Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Classificação supervisionada baseada em árvore geradora mínima

Letícia Cavalari Pinheiro[1,3]
Renato Martins Assunção[2]

1 Introdução

Classificação supervisionada é um dos problemas mais estudados na área de aprendizagem de máquina, com excelentes algoritmos disponíveis: desde mais simples como a Regressão Logística até mais complexos como as Florestas Aleatórias e as Máquinas de Suporte de Vetores (SVM). Neste trabalho serão considerados os seguintes métodos: Regressão Logística, Árvores de Regressão e Classificação (CART) [1], Florestas Aleatórias [4] e Máquinas de Suporte de Vetores (SVM)[5] [6].

Cada método tem suas qualidades, assim como seus pontos fracos. Com o objetivo de propor um método simples e eficiente para classificação supervisionada, principalmente nos casos em que alguns dos métodos tradicionais não têm bom desempenho, apresentamos o método "SKATER Não-Espacial" (NSS). Nosso método tem a vantagem de detectar clusters com formatos diversos e apresentou resultados satisfatórios diante dos métodos tradicionalmente utilizados.

2 Materiais e métodos

O Método "Non-Spatial SKATER"(NSS) utiliza o conceito de árvore geradora mínima. Sua idéia foi baseada no "Spatial Clustering Analysis Through Edge Removal"(SKATER) [2], que é um método de agrupamento (clustering) espacial para dados de área, baseado no conceito de árvore geradora mínima.

Suponha que os dados são compostos de n elementos, cada um com k atributos e uma resposta binária para a característica que prentede-se classificar. O vetor que caracteriza um elemento é composto por k atributos x_i's e uma resposta binária y_i: (x_{i1}; ...; x_{ik}; y_i). Cada elemento é identificado como um nó em um grafo não direcionado G = (V ;E). O custo associado à aresta (v_i; v_j) é 1/d(i; j), onde d(i; j) é uma das distâncias estatísticas entre vetores existentes, tal como a distância Euclidiana ou a de Mahalanobis. Essa distância é calculada com base nos vetores de atributos (x_{i1}; ...; x_{ik}), representando a distância par a par entre os elementos.

1 EST – ICEX/UFMG. E-mail: leticia.ufmg@gmail.com
2 DCC – ICEX/UFMG
3 CPqRR – FIOCRUZ/MG

Definimos o número de clusters a serem identificados como C. A partir do grafo obtido, uma árvore geradora mínima (AGM) é construída com base nos custos calculados entre os pares de elementos. A poda da árvore é feita sequencialmente. A cada iteração, é calculada a medida de heterogeneidade para cada aresta possível de ser retirada:

$$Q(G_1;...;G_C) = \sum_{q=1}^{C} \sum_{i \in G_q} (y_i - \bar{y}_q)^2 \quad (1)$$

onde p_q é a proporção de 1's no grupo q. A aresta que mais decresce essa medida é selecionada. Após C-1 arestas cortadas, C clusters estarão formados. Cada um desses clusters formados será classificado de acordo com a proporção de respostas 0's e 1's de seus elementos. Um grupo com maior proporção de elementos cuja resposta é 0 será classificado como 0 e, analogamente, um grupo com maior proporção de elementos cuja resposta é 1 será classificado como 1.

Formados os clusters, deve-se definir a forma de classificação para novos elementos sem rótulo (ou seja, sem o valor da resposta binária y_i). Esses novos elementos serão classificados no grupo que contém o seu vetor de atributos. Este grupo é o que contém o vetor de atributos mais próximo ao novo vetor. Por exemplo, observa-se que o novo vetor x_{novo} tem o vetor de atributos mais próximo do vetor de atributos x_a, e o x_a pertence a um cluster que foi rotulado como 1 por ter maior proporção de 1's entre as respostas dos elementos que o compõem. Portanto, x_{novo} será classificado como 1. Este é o formato básico do método proposto neste trabalho. Porém, durante sua implementação, surgiram algumas idéias que poderiam melhorar ainda mais os resultados alcançados para os problemas de classificação a serem resolvidos. A partir dessas idéias foram introduzidas pequenas modificações no método. Essas modificações influenciam em alguns detalhes do método proposto e são chamadas de NSSY, RNSS e RNSSY. A seguir, apresentaremos resultados das 4 versões desenvolvidas comparados aos resultados dos métodos já existentes.

