7CCENDIPET06

ATRIBUIÇÃO DE AUTORIA COM WEKA

Tatiane Cruz de Souza Honório[1]; Francisco Dantas Nobre Neto[1]; Thallys Pereira de Almeida[1]; Rodrigo Cartaxo Marques Duarte[2]; Yuri de Almeida Malheiros Barbosa[2]; Vinicius de Melo Rocha[2]; Leonardo Vidal Batista[3]
Departamento de Informática/PET

Abstract. *The research guideline about authorship attribution of texts from Portuguese literature is a huge universe which has a lot to explore. Many techniques have been used trying to minimize this problem, in several languages. Were chosen and compared between them two possible approach to solve this problem. This paper presents the methods decision tree and Support Vector Machine, or just SVM.*

Resumo. *A linha de pesquisa sobre atribuição de autoria de textos da literatura portuguesa é um universo no qual ainda há muito a explorar. Várias técnicas foram utilizadas tentando auxiliar na resolução desse problema, nos mais variados idiomas. Foram selecionadas e comparadas entre si duas possíveis formas de solucionar essa questão. Este trabalho apresenta os métodos de árvore de decisão e a técnica Support Vector Machine, ou apenas SVM.*

Palavras-chave: Atribuição de Autoria, Weka, Textos da Literatura Portuguesa

Introdução

Classificar um texto consiste no trabalho de atribuir um dado texto a uma classe identificada pelo classificador. Essas classes variam de autores a estilos de escolas literárias, dependendo apenas do caminho a ser trilhado pelo pesquisador.

Na área de atribuição de autoria, temos alguns casos de sucesso, como a utilização do algoritmo de compressão de dados *Prediction by Partial Match* (PPM). Outras técnicas também já foram experimentadas nesse campo, tal como redes neurais e classificadores *bayesianos.*

Em uma classificação supervisionada, são conhecidos previamente quais textos pertencem a quais classe. Cada corpus pré-classificado é usado para construir modelos ou para definir quais atributos são distintos por classe. Este processo é chamado de aprendizagem. Características singulares e modelos precisos são a chave para obter eficiência dos classificadores.

A ferramenta utilizada para realizar os testes neste trabalho, tanto com árvore de decisão C4.5 como para o Support Vector Machine (SVM) foi o Weka (Waikato Environment for Knowledge Analysis). A motivação para utilização da mesma se dá pelo fato de ser facilmente manipulada e poder incorporar alguns recursos de filtros de forma rápida, otimizando, assim, a quantidade de testes.

A implementação para a árvore de decisão disponível no Weka é o algoritmo J4.8. Já para a poderosa técnica SVM, o algoritmo SMO(Sequential Minimal Optimization) entra em ação para prover a implementação usando polinômios ou núcleos Gaussianos.

Descrição Metodológica

O Weka é um software composto por uma coleção de algoritmos de aprendizagem de máquina e de ferramentas de processamento de dados para mineração de dados, como árvore de decisão e SVM.

O método usado pela árvore de decisão para solucionar o problema de aprendizagem é o *dividir-para-conquistar*. Nessa abordagem, os nós da árvore são responsáveis por testar os atributos. As folhas da árvore são as classes. Para realizar a classificação de uma instância desconhecida, a árvore é abordada de cima para baixo, de acordo com os atributos que são testados nos nós.

A árvore de decisão tem diversas vantagens quando comparada com outras técnicas. Algumas delas é que, ela é de fácil entendimento e de interpretação do resultado, e é rápido o processamento para atribuição dos textos. O C4.5, algoritmo para árvore de decisão, contém varias melhorias, como escolher atributos apropriados.

[1] Aluno(a) Bolsista(a); [2] Aluno(a) Voluntário; [3]Prof(a) Orientador(a)/Coordenador(a);

A técnica SVM pertence a uma categoria de classificadores lineares e foi desenvolvida por Vapnik, com o objetivo de auxiliar na solução de problemas, não só de classificação, mas também no reconhecimento de padrões. O seu conceito é baseado na idéia de minimizar riscos estruturais, ou seja, minimizar erros da classificação empírica e maximizar a margem geométrica entre os resultados.

SMO (Sequential Minimal Optimization) é um algoritmo eficiente para implementação da técnica SVM. Com a utilização desse algoritmo, a utilização de memória é linear para realizar os treinamentos. Com isso, o SMO permite lidar com grande quantidade de arquivos para treinamento.

O arquivo .arff, padrão suportado pelo Weka, foi formado após o pré-processamento dos textos dados. Foram selecionados, inicialmente, dezessete atributos numéricos para realização dos testes [7]. O arquivo foi gerado, então, contendo os dezessete atributos e as obras usadas tanto na fase de treinamento quanto na fase de classificação dos textos.

