Um Classificador de Renda para o Adult Data Set

Luis Felipe Müller
lhenriques@inf.puc-rio.br

Rafael dos Reis Silva
rrsilva@inf.puc-rio.br

30 de Junho de 2015

Resumo

Este trabalho tem como objetivo melhorar a acurácia de um classificador binário de renda para o *Adult Data Set*. Foram utilizados alguns modelos de aprendizado de máquina gerados a partir de diferentes algoritmos. Aqueles baseados em árvore de decisão foram os que obtiveram os melhores resultados, atingindo uma acurácia de mais de 99

1 Introdução

Classificadores binários são modelos clássicos de aprendizado de máquina. Um exemplo é a tarefa de classficar uma mensagem como *spam* ou não-*spam*. Esses problemas pertecem à categoria de aprendizado *supervisionado*, que consiste em predizer, ou estimar, uma *saída* baseada e uma ou mais *entradas* [1].

2 Base de Dados e Tarefa

O *Adult Data Set* [2] é uma base de dados do *UCI Machine Learning Repository* [3], gerada a partir de dados do censo americano de 1994. O *data set* possui 15 atributos, sendo 9 categóricos e 6 numéricos, como idade, educação, estado civil, dentre outros. São disponibilizados 2 arquivos, um para o conjunto de treino (com 32.561 registros) e o outro de teste (com 16.281).

A tarefa associada a esta base é uma classficação binária: predizer se uma pessoa possui renda maior do que \$50K ao ano. O atributo *over50K* contém essa informação: ele possui valor "$>$50K" se a pessoa apresenta renda maior do que \$50K ou "$<=$50K" caso contrário.

O *data set* foi citado pela primeira vez em [4] e, desde então, é utilizado, principalmente, para medir a qualidade de novos métodos de aprendizado de máquina [5] [6] [7].

3 Preparação dos Dados

Os arquivos com os dados foram tratados ...

4 Classificadores

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.935	0.482	0.859	0.935	0.896	0.893	<=50K
0.518	0.065	0.716	0.518	0.601	0.893	>50K

Tabela 1: Resultados por Classe: Naive Bayes

a	b	← classified as
23112	1608	a = <=50K
3782	4059	b = >50K

Tabela 2: Matrix de Confusão Naive Bayes

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.932	0.339	0.897	0.932	0.914	0.921	<=50K
0.661	0.068	0.755	0.661	0.705	0.921	>50K

Tabela 3: Resultados por Classe: NBTree

a	b	← classified as
23037	1683	a = <=50K
2658	5183	b = >50K

Tabela 4: Matrix de Confusão NBTree

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.999	0.006	0.998	0.999	0.999	1	<=50K
0.994	0.001	0.997	0.994	0.996	1	>50K

Tabela 5: Resultados por Classe: Random Tree

a	b	← classified as
24697	23	a = <=50K
46	7795	b = >50K

Tabela 6: Matrix de Confusão Random Tree

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.006	0.998	1	0.999	1	<=50K
0.994	0	1	0.994	0.997	1	>50K

Tabela 7: Resultados por Classe: Random Forest

a	b	← classified as
24717	3	a = <=50K
44	7797	b = >50K

Tabela 8: Matrix de Confusão Random Forest

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.002	0.999	1	1	1	<=50K
0.998	0	1	0.998	0.999	1	>50K

Tabela 9: Resultados por Classe: Random Committee

a	b	← classified as
24719	1	a = <=50K
18	7823	b = >50K

Tabela 10: Matrix de Confusão Random Committee

Referências

[1] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN: 1461471370, 9781461471370.

[2] Barry Becker. *Adult Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/Adult`.

[3] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: `http://archive.ics.uci.edu/ml`.

[4] Ron Kohavi. "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid". Em: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, to appear.

[5] Dmitry Pavlov, Jianchang Mao e Byron Dom. "Scaling-up support vector machines using boosting algorithm". Em: *Pattern Recognition, 2000. Proceedings. 15th International Conference on*. Vol. 2. IEEE. 2000, pp. 219–222.

[6] José Ramón Cano, Francisco Herrera e Manuel Lozano. "Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study". Em: *Evolutionary Computation, IEEE Transactions on* 7.6 (2003), pp. 561–575.

[7] Moritz Hardt, Katrina Ligett e Frank McSherry. “A simple and practical algorithm for differentially private data release”. Em: *Advances in Neural Information Processing Systems*. 2012, pp. 2339–2347.