8º Congresso de extensão universitária da UNESP

"Diálogos da Extensão: do saber acadêmico à prática social"

Realização: unesp UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO" PROEX

Comparação de técnicas de aprendizado de máquina para a classificação da qualidade da madeira

Mateus Roder[1], André Luis Debiaso Rossi, Carlos de Oliveira Affonso
Unesp, Itapeva, Engenharia de Produção
E-mail: mateusroder@hotmail.com
Bolsa: Bolsista de extensão[1]

Eixo 3 - "Novas Tecnologias: Perspectivas e desafios"

Resumo

A classificação da qualidade da madeira é uma tarefa crucial para que as empresas fornecedoras dessa matéria prima logrem êxito no seu desenvolvimento econômico. Essa classificação é realizada, geralmente, por especialistas, o que torna essa tarefa subjetiva e de alto custo. Assim, buscou-se na computação, em específico na área de Aprendizado de Máquina (AM), soluções automatizadas com altas taxas de acerto e maior eficiência. Atualmente, diversas técnicas de AM estão disponíveis e podem ser usadas para esse propósito. Porém, escolher o melhor algoritmo para o problema sob análise não é uma tarefa trivial. Sendo assim, o objetivo deste artigo é comparar diferentes algoritmos de AM para o problema de classificação da qualidade da madeira. Para isso, realizaram-se experimentos usando três algoritmos de AM: Redes Neurais Artificiais, Máquinas de Vetores de Suporte e Árvores de Decisão. Esses algoritmos representam diferentes vieses de classificação. Um total de 374 instâncias obtidas a partir da análise de imagens de tábuas de madeira foram utilizadas nessa avaliação. Cada instância é descrita pelas características da imagem e por um atributo alvo, que identifica a qualidade da madeira. A partir das análises dos resultados, observou-se que os desempenhos preditivos dos algoritmos foram similares, com pequena vantagem para a máquina de vetores de suporte.

Palavras Chave: *Mineração de dados, Classificação, Qualidade da madeira.*

Abstract:

The wood quality classification is a crucial task for the suppliers of this raw material to succeed in their economic development. This classification is usually performed by experts, which makes this task subjective and expensive. Thus, computing tools have been developed, in particular in machine learning (ML) area, to accomplish this task efficiently and with high accuracy rates. Currently, several ML techniques are available and can be employed for this purpose. However, choosing which is the best algorithm for the problem under analysis is not a trivial task. Therefore, the purpose of this article is to compare different ML techniques for the wood quality classification problem. For such, we carried out experiments using three ML algorithms: Artificial Neural Networks, Support Vector Machines and Decision Trees. These algorithms represent different search bias. An amount of 374 instances obtained from wood boards images were used for training and test of the models. Each instance is described by the features of an image and a target attribute, which identifies the quality of this timber. The results showed that the predictive performance of the algorithms were similar, with slightly superiority of the support vector machines.

Keywords*: Data Mining, Classification, Wood Quality Classification.*

Introdução

A produção de matéria prima de qualidade a partir de um rigoroso critério de seleção é um dos fatores determinantes para o sucesso das empresas fornecedoras. Sendo assim, há a constante necessidade de buscar meios ou ferramentas que possibilitem a oferta de matérias primas de melhor qualidade concomitante à redução de custos (temporais e monetários).

A madeira, a qual é objeto de estudo deste artigo tem grande importância econômica para a cidade e região de Itapeva/SP, onde este trabalho está sendo desenvolvido. Um dos problemas enfrentados por pequenas e médias empresas fornecedoras dessa matéria prima é a classificação da sua qualidade. A

madeira pode apresentar vários problemas que comprometem o seu uso nas mais diversas aplicações. Segundo Quoirin (2004), os defeitos da madeira podem advir naturalmente da formação da árvore ou ocasionados durante o processo de manejo e tratamento, tais como o corte, a secagem e a usinagem. Stanalker e Harris (1989) observaram que os defeitos mais comuns são nós, empenamentos, rachaduras, inclinação das fibras e fendas. Além disso, os defeitos podem ser internos ou visíveis externamente, sendo que cada qual pode ser visualizado utilizando diferentes técnicas (VILLIERE, 1966; TOMASELLI & KLITZKE, 2000; JANKOWSKY, 2002).

Na Figura 1 é mostrada a imagem de uma tábua de madeira sem nenhum defeito, ou seja, de qualidade elevada, enquanto que a imagem da tábua de madeira mostrada na Figura 2 possui um defeito conhecido como nó, sendo assim classificada como de qualidade intermediária.