A tabela 1 mostra os atributos selecionados.

Tabela 1: Atributos selecionados

Atributos	DESCRIçÃO
Word Length Average	Tamanho médio das palavras
Short Words	Número de palavras pequenas (número de caracteres ≤3) / N
Vocabulary Richness	V / N
Hapax Legomena N	Número de hapax legomenas/N
Hapax Legomena V	Número de hapax legomenas/V
Hapax Dislegomena V	Número de hapax dislegomenas / V
Entropy Words	Entropia de palavras
Entropy 2-Grams	Entropia de 2-gramas
Entropy 3-Grams	Entropia de 3-gramas
Entropy 4-Grams	Entropia de 4-gramas
Guirad R	$R = V / \sqrt{N}$
Herdan C	$C = \log10\ V / \log10\ N$
Rubert K	$K = \log10\ V / \log10\ (\log10\ N)$
Maas A	$A = \sqrt{\log10\ N - \log10\ V / (\log10\ N)^2}$
Dugast U	$U = (\log10\ N)^2 / \log10\ N - \log10\ V$
Luk Janenkov	$1 – (V^2 / V^2 * \log10\ N)$
Honore H	H = 100 * log10 N / (1-(número de hapax legomenas / V))

Onde:
N = número total de palavras;
V = número total de palavras distintas.

Dez classes de autores foram formadas cada uma contendo quatro obras limitadas em até 150000 caracteres. Para normalizar os atributos, os textos foram divididos em blocos com 1000 palavras e processados à geração dos valores dos atributos. Caso o último bloco não obtivesse as 1000 palavras, ele seria descartado.

Os testes foram feitos utilizando validação cruzada, e em cada rodada são classificados exclusivamente textos que não participaram da etapa de aprendizagem. A classificação consiste em atribuir cada texto a uma das classes através da análise dos valores dos atributos deles colhidos.

Os textos escolhidos para realizar os testes de classificação foram:

Tabela 2: Textos escolhidos por autor

Autor	Textos
Adolfo Caminha	A Normalista; Bom-Crioulo; No País do Ianques; Tentação;
Aluisio de Azevedo	A Mortalha Alzira; Casa de Pensão; O Cortiço; O Mulato;
Euclides da Cunha	Á Margem da História; Contrastes e Confrontes; Os Sertões; Peru Versus Bolívia;
Joaquim Manuel Macedo	A Luneta Mágica; A Moreninha; Memórias da Rua do Ouvidor; O Moço Louro;
José de Alencar (Romance Regional)	O Garatuja; O Gaúcho; O Sertanejo; Til;
José de Alencar (Romance Urbano)	Diva; Encarnação; Lucíola; Senhora;
Lima Barreto	Histórias e Sonhos; Os Bruzundangas; Recordações do Escrivão Isaías Caminha; Triste Fim de Policarpo Quaresma
Machado de Assis (Realismo)	Dom Casmurro; Memorial de Aires; Memórias Póstumas Brás Cubas; Quincas Borba;
Machado de Assis (Romance)	Helena; Iaiá Garcia; A Mão e a Luva; Ressurreição;
Visconde de Taunay	A Retirada da Laguna; Ao Entardecer; Inocência; No Declínio;

A divisão dos autores Machado de Assis e José de Alencar se dá devido a seus textos possuírem mais de um estilo literário bem característico. A escolha desses autores se deu pela facilidade de encontro de obras dos mesmos em formato eletrônico e em domínio público, e também porque os autores listados acima são consagrados em nosso país.

Resultados

O método usado para selecionar os atributos, em ambas as abordagens de classificação, foi simples. Inicialmente, apenas um único atributo foi escolhido para participar dos testes em cada rodada. Depois, todos os atributos foram marcados e excluído apenas um, por rodada. Logo após, os atributos que obtiveram melhor eficiência foram escolhidos para o teste final. Foram selecionados entre os três e sete melhores atributos, para cada método.

Com o auxílio do filtro Discretize em ambos os classificadores, os resultados seguintes mostram a porcentagem de acertos.

