao levantamento das técnicas a serem implementadas no toolbox. Para tanto, é necessário realizar uma busca sobre os principais modelos probabilísticos e classificadores existentes na literatura, suas diferenças e particularidades, bem como métodos auxiliares as etapas de classificação, como o z-score (o qual permite normalizar os dados para média igual a zero e desvio padrão igual a um). Além disso, pretende-se também investigar os diferentes métodos de redução de dimensionalidade, para seleção e extração de características, o que torna mais simples a etapa de classificação das amostras [3-9,11-12].

Implementação de diferentes classificadores estatísticos

Na literatura, é possível encontrar uma série de diferentes classificadores. Muitos deles são baseados no Teorema de Bayes, o qual descreve a relação entre uma probabilidade condicional e sua inversa, i.e., a probabilidade de uma hipótese dada a observação de uma evidência e a probabilidade da evidência dada pela hipótese:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

onde P(A) e P(B) são as probabilidades a priori de A e B, e P(B|A) e P(A|B) são as probabilidades a posteriori de B condicional a A e de A condicional a B respectivamente [3-9,11-12].

Dentre os classificadores baseados no teorema de Bayes, podemos facilmente citar o próprio classificador Bayesiano, o método de LDA (Linear Discriminant Analysis) e QDA (Quadratic Discriminant Analysis). Essas três abordagens diferem principalmente pelo fato dos dois últimos aplicarem transformações sobre os dados sob análise, enquanto o classificador Bayesiano não. Essas tranformações tem o intuito de encontrar uma projeção dos dados em um espaço de características onde a variação entre classes é maximizada, enquanto a variância dentro da classe é minimizada. Para tanto, o método de LDA utiliza uma única matriz de covariâncias para todas as

observações para realizar essa transformação, enquanto o QDA utiliza matrizes de covariâncias diferentes para cada classe, tendo, por esse mesmo motivo, um custo computacional maior em relação aos demais citados [3-9,11-12].

Implementação de diferentes estratégias de Validação Cruzada

A validação cruzada é uma técnica muito utilizada para determinar como os resultados de uma análise estatística serão generalizados para conjuntos de dados independentes. Basicamente, essa técnica consiste em dividir um conjunto de observações em subconjuntos complementares: o conjunto de treinamento (onde será realizada a análise) e o de teste (validação da análise) [3-9,11-12].

Na literatura, podemos encontrar várias estratégias de validação cruzada. Abaixo, são apresentadas algumas das quais se pretende implementar neste trabalho, bem como uma breve descrição delas.

- Hold-out validation: nele, um subconjunto de observações é escolhido aleatoriamente a partir da amostra inicial para formar um conjunto de teste (normalmente, menos de um terço da amostra inicial), sendo as observações restantes mantidos como dados de treinamento.
- K-fold cross-validation: neste caso, o conjunto de observações é dividido em K subconjuntos disjuntos. Para cada partição, as observações nela contidas são utilizadas para validação, enquanto os dados das K-1 partições restantes são designados ao grupo de treinamento. Este processo é repetido K vezes, de modo que cada partição seja utilizada uma vez para validação dos dados.
- Leave-one-out cross-validation: trata-se de um caso particular tanto do K-fold cross-validation. Neste caso, K é igual ao número total de observações. Desse modo, cada partição possui um único elemento, o qual será utilizado na etapa de validação, enquanto o conjunto de treinamento é composto por todas as observações restantes.

Validação dos diferentes classificadores estatísticos implementados

A etapa de validação de cada abordagem implementada se dará por meio do uso de bases de dados conhecidas e disponíveis na literatura e, obviamente, pela comparação com softwares estatísticos conhecidos (e.g., Weka e R Project).

CRONOGRAMA

Obs.: Vide Legenda

Atividades	2011										2012	
	03	04	05	06	07	08	09	10	11	12	01	02
A												
B												
C												
D												
E												
F												

Legenda:

A- Conhecimento de Matlab e conceitos de softwares R e Weka.

B- Levantamento das bases estatísticas e das técnicas a serem implementadas no toolbox.

C- Investigação dos diferentes métodos de redução de dimensionalidade, para seleção e extração de características, e classificação de amostras.

D- Implementação e teste dos métodos que compõem o Toolbox.

E- Teste e validação com dados conhecidos e comparação com outros softwares estatísticos.

F- Relatório Final.

RECURSOS NECESSÁRIOS

Para realização do projeto será utilizado o software Matlab, recursos imprescindíveis como acesso à internet, laboratório de informática e referências literárias tanto referentes ao programa quanto às bases estatísticas.

RESULTADOS ESPERADOS

Ao fim do projeto, espera-se como resultado uma implementação eficiente de um conjunto amplo de métodos estatísticos para classificação e análise dos dados no Matlab, através de toolbox, que proporcione ao usuário satisfação quanto as necessidades estatísticas na resolução de problemas e promovendo alto desempenho na etapa de análise de dados multivariados.

BIBLIOGRAFIA

[1] Blanchet, G., Charbit, M., Digital Signal and Image Processing Using MATLAB, Wiley-ISTE (2006)

[2] Chamber J.M., Software for Data Analysis - Programming with R. (first ed.), Springer (2008).

[3] Everitt , B.S. Dunn, G., Applied Multivariate Analysis (second ed.), Arnold (2001).

[4] Ferreira, D. F.. Apostila de análise multivariada, Universidade Federal de Lavras (1996).

[5] Fukunaga , K., Introduction to Statistical Pattern Recognition (second ed.), Academic Press (1990).

[6] Johnson, R. A.., Wichern, D. W.., Applied Multivariate Statistical Analysis. 4ª Edição, Amazon (1998).

[7] Manly , B. J. F.. Métodos estatísticos multivariados, 3ª Edição Editora: ARTMED -.

[8] Mesquita, J. M. C. de. Estatística multivariada aplicada à Administração: guia prático para utilização do SPSS. Editora: CRV (2010).

[9] Mingoti, S. A.. Análise de Dados Através de Métodos de Estatística Multivariada. Editora UFMG (2005).

[10] Silva M. P. S., Apostila Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka, Universidade do Estado do Rio Grande do Norte (2004).

[11] Theodoridis , S., Koutroumbas , K., Pattern Recognition (second ed.), Academic Press (2003).

[12] Timm, N.H.., Applied Multivariate Analysis. 1ª Edição, Springer (2002).

[13] Torgo L., Linguagem R - Programação para a Análise de Dados. 1ª Edição, Escolar (2009).

[14] Witten I. H. e Frank E., Data Mining: Practical Machine Learning Tools and Techniques, (Second Ed.), Amazon (2005).