A identificação de defeitos e a classificação da qualidade da madeira são de suma importância para o setor madeireiro, pois permitem estabelecer o seu destino e uso correto como matéria prima. Em geral, esse processo é realizado manualmente por especialistas e, portanto, sujeito a diversas fontes de subjetividade, como a experiência e o cansaço desses profissionais, além das recomendações do setor de qualidade da empresa.

Figura 1. Imagem de uma tábua de madeira sem defeitos (qualidade superior).

Outra questão está relacionada ao custo e à eficiência dos especialistas que realizam essa tarefa. Quanto maior o número de especialistas, maior a quantidade de matéria prima analisada, porém, maior também será o custo de produção.

Tendo em vista essas dificuldades, sistemas que analisam automaticamente os defeitos presentes na madeira e a classificam de acordo com a sua qualidade foram desenvolvidos (AFFONSO, 2015; CASTELLANI; ROWLANDS, 2009; CHARNIYA; DUDUL, 2011; JORDAN; FEENEY; NESBITT; EVERTSEN, 1998; KHASHEI; HAMADANI; BIJARI, 2012). Entretanto, a automatização desse processo ainda é muito custoso para pequenas e médias empresas.

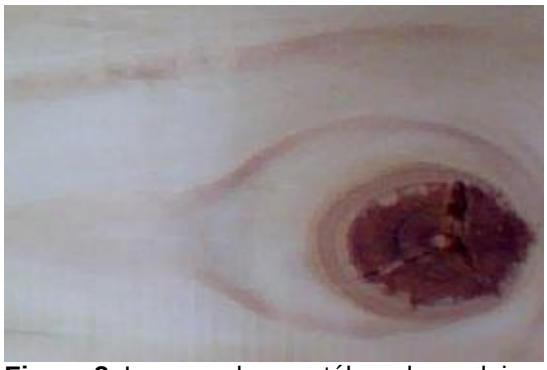

Figura 2. Imagem de uma tábua de madeira com nó (qualidade intermediária).

Muitos desses classificadores automáticos são construídos empregando-se técnicas da área de Aprendizado de Máquina (AM), que tem como objetivo principal o estudo de sistemas que podem aprender e melhorar por meio da experiência. Portanto, um sistema ou modelo de AM é capaz de tomar determinadas decisões a partir de experiências anteriores, sejam elas bem sucedidas ou não (MITCHELL, 1997). Essa experiência é fornecida por meio de uma base de dados, em que cada exemplo, ou instância, é formado por atributos inerentes ao problema e um rótulo, que identifica a classe a qual esse exemplo pertence.

Uma das técnicas de AM muito conhecida e usada para problemas de classificação são as Redes Neurais Artificiais (RNA), que mimetizam o funcionamento do cérebro humano, visto que possuem "neurônios artificiais" para o processamento de informações, buscando rapidez e eficiência no tratamento de problemas (HAYKIN, 2001).

Outra técnica de AM são as Máquinas de Vetores de Suporte (SVM, do inglês, *Support Vector Machines*). Para o problema de classificação, as SVM procuram encontrar um hiperplano que maximiza a margem de separação entre as classes. Os vetores de suporte utilizados pelas SVM são os exemplos que estão próximos da superfície de decisão e, por isso, são mais difíceis de serem classificados. As SVM conseguem lidar com problemas não lineares, mapeando o espaço de dimensão original dos dados para outro de maior

dimensão, denominado espaço de características (HAYKIN, 2001).

Uma das limitações das RNAs e das SVMs é em relação à interpretabilidade. As duas técnicas geram modelos que não podem ser trivialmente interpretados, dificultando uma maior compreensão do problema e de como os modelos realizarão a classificação de novos dados.

Por outro lado, técnicas de AM baseadas em Árvores de Decisão (AD) induzem modelos que, geralmente, podem ser facilmente interpretados. O processo de geração envolve, primeiramente, a criação de um ranking dos atributos que descrevem o problema. Aqueles com maior importância são utilizados no topo da árvore, ou seja, a classificação dos dados inicia-se observando esses atributos para depois passar pelos demais. Os modelos resultantes de AD podem ser representados por conjuntos de regras "Se, Então", o que facilita o seu uso (MITCHELL, 1997).

Embora a interpretação seja uma característica importante, o desempenho preditivo dos modelos, ou seja, sua capacidade de generalização, pode ser o propósito mais requerido em alguns casos. Esse desempenho depende do viés de cada algoritmo e das características da base de dados usada no treinamento dos modelos (MITCHELL, 1997). Entretanto, a escolha do melhor viés para cada problema não é uma tarefa trivial.