A tabela 3 mostra a matriz de confusão para o melhor resultado da classificação quando utilizado árvore de decisão C4.5:

Tabela 3: Matriz de confusão com C4.5

a	b	c	d	e	f	g	h	i	j	Classificado como:
4	0	0	0	0	0	0	0	0	0	a= JA(RR)
0	2	0	2	0	0	0	0	0	0	B= JMM
0	0	3	1	0	0	0	0	0	0	c= AA
0	0	0	4	0	0	0	0	0	0	d= MA(RE)
0	0	0	0	4	0	0	0	0	0	e= MA(RO)
0	0	0	0	0	3	0	0	1	0	f= VT

0	0	0	0	0	0	4	0	0	0	g= LB
1	1	1	0	0	0	1	0	0	0	h= JA(RU)
0	0	0	0	0	0	0	0	4	0	i= EC
0	1	0	1	0	0	0	0	0	2	j= AC

Os atributos que foram usados para teste foram: o tamanho médio das palavras, Hapax Legomena N, Hapax Dislegomena V e entropia de palavras. Com eles foi obtida uma taxa de acerto de 75%, com 30 atribuições corretas de 40.

A tabela 4 mostra a matriz de confusão para o resultado da classificação quando utilizada o método SVM:

Tabela 4: Matriz de confusão com SVM

a	b	c	d	e	f	g	h	i	j	Classified como:
4	0	0	0	0	0	0	0	0	0	a= JA(RR)
0	4	0	0	0	0	0	0	0	0	b= JMM
0	0	3	0	1	0	0	0	0	0	c= AA
0	1	1	2	0	0	0	0	0	0	d= MA(RE)
0	0	1	0	3	0	0	0	0	0	e= MA(RO)
0	0	0	0	0	4	0	0	0	0	f= VT
0	0	0	0	0	0	4	0	0	0	g= LB
0	0	0	0	0	0	0	4	0	0	H= JA(RU)
0	0	0	0	0	0	0	0	4	0	i= EC
0	0	0	0	0	1	1	0	0	2	I= AC

O melhor resultado ao utilizar SVM foi encontrado com os seguintes atributos: tamanho médio das palavras, Hapax Legomena N, Hapax Dislegomena N, Hapax Dislegomena V, entropia de bigramas, entropia de trigramas, entropia de quadrigramas, Guirad R e Herdan C. Estes atributos alcançaram 85% de taxa de acerto.

Conclusão

O processo de atribuição de autoria não é simples. Autores apresentam variações de estilos que os levam a serem confundidos com outros escritores. A seleção cuidadosa de atributos é uma etapa fundamental para o desempenho do classificador.

Autores como José de Alencar (período urbano), Lima Barreto e Euclides da Cunha, possuem um estilo literário singular, o que levou a uma taxa de acerto de 100% em ambos os métodos.

Com relação à complexidade computacional, houve vantagem por parte do J4.8. A construção da árvore e a classificação obtiveram cerca de 90% menos do tempo exigido pelo SMO. A taxa de acerto com a utilização da árvore de decisão alcançou 75%, com baixo custo computacional, enquanto que o SVM obteve 85% de acerto, mas requerendo um maior custo computacional. Portanto, a escolha do classificador depende do grau de exigência e dos recursos computacionais disponíveis.

Os resultados obtidos são competitivos em relação a alguns dos melhores classificadores reportados na literatura especializada, que não utilizam seleção de atributos, que atingem em torno de 86% de acerto. A vantagem das técnicas aqui apresentadas é que os atributos extraídos podem lançar luz sobre o que caracteriza o estilo literário dos escritores, o que pode ser valioso para o profissional de Letras.

Referências

[1]KJELL, Bradley. *Autorship Attribution of Text Samples using Neural Networks and Bayesian Classifiers*. IEEE: International Conference on Systems Man and Cibernetic, San Antonio, USA; pp: 1660-1664; ISBN 0-7803-2129-4;1994.

[2]SHANNON, C. E. "*A Mathematical Theory of Communication*" Bell Syst. Tech. J., vol. 27, pp. 379-423, (1948).

[3]BATISTA, L. V., MEIRA, Moab Mariz; *Texture Classification Using the Lempel-Ziv-Welch Algorithm*. Lecture Notes in Computer Science. Berlin, v.3171, p.444 - 453, 2004.

[4]BATISTA, L. V., MEIRA, Moab Mariz. *Texture Classification using Histogram Equalization and the Lempel-Ziv-Welch Algorithm*. Anais do XXI Simpósio Brasileiro de Telecomunicações - SBT'2004, v.1. p.1-6, Belém, 2004.

[5] Witten, Ian H.; Frank, Eibe. Data *mining: practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2005.

[6] Pyle, Dorian; *Data preparation for data mining*. San Diego: Morgan Kaufmann, 1999.

[7] Corney, M.; *Analysing E-mail Text Authorship for Forensic Purposes*. Brisbane: QUT, 2003. 165 p. Dissertação (Mestrado) - Queensland University of Technology, Brisbane, 1983.