Detectando similaridade entre diferentes representações de entidades usando Classificadores Bayesianos[1]

Daniel S. de Oliveira, Carina F. Dorneles

Dpto. de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.049-900 – Florianópolis – SC – Brazil

{oliveirads,dorneles}@inf.ufsc.br

Resumo. *Este artigo apresenta os resultados iniciais de um experimento realizado sobre o Apache Mahout com o objetivo de detectar similaridade entre bases relacionais, utilizando Classificadores Bayesianos. Foram realizados testes com os algoritmos disponíveis – Bayes e CBayes – para classificar os registros de duas bases de dados com conteúdo semelhante.*

1. Introdução

Na Web, as entidades do mundo real podem estar representadas de formas diferentes, tornando a integração de dados uma tarefa árdua. Muitas vezes se deseja encontrar uma entidade e o ideal seria encontrar as representações mais significativas, de modo a enriquecer a consulta. Dada uma representação de uma entidade, por exemplo, "Universidade Federal de Santa Catarina", afirmar que outra representação, por exemplo, "UFSC", diz respeito à mesma entidade pode ser um problema de classificação binária: ou a segunda representação diz respeito à primeira, ou não. Dentre as técnicas de classificação binária, destacam-se os Classificadores Bayesianos, por sua simplicidade e eficiência. [Rennie ET AL 2003] sugerem correções que tornam o Classificador Textual Naive Bayes praticamente tão eficiente quanto o Support Vector Machine (SVM), um classificador muito usado pela confiabilidade nos resultados oferecidos [H and Thor 2010, Bilenko and Mooney 2003, Christen 2008].

Neste trabalho, são analisados experimentos de classificação de representações de entidades usando os Classificadores Bayesianos do Apache Mahout. Os algoritmos são treinados com entidades de uma, ou mais classes. Em seguida, são inseridas entidades diversas para serem classificadas. Por fim, a ferramenta aponta o percentual de classificações corretas. Dessa forma, pode-se afirmar quais são as entidades mais similares entre si. A Seção 2 deste artigo fala sobre os Classificadores Bayesianos e a implementação destes pelo Apache Mahout. A Seção 3 mostra os experimentos realizados. Os trabalhos futuros são apresentados na Seção 4.

2. Apache Mahout – Classificadores Bayesianos

Os Classificadores Bayesianos são capazes de determinar a classe de uma entidade, uma vez fornecidos dados de treino com entidades rotuladas. Suas limitações compreendem basicamente a necessidade de quantidade semelhante de registros nas classes de treino e

[1] Este trabalho é parcialmente financiado pelo CNPQ (Bolsa PQ-Nível 2 - Nro. processo: 307992/2010-1) e pela Capes (Bolsa do Programa REUNI).

a representatividade destes. O Mahout possui dois Classificadores Bayesianos: Naive Bayes (Bayes) e Complementary Naive Bayes (CBayes). O primeiro já possui algumas correções que levam em conta a quantidade de palavras que aparecem em um documento, representando uma classe e o somatório de ocorrências da classe ao longo do corpo de documentos. O segundo corrige erros sistêmicos do primeiro, que causam classificações errôneas, como dados de treino desbalanceados e classes com maior peso do que outras, causado pela assunção de independência entre as características. Além de modelar melhor o texto, aplicando técnicas de processamento de texto, como a métrica de similaridade Tf-Idf. Ambos possuem um parâmetro para ajuste fino manual dos pesos, *alpha_i*, conhecido como Suavizador de Laplace [Mitchell 2010], que serve para evitar superajustes que invalidam o modelo.

O modelo é gerado com base nos dados de treino, onde são apresentadas classes contendo entidades rotuladas. Depois são inseridos os documentos contendo representações de entidades e o Classificador, baseado no modelo gerado, aloca as representações nas classes com maior verossimilhança. A ferramenta dispõe de um Testador que verifica se as classificações foram realizadas corretamente. Como os dados de teste também são rotulados, é possível checar se uma representação foi alocada na classe de mesmo rótulo. Observa-se que este rótulo não é utilizado durante uma classificação normal, pois esta ocorre apenas com o processamento do Classificador sobre o modelo. A Figura 1 mostra a arquitetura do Classificador Bayesiano do Mahout.