3 Resultados

Para avaliar o método, foram realizadas simulações de cinco formatos de conjuntos de dados com atributos em apenas 2 dimensões. Essa escolha das dimensões se justifica pela facilidade em observar a disposição dos pontos que correspondem aos valores dos atributos no plano cartesiano. Para efeitos de comparação e avaliação da eficiência do método desenvolvido, foram aplicados aos dados os métodos propostos NSS, RNSS, NSSY e RNSSY e outros quatro métodos populares de classificação supervisionada: Regressão Logística, Árvores de Classificação e Regressão(CART), Florestas Aleatórias e Máquinas de Vetores de Suporte (SVM). As medidas observadas para avaliação foram: Percentual de acertos, Sensibilidade (Revocação), Especificidade e Precisão. A seguir, são apresentados os resultados de percentual de acertos de cada método.

Percentual de acertos								
	NSS	NSSY	RNSS	RNSSY	Reg. Log.	CART	Rand. For.	SVM
Conjunto Simulado 1	0,96	1	0,96	1	0,61	0,87	0,95	0,98
Conjunto Simulado 2	0,845	0,905	0,87	0,905	0,555	0,84	0,825	0,9
Conjunto Simulado 3	0,85	0,96	0,86	0,95	0,72	0,98	0,99	0,94
Conjunto Simulado 4	0,91	0,95	0,91	0,95	0,61	0,82	0,88	0,95
Conjunto Simulado 5	0,89	1	0,94	1	0,61	0,95	0,99	0,98

Os problemas reais de classificação com os quais costumamos nos deparar envolvem muitas dimensões, de forma a não ser possível uma clara visualização do "formato"em que os pontos estão distribuídos no espaço, como foi possível com os dados simulados em duas dimensões. Dessa forma, fica difícil prever qual dos métodos de classificação seria supostamente mais adequado a cada conjunto de dados analisado. Diante dessa situação, foram selecionados 4 bancos de dados disponíveis na internet [3], aos quais foram aplicadas as quatro variações do método criado, além da Regressão Logística, do CART, das Florestas Aleatórias e do SVM, para efeitos de comparação dos resultados.

a) Conjunto de Dados Ionosphere: Os dados deste conjunto foram coletados em Goose Bay, Labrador, no Canadá, por um sistema composto por uma matriz de fases de 16 antenas de alta frequência. "Bons"retornos de radar são aqueles que mostram evidência de algum tipo de estrutura na ionosfera. Retornos "ruins"são aqueles que não o fazem, seus sinais passam através da ionosfera. Os sinais recebidos foram processados usando uma função de autocorrelação cujos argumentos são o tempo de um pulso e o número do pulso. Houve 17 números de pulsos para o sistema de Goose Bay. Instâncias nesta base de dados são descritas por dois atributos por número de pulso, correspondendo aos valores complexos devolvidos pela função resultante a partir do sinal eletromagnético complexo. Dessa forma, o conjunto de dados é composto por 34 atributos contínuos e por uma resposta binária "bom"ou "ruim". Este banco tem 350 elementos.

b) Wisconsin Breast Cancer Dataset: Os dados deste conjunto foram obtidos da Universidade de Wisconsin, Madison, pelo Dr. William H. Wolberg. Foram analisados casos de tumores de mama entre janeiro de 1989 e outubro de 1991. Para cada elemento do conjunto de dados há nove medidas (relativas a tamanho de célula, formato de célula, etc) que variam nos números inteiros entre 1 e 10, e uma resposta binária "maligno"ou "benigno". Este banco é composto por 680 elementos.

c) Wisconsin Diagnosis Breast Cancer (WDBC): As características presentes nesses dados foram calculadas a partir de uma imagem digitalizada de um aspirado de agulha fina de uma massa de mama. Eles descrevem as características dos núcleos das células presentes na imagem. Para cada elemento do conjunto de dados há 30 atributos numéricos e uma resposta binária "maligno"ou "benigno". Este banco é composto por 565 elementos.

d) German Credit Data: Os dados deste banco são compostos por diversas características de candidatos a crédito na Alemanha, como posses, sexo, estado civil, tempo no atual emprego, etc. O banco de dados originais é composto por 20 atributos, que foram

transformados para possibilitar a aplicação direta dos métodos de classificação supervisionada. Dessa forma, cada elemento é composto por 30 atributos (numéricos, binários e categóricos ordinais) e uma resposta binária "bom pagador"ou "mau pagador". Este banco é composto por 1000 elementos.

