FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2013

Machine Learning for Sentiment Analysis

Michal Patočka[1]

1 Introduction

Sentiment analysis is one of the tasks from natural language processing (NLP). The main idea behind this task is determining so called sentiment from the given text. Sentiment of text reflects opinions of the text author. This work uses automatic sentiment detection in the way described in Pang, B. and Lee, L. (2008).

Main focus of our work is improving vectors of features for machine learning algorithms by using feature selection algorithms. These algortihms can identify and filter out least relevant features and increase performance of machine learning algorithms. Five different algorithms are used and tested on dataset extracted from Czech server heureka.cz, that contains user reviews for products from e-shops.

2 Feature Selection Algorithms

Every feature selection algorithm works on similar basics. It uses probability of feature occurrence in specific class to determine feature weight. Feature weight is higher, when the feature is less evenly occurring throughout all classes (Forman et al. (2003)).

Feature selection algorithms are used to achieve main goal - reduce size of feature vector while not harming classification success rate. In ideal case we will achieve higher performance by filtering non-relevant features (Mejova, Y. and Srinivasan, P. (2011)).

3 Evaluation

We use data-specific approach to determine success rate of algorithms. Selected dataset is highly imbalanced, so metrics based on accuracy are not applicable. Instead we use F1-measure, that computes success rate for every category and uses average value as a result.

Maximum Entropy algorithm is used for all results listed below, as it provides best success rate to evaluation time ratio.

Performance is determined by using a standard cross-validation method. 90% of the dataset is used as a training set and 10% as a test set.

3.1 Results

We were able to identify best setups for the following algorithms Odds Ratio (OR), Relevancy Score (RS), Information Gain (IG) and Mutual Information (MI). The setup parameter is stated in brackets. No usable setup of algorithm Chi Square was identified.

Best results are shown on table 1. Relevancy Score was able to achieve higher F1-measure then baseline, while filtering 18.6% of all features. In all other cases F1-measure

[1] student of master program Computer Science and Engineering, study field Software Engineering, e-mail: thelama@students.zcu.cz

Algorithm	Accuracy	F1 measure	Feature count	Feature vector reduction	Computation time [s]
no selector	0,75	0,63	375470	0	4505
RS(0,1)	**0,77**	**0,65**	375470	18,6%	4172
RS(0,5)	0,61	0,46	375470	93,7%	2961
RS(0,6)	0,36	0,34	375470	98,2%	3779
MI(0,6)	0,73	0,59	375470	0,3%	4863
MI(0,7)	0,71	0,55	375470	2,6%	4695
MI(0,8)	0,68	0,50	375470	20,4%	5772
IG(0,7)	0,71	0,55	375470	2,6%	3634
IG(0,8)	0,68	0,50	375470	20,4%	4022
IG(0,9)	0,51	0,36	375470	71,5%	3847
OR(0,9)	0,74	0,56	375470	31,8%	5090

Table 1: Performance of best identified algorithm setups. All results are statistically significant on 95% level.

was lower than baseline, but good reduction of feature vector size and/or computation time (computed on Intel Core I7 (3612QM) 2,10GHz, 8GB RAM, WIN7 machine) was achieved. This is important as well, because sentiment analysis is quite computation-heavy, and reducing computation time can be very useful for some large experiments.

4 Conclusion

The evaluation demonstrates, that using feature selection algorithms in combination with machine learning algorithms can outperform the baseline classifier with all features as well as increase the computation efficiency.

Acknowledgement

The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated.

References

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1–135.

Forman, G., Guyon, I., and Elisseeff, A. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3:1289–1305.

Mejova, Y. and Srinivasan, P. (2011). Exploring feature definition and selection for sentiment classifiers. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, ICWSM. The AAAI Press