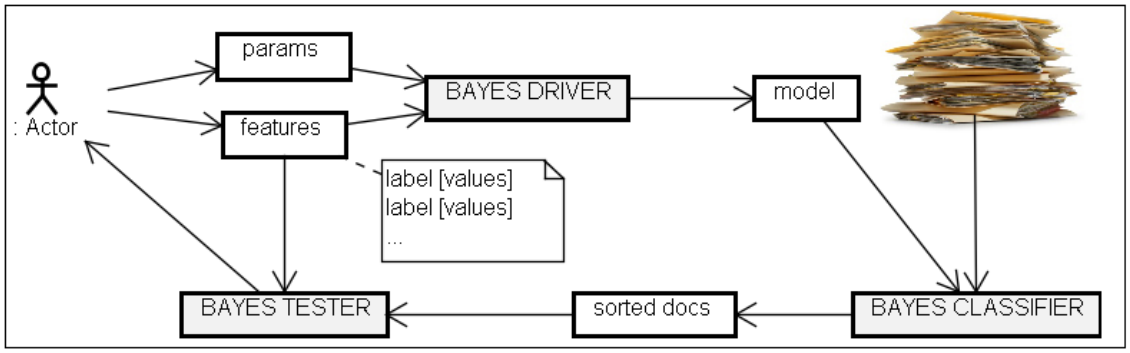

Figura 1. Arquitetura do Classificador Bayesiano do Apache Mahout

3. Experimentos

O objetivo dos experimentos é comprovar a eficácia dos Classificadores Bayesianos do Mahout quando aplicados sobre dados tabelados, a fim de indicar similaridade entre valores de atributos de representações de entidades simples – sem relacionamentos.

3.1. Configuração

Os experimentos foram realizados utilizando uma base de dados de ruas, contendo aproximadamente 3500 registros; e uma base de dados de inscrição em concurso, contendo aproximadamente 6500 registros: RUAS(ender), CONCURSO(chave, dtnasc, sexo, unidade, ender, comple, bairro, cidade, uf). A ferramenta foi treinada com metade (3250) dos registros da base de concursos, considerando três fatores: algoritmo (Bayes e CBayes), parâmetro de sensibilidade dos pesos – *alpha_i* – e conjunto de atributos (Completo e Seletivo). Este terceiro fator é empregado suprimindo atributos de códigos

julgados pouco significativos para representar a entidade real, neste caso “chave” e “unidade”, seguindo orientações do manual [Owen ET AL 2011]. O Testador foi utilizado após cada treino para checar a eficácia da classificação executada sobre a outra metade da base de concursos (3250 registros) e sobre a base de ruas (3500 registros).

3.2. Resultados

A Figura 2 mostra os gráficos dos resultados dos testes realizados com treinamento completo. No primeiro conjunto de testes, são classificadas 3250 representações da entidade Concurso, utilizando os dois algoritmos para comparar todos os atributos desta base com todos os da base de treino. No segundo conjunto de testes, são classificadas 3500 representações da entidade Rua, também utilizando os dois algoritmos para comparar todos os atributos das duas bases. A Figura 3 mostra os gráficos dos resultados dos testes realizados com treinamento seletivo, onde foram excluídos os atributos de códigos. O teste do terceiro conjunto é similar ao primeiro, assim como o quarto conjunto com relação ao segundo. O percentual de classificações corretas se refere aos atributos que foram alocados na classe esperada. No caso de Concurso sobre Concurso, cada atributo deve corresponder a si próprio. No caso de Rua sobre Concurso, o atributo “Rua.ender” deve corresponder ao atributo “Concurso .ender”.