Cada um dos conjuntos de dados foi dividido em um conjunto de construção do modelo e outro de avaliação. Para evitar qualquer tipo de tendência nessa divisão, utilizamos amostragem sistemática para selecionar o grupo de avaliação do modelo. Os métodos foram aplicados 5 vezes a cada conjunto de dados, variando os conjuntos de construção e validação do modelo. Os resultados das 5 aplicações a cada conjunto de dados são armazenados e utilizados para o cálculo das medidas (médias) de avaliação do modelo (Percentual de Acertos, Sensibilidade (Revocação), Especificidade, Taxa de Falso Positivo e Precisão). A seguir, são apresentados os resultados de percentual de acertos de cada método.

Percentual de acertos								
	NSS	NSSY	RNSS	RNSSY	Reg. Log.	CART	Rand. For.	SVM
Conjunto Real 1	0,897	0,894	0,906	0,9	0,88	0,866	0,931	0,926
Conjunto Real 2	0,963	0,92	0,954	0,912	0,95	0,935	0,962	0,964
Conjunto Real 3	0,954	0,958	0,961	0,965	0,95	0,926	0,963	0,972
Conjunto Real 4	0,683	0,683	0,707	0,665	0,7	0,708	0,693	0,716

4 Conclusão

Para os conjuntos de dados simulados: Foram simulados diversos cenários, alguns favoráveis e outros desfavoráveis a algum dos métodos já existentes. Para todos eles, foi realizada a comparação do desempenho das variações apresentadas do método proposto neste trabalho com o desempenho dos outros métodos popularmente conhecidos atualmente. A partir dos resultados apresentados, foi possível observar que, no geral, o método proposto é bastante eficiente. Nos cenários 1 e 5 tivemos duas variações do nosso método com desempenho perfeito, classificando todos os pontos corretamente. Nos outros três cenários, foram observados valores razoáveis para as medidas avaliadas na maior parte das variações do nosso método. No geral, os resultados são sempre os melhores ou então estão próximos dos melhores. Dessa forma, concluímos que para os dados simulados o método proposto obteve grande sucesso nos testes realizados, se mostrando uma ótima opção para classificação supervisionada em casos em que os métodos tradicionais não funcionam tão bem ou, algumas vezes, até mesmo em cenários onde eles funcionam bem.

Para os conjuntos de dados reais: É interessante pensar que foram realizados testes com conjuntos de dados disponíveis e já utilizados anteriormente em testes de métodos de classificação supervisionada. Isso poderia ser um indício de que os bancos teriam sido usados como exemplos de sucesso para algum dos métodos que estamos comparando ao que criamos. Mesmo assim, foi observado que as variações do método proposto não ficaram com resultados muito inferiores em nenhum dos casos. Mais uma vez o nosso método se mostrou bastante competitivo com os demais. Além disso, podemos observar uma regularidade dos resultados, de forma que em nenhum dos casos apresentados, houve um desempenho consideravelmente ruim do nosso método em relação aos outros testados.

Concluímos observando que para alguns conjuntos de dados, os métodos propostos tiveram resultados mais satisfatórios que os demais métodos. Esses foram apenas alguns casos, mas na maior parte dos conjuntos de dados, o desempenho dos métodos propostos foi bastante satisfatório, ficando bem próximo dos que obtiveram melhores resultados.

Dessa forma, tais métodos se mostraram uma boa opção para a realização de classificação supervisionada, visto que obtiveram bom desempenho e que foram constantes, característica que não foi observada em alguns outros métodos, que algumas vezes apresentavam ótimos resultados e outras apresentavam resultados ruins.

5 Referências

[1] G. Boulesteix, A.L. & Tutz. Identification of interaction patterns and classification with applications to microarray data. Computational Statistics & Data Analysis, 50:783–802, 2006.

[2] J.P. Lage R. M. Assunção and E.A. Reis. Análise de conglomerados espaciais via Árvore geradora mínima. Revista Brasileira de Estatística, 63:7–24, 2002.

[3] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[4] Leo Breiman and Adele Cutler. Random forests, 2013.

[5] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[6] Y. Pirooznia, M. & Deng. Svm classifier - a comprehensive java interface for support vector machine classification of microarray data. BMC Bioinformatics, 7, Suppl 4:S25, 2006.

6 Agradecimentos

Agradecemos à FAPEMIG pelo apoio financeiro para a execução deste trabalho.