Objetivos

Este artigo tem como objetivo geral comparar o desempenho preditivo de diferentes técnicas de AM no que tange à classificação da qualidade da madeira. Essa comparação resultará na indicação de qual viés é mais apropriado de acordo com as características do problema sob análise.

Os modelos gerados por essas técnicas poderão ser usados em equipamentos de baixo custo que realizem o processo de classificação da madeira, tornando-os acessíveis a um maior número de empresas.

Material e Métodos

Os dados usados para treinamento e classificação da qualidade da madeira incluíram 374 instâncias, que foram obtidas a partir da análise de imagens de 374 tábuas de madeira em uma empresa da cidade de Itapeva, São Paulo.

Cada instância é composta dos valores dos atributos preditivos e de um atributo alvo (*target*). Os valores dos atributos preditivos foram obtidos por meio da extração de característica de imagens das tábuas de madeira usando as medidas propostas por Haralick (1973). No total, 10 atributos numéricos extraídos com essas medidas foram usados para descrever essas imagens. O atributo alvo identifica a qualidade das tábuas de madeira, de acordo com um padrão proposto por especialistas, em que "A" define um elevado padrão qualidade, "B" uma qualidade intermediária e "C" uma qualidade inferior. Do total de 374 instâncias (imagens) analisadas, 144 são da classe "A", 177 são da classe "B" e 53 da classe C. Portanto, pode-se dizer que há um desbalanceamento das classes, pois não há o mesmo número de instância para cada uma delas, principalmente da classe "C" em relação às demais.

Para a classificação das amostras de madeira serão investigadas três técnicas de AM: RNA, SVM e AD. Essas técnicas estão implementadas em diferentes linguagens de programação e podem ser encontradas em diferentes bibliotecas ou ferramentas. Para este estudo, será utilizada a linguagem de programação estatística R (R CORE TEAM, 2014).

A rede neural utilizada neste estudo é a rede perceptron de múltiplas camadas (MLP, do inglês, *MultilayerPerceptron*). Essa rede generaliza o modelo perceptron com uma única camada de neurônios e o treinamento do modelo é realizado, geralmente, pelo algoritmo de retropropagação de erro *back-propagation* (BP) (HAYKIN, 2001). Esse algoritmo será utilizado a partir do pacote "nnet", disponível para a linguagem R. O número neurônios na camada oculta foi definido como a metade do número de atributos. Os demais parâmetros foram mantidos como definidos por padrão pelo pacote que implementa o algoritmo.

As SVM serão utilizadas aqui com o kernel Gaussiano (RBF, do inglês, *Radial-Basis Function*). Esse kernel obteve bons resultados em muitos estudos e é capaz de tratar fronteiras de decisão não lineares possuindo menores dificuldades numéricas quando comparado a outros kernels (HSU, 2007). Esse algoritmo é implementado em R pelo pacote "e1071". Os valores padrão dos parâmetros definidos pelo pacote não foram alterados.

Por último, a árvore de decisão investigada neste estudo foi a CART (*Classification and Regression Trees*) (BREIMAN, 1984). A CART é uma técnica não paramétrica que pode selecionar, dentre um grande número de variáveis e suas interações, as mais importantes na determinação da variável de

resultado (YOHANNES& HODDINOTT, 1999). Essa árvore usa um procedimento de particionamento recursivo binário em que um atributo pode ser usado mais de uma vez. Esse procedimento é realizado até a árvore atingir um tamanho máximo. Em seguida ocorre a poda da árvore, que pode eliminar um ou mais particionamentos realizados anteriormente. O pacote "rpart" para a linguagem R, que implementa esse algoritmo, foi utilizado neste estudo. Os valores padrão dos parâmetros não foram modificados.

Para que a comparação entre diferentes algoritmos seja justa e as estimativas de desempenho as mais próximas possíveis do real, o método de reamostragem de validação cruzada (VC) estratificado foi utilizado. Nesse método, os dados são divididos em *k* partições disjuntas. Para cada uma das k iterações desse procedimento, k-1 partições são usadas para o treinamento do modelo, enquanto que a partição restante é separada para teste, ou seja, para estimar o desempenho preditivo do modelo gerado.