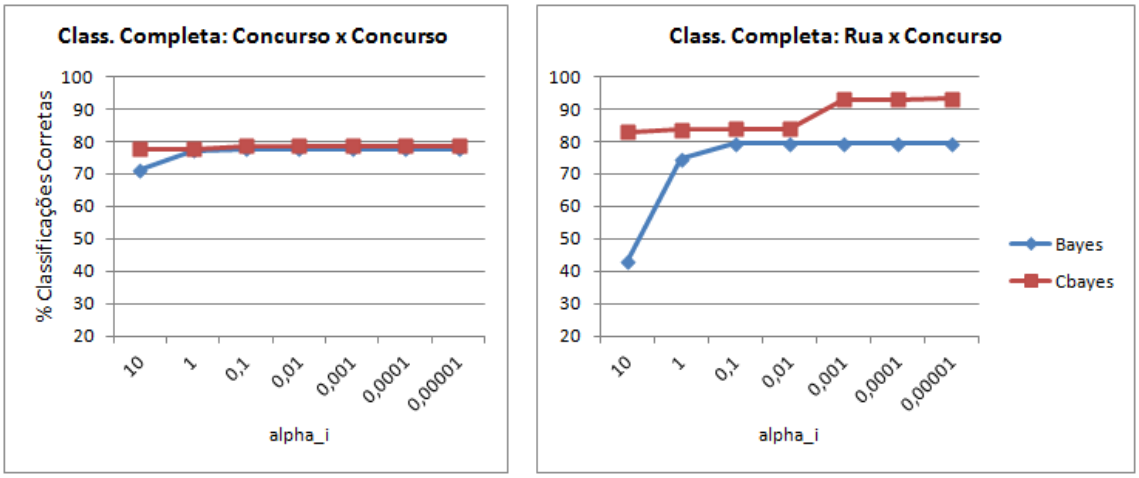

Figura 2. Classificação com todos os atributos

Figura 3. Classificação apenas com atributos representativos

Percebe-se que, em geral, o percentual de acerto é maior na classificação seletiva, pois os algoritmos não são confundidos pelos atributos de código. Percebe-se também que o algoritmo CBayes tem rendimento diferenciado nos dois casos em que as bases de treino e teste são diferentes, pois, nos outros dois casos, é mais fácil afirmar que a entidade testada é similar às entidades de treino, então o Bayes também apresenta resultado satisfatório. Ainda, percebe-se que o percentual de acerto aumenta à medida que o alpha_i diminui, pois ele não interfere tanto no processamento dos algoritmos. As classificações mal sucedidas se devem, basicamente, à semelhança entre valores de atributos distintos, como por exemplo, endereços completos que contêm o nome do bairro. Importante salientar que os resultados positivos na comparação entre as bases Rua e Concurso não indicam similaridade entre elas, mas sim entre os atributos comuns a ambas. Com base nessa capacidade de reconhecer atributos similares, pode-se proceder à indicação de entidades similares.

4. Conclusões e Trabalhos Futuros

Após estudo teórico e análise dos experimentos, conclui-se que os Classificadores Bayesianos do Apache Mahout são eficazes para detectar similaridade entre diferentes representações de entidades, tomando como base a similaridade entre seus atributos. Para aperfeiçoar os resultados gerados pelo Mahout, os algoritmos de Classificação Bayesiana serão adaptados, fazendo uso de métricas de similaridade já conhecidas no campo de recuperação de texto [Dorneles ET AL 2011], de modo a indicar a melhor função, ou conjunto de funções, para comparar cada atributo. Ainda, pretende-se automatizar a tarefa de testes, a fim de descobrir as melhores combinações de algoritmo e parâmetros de sensibilidade para uma determinada base de dados.

Referências

Bilenko, M. and Mooney, R. J. (2003). Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Christen, P. (2008). Febrl: a freely available record linkage system with a graphical user interface. In *Proceedings of the second Australasian workshop on Health data and knowledge management-Volume 80*. Australian Computer Society, Inc.

Dorneles, C. F., Gonçalvez, R., Mello, R. S. (2011). *Approximate data instance matching: a survey*. Knowledge and Information Systems, v. 27, i. 1.

H and Thor, A. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, v. 3, n. 1, p. 484-493.

Mitchell, T. M. (2010). Chapter 1 Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression Learning Classifiers based on Bayes Rule. *Machine Learning*. p. 1-17.

Owen, S., Anil, R., Dunning, T. and Friedman, E. (2011). *Mahout in Action*. Manning Publications.

Rennie, J. D., Shih, L., Teevan, J. and Karger, D. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Machine Learning-International Workshop Then Conference*.