

CLASSIFICAÇÃO MULTIRRÓTULO DE DOCUMENTOS TEXTO UTILIZANDO A RELEVÂNCIA BINÁRIA E O ALGORITMO NAÏVE BAYES

José Maurício Karl Ururahy¹, José Lucas de Godoy Viana Bastos²,
Alexandre Cesar Maretto Federici³ e Eduardo Corrêa Gonçalves⁴

Introdução

Classificação multirrótulo (CMR) ou *multi-label classification* é um dos tópicos de pesquisa mais relevantes nas áreas de Big Data e Aprendizado de Máquina. Neste problema, o objetivo é realizar a associação automática de objetos a uma ou mais classes, pertencentes a um conjunto pré-definido de classes. Existem muitas aplicações modernas e importantes para a CMR, tais como a genômica funcional (determinar as funções biológicas de genes e proteínas) e a categorização de textos (associar documentos texto a tópicos). Este artigo descreve as etapas de um experimento de CMR no ambiente R utilizando uma base de dados que armazena informações sobre mais de 264.000 filmes.

Objetivos

Este trabalho descreve um experimento realizado sobre uma base de dados real: a *Internet Movie Database* [IMDb, 2018]. A partir desta base, foi construído um sistema de classificação multirrótulo [GIBAJA; VENTURA, 2015] capaz de inferir os gêneros de um filme (“drama”, “comédia”, “ação”, etc.) a partir do texto contendo o seu resumo.

Material e Método

A base de dados do IMDb é fornecida em vários arquivos texto. O experimento envolveu os arquivos “plot.list” (resumos dos filmes em inglês) e “genres.list” (gêneros dos filmes) que, em conjunto, armazenam informações sobre 264.301 filmes. A construção do

¹ Escola Nacional de Ciências Estatísticas (ENCE-IBGE), zemaauricioku96@gmail.com

² Escola Nacional de Ciências Estatísticas (ENCE-IBGE), maher.dae@gmail.com

³ Escola Nacional de Ciências Estatísticas (ENCE-IBGE), alexandre.federici@gmail.com

⁴ Escola Nacional de Ciências Estatísticas (ENCE-IBGE), eduardo.correa@ibge.gov.br

classificador contemplou duas etapas: modelagem da base de dados e criação do modelo classificador. Na primeira etapa, o pacote “tm” [FEIRENER; HORNIK; MEYER, 2008] foi utilizado para transformar o arquivo de resumos de filmes (dado não-estruturado) em uma matriz de palavras (dado estruturado). Neste processo, inicialmente as palavras irrelevantes (*stop words*) – como artigos, pronomes e preposições – são removidas. Em seguida, os resumos são modelados como uma matriz binária X , de dimensão $d \times n$ (*document-term matrix*), onde d é o número de filmes e n a quantidade de termos (palavras que não são *stop words*). O valor 1 em uma célula $C_{i,j}$ de X indica a presença do termo t_j no resumo do filme d_i , enquanto 0 indica que t_j não ocorre em d_i . O arquivo de gêneros foi importado para uma segunda matriz binária Y , de dimensão $d \times q$, onde q é o número de classes.

A segunda etapa consiste na construção do classificador multirrótulo propriamente dito, que corresponde a uma função alvo $Y=f(X)$, capaz de mapear as palavras presentes em um resumo em uma lista de gêneros. A coleção de 264.301 filmes forneceu 328.881 termos distintos. Apenas os 500 termos mais frequentes foram selecionados para tomar parte na construção do classificador. Neste processo, inicialmente o pacote “mldr” [CHARTE; CHARTE, 2015] foi utilizado para dividir as matrizes X e Y em duas partições: treino (2/3 das observações) e teste (1/3 das observações). Em seguida, a construção do classificador foi realizada aplicando-se o método de Relevância Binária [GIBAJA; VENTURA, 2015] sobre a partição de treinamento. Nesta abordagem, q classificadores binários são treinados de forma independente, um para cada classe. Para classificar um novo objeto (ou seja, inferir a lista de gêneros de um novo resumo), basta combinar as saídas produzidas por cada classificador binário. Os classificadores binários foram treinados com o uso do algoritmo Naïve Bayes (pacote “naivebayes” [MAJKA, 2018]). Trata-se de um classificador probabilístico que, para cada rótulo de classe, emprega o Teorema de Bayes para gerar uma estimativa de o novo objeto pertencer à mesma [GONÇALVES, 2014].

Resultados e Discussão

A acurácia de cada classificador binário foi mensurada utilizando-se a partição de teste. Os resultados são apresentados na **Tabela 1**. É possível observar que o classificador do gênero “Documentário” obteve o melhor desempenho (76% de acurácia) e o do gênero “Aventura”, o pior (62%). A acurácia média dos classificadores foi igual a 66%. Estes resultados podem ser considerados promissores, tendo em vista que o número de termos utilizados para construir o modelo foi bastante reduzido (menos de 1% do total) e o modelo foi construído com o emprego de uma abordagem *baseline* [GONÇALVES, 2014].

Tabela 1 – Acurácias obtidas por cada classificador binário na partição de teste.

	Classe								
	Ação	Aventura	Comédia	Crime	Documentário	Drama	Horror	Romance	Suspense
Acurácia	0,67	0,62	0,65	0,66	0,76	0,65	0,63	0,65	0,64

Conclusão

Este trabalho apresentou um estudo inicial sobre CMR no ambiente R, onde a abordagem básica para a construção de um classificador multirrótulo (Relevância Binária com Naïve Bayes) foi aplicada a dados textuais contendo informações sobre filmes. Visando melhorar o desempenho preditivo do modelo de classificação, como trabalho futuro pretende-se realizar novos experimentos utilizando um número maior de termos (acima de 500) e empregando técnicas de CMR mais sofisticadas, como as recentemente disponibilizadas no pacote “utiml” [RIVOLLI, 2018].

Referências

- CHARTE, F.; CHARTE, D. Working with multi-label datasets in R: the mldr package. **The R Journal**, v. 7, n. 2, p. 149–162, 2015.
- FEIRENER, I.; HORNIK, K.; MEYER, D. Text mining infrastructure in R. **Journal of Statistical Software**, v. 25, n. 5, p. 1–54, 2008.
- GIBAJA, E.; VENTURA, S. A tutorial on multilabel learning. **ACM Computing Surveys**, v. 47, n. 3, p. 1–52, 2015.
- GONÇALVES, E. C.. NBBR: a baseline method for the evaluation of Bayesian multi-label classification algorithms. In: IEEE 14 th International Conference on Computational Science and Its Applications, 2014, Guimarães. **Anais... Guimarães: IEEE**, 2014, p. 245–247.
- IMDb. **The internet movie database**. Disponível em: <<http://www.imdb.com/interfaces>>. Acesso em: 19 mar. 2018.
- MAJKA, M. **Package ‘naivebayes’**. Disponível em: <<https://cran.r-project.org/web/packages/naivebayes/naivebayes.pdf>>. Acesso em: 13 mar. 2018.
- RIVOLLI, A. **Package ‘utiml’**. Disponível em: <<https://cran.r-project.org/web/packages/utiml/utiml.pdf>>. Acesso em: 18 mar. 2018.