O desempenho preditivo dos modelos gerados será obtido pela média sobre as k partições de teste usando as medidas de taxa de acerto (acurácia) e a Área sob a curva (AUC, do inglês, *Area Under Curve*). A primeira é muito utilizada em estudos de AM e é calculada como o número de acertos sobre o total de exemplos preditos. Porém, essa medida não é adequada quando há um desbalanceamento de classes. Por outro lado, a AUC não é influenciada por esse fator (FAWCETT, 2006). Portanto, como a base de dados investigada é desbalanceada, a medida AUC também deve ser considerada na análise do desempenho preditivo das técnicas de AM.

Além disso, o teste estatístico de Wilcoxon será aplicado para avaliar se a diferença de desempenho dos classificadores é significativa. O teste de Wilcoxon é não paramétrico, ou seja, não faz nenhuma suposição sobre a distribuição dos dados. Neste artigo, o teste pareado é aplicado, considerando cada partição de teste do método de VC (SEWARD & DOANE, 2014).

Resultados e Discussão

Na Tabela 1 são apresentados os desempenhos preditivos dos três algoritmos investigados nesse artigo (linhas) para as medidas de taxa de acerto e AUC (colunas). O desvio padrão também é apresentado e refere-se à variação para as 10 partições testadas com o método de amostragem de VC. Além disso, para facilitar a visualização, esses resultados são apresentados por meio de um gráfico de barras mostrado no Anexo 1 (Figura 3).

Tabela 1. Médias das taxas de acerto das predições para as técnicas estudadas.

Técnicas de AM	Taxa de acerto	AUC
SVM	0,836 ± 0,046	0,763 ± 0,073
MLP	0,820 ± 0,058	0,724 ± 0,064
CART	0,799 ± 0,068	0,757 ± 0,060

Analisando a Tabela 1, nota-se que todos os algoritmos conseguiram taxas de acerto próximas ou maiores que 0,8 e AUC de aproximadamente 0,75, demonstrando que foram capazes de extrair algum conhecimento dos dados, gerando modelos com grande capacidade de generalização.

O algoritmo SVM obteve os melhores resultados de predição em relação às outras técnicas de AM para as duas medidas utilizadas. Porém, o teste de Wilcoxon mostrou que os desempenhos dos algoritmos não foram significativamente diferentes para as duas medidas utilizadas considerando um nível de confiança de 95%.

Outro ponto interessante é que a pequena quantidade de atributos extraídos das imagens foi capaz de fornecer informações úteis para a classificação da qualidade da madeira. É importante ressaltar também que a árvore de decisão possui a vantagem de ser mais fácil de interpretar em relação aos demais modelos gerados.

Conclusões

A partir das predições dos diferentes algoritmos e posteriores análises dos resultados obtidos, observou-se, de maneira sutil, um melhor desempenho preditivo do algoritmo SVM, seguido das AD e então pela MLP. Tal fato sugere que para problemas semelhantes, ou seja, do mesmo domínio, o algoritmo SVM deve ser a primeira escolha em relação ao desempenho preditivo.

Porém, ao aplicar o teste de Wilcoxon com um nível de confiança de 95%, o mesmo mostrou que os três algoritmos não possuem diferenças significativas para a predição deste problema, ou seja, os algoritmos possuem desempenhos similares.

Com isso, a classificação da madeira por meio das técnicas empregadas beneficia o setor produtivo e

também os consumidores, que terão produtos com melhor qualidade e de menor custo.

Como trabalho futuro pretende-se realizar um estudo aprofundado das AD geradas pelo algoritmo CART a fim de analisar quais atributos tiveram prevalência na geração desse modelos e, com isso, propor ou investigar outras características que podem ser importantes para a classificação dos dados.

Adicionalmente, pretende-se investigar a influência da otimização dos parâmetros de um ou mais algoritmos utilizados para esse problema. Entende-se otimização como o processo de encontrar valores para os parâmetros livres dos algoritmos de AM que resultem em modelos com melhor capacidade de generalização. Além disso, outros algoritmos, com diferentes vieses de busca poderão ser investigados para o problema, assim como procedimentos de pré-processamento dos dados de entrada.

Agradecimentos

Os autores agradecem à Pró-reitoria de Extensão (PROEX) pela bolsa concedida para a realização deste trabalho.

AFFONSO, C., SASSI, R.J., BARREIROS, R.M. **Biological image classification using rough-fuzzy artificial neural network**, Expert Systems with Applications, Available online 7 August 2015, ISSN 0957-4174.

BRAGA, A., CARVALHO, A. C. P. L. F., LUDEMIR, T.. **Redes Neurais Artificiais**: Teoria e Aplicações. Editora LTC, 2000.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J. **Classification and Regression Trees**, Wadsworth, Belmont, CA. Republished by CRC Press, 1984.

CASTELLANI M., ROWLANDS H. **Evolutionary Artificial Neural Network Design and Training for wood veneer classification**, Engineering Applications of Artificial Intelligence, Volume 22, Issues 4–5, Pages 732-741, ISSN 0952-1976, June, 2009.

CHARNIYA, N. N., DUDUL, S. V. **Classification of material type and its surface properties using digital signal processing techniques and neural networks**, Applied Soft Computing, Volume 11, Issue 1, Pages 1108-1116, ISSN 1568-4946, January, 2011.

FAWCETT, T. An introduction to roc analysis. **Pattern Recognition Letters**, v. 27, n. 8,p. 861–874, ROC Analysis in Pattern Recognition, 2006.

HAYKIN, S. **Redes Neurais - Princípios e Prática,** 2ª Ed., Porto Alegre: Bookman, 2001.

HSU, C.W., CHANG, C. C., LIN, C.J. **A Practical Guide to Support Vector Classification**, Department of Computer Science - National Taiwan University, Taipei, Taiwan, 2007.

JANKOWSKY, I. P. **Defeitos na Secagem de Madeiras.** 2002. Disponível em: <http://www.ipef.br/tecprodutos/defsecagem.asp>. Acesso em: 10 jul. 2015.

JORDAN R., FEENEY F., NESBITT N., EVERTEN J. A. **Classification of wood species by neural network analysis of ultrasonic signals**, Ultrasonics, Volume 36, Issues 1–5, Pages 219-222, ISSN 0041-624X, February, 1998.

HARALICK, R.M., SHANMUGAM,K., DISTEIN, I. **Textural Features for Image Classification**, IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, Vol. SMC-3, n. 6, p. 610-621, Institute of Electrical and Electronics Engineers, Inc., November 1973.

KHASHEI, M., HAMADANI,A. Z., BIJARI,M.**A novel hybrid classification model of artificial neural networks and multiple linear regression models**, Expert Systems with Applications, Volume 39, Issue 3, Pages 2606-2620, ISSN 0957-4174, 15 February, 2012.

MITCHELL, T. M. **Machine Learning**, McGraw-Hill, 1997.

QUOIRIN, N. S. R. **Diagnóstico de defeitos em madeira por tomografia de raios X.** 2004. 130 f. Dissertação (Mestrado) - Curso de Engenharia de Materiais e Processos, Universidade Federal do Paraná, Curitiba, 2004.

R Core Team (2014).**R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL: http://www.R-project.org/.

SEWARD, L. E., DOANE, D. P. **Estatística Aplicada à Administração e Economia,** 4. ed. Porto Alegre: Bookman, p. 840, 2014.

STANALKER, J. J., HARRIS, E. C. **Structural design in wood**. New York: Chapman and Hall, p. 426, 1989.

TOMASELLI, I.; KLITZKE, R. A.**Secagem da madeira**. Apostila do curso de Engenharia Florestal - Universidade Federal do Paraná, 2000.

VILLIERE, A. **Séchagedes bois**, 6.ed. Paris: Editora Dunod, 1966.

YOHANNES, Y., HODDINOTT, J. **Classification and regression trees: an introduction.** Washington: International Food Policy Research Institute, 1999. 29 p. Disponível em: <http://pdf.usaid.gov/pdf_docs/Pnach725.pdf>. Acesso em: 21 jul. 2015.

Anexo 1

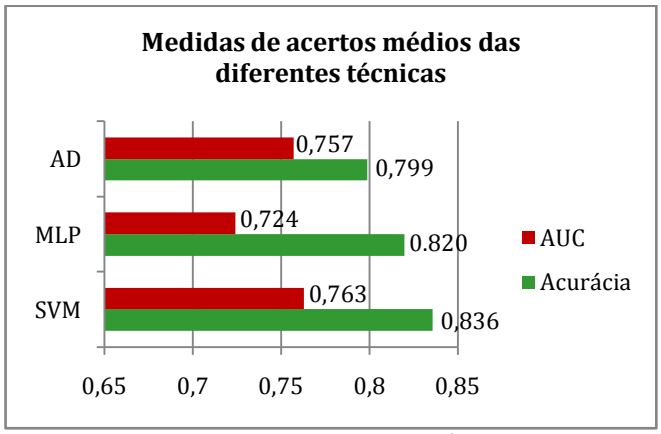

Figura3. Taxa de acerto e AUC das técnicas investigada.