Introduction //ADD

Some prediction approaches are based on ana- lyzing early users’ comments [1], or features about post contents and domains [2]. Another proposed method [3] predicted the article’s popularity not only based on its own appeal, but also other ar- ticles that it is competing with. Prediction models with SVMs, Ranking SVMs [3], Naive Bayes [2] are investigated, and more advanced algorithms such as Random Forest, Adaptive Boosting [4] could increase the precision. This report however, incorporates a broader and more abstracter set of features, and starts with basic regression and classification models to advanced ones, with elaborations about effective feature selection.

Reference:

[1] Tatar, Alexandru, et al. "Predicting the popularity of online articles based on user comments." Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.

[2] "Predicting the Popularity of Social News Posts." 2013 cs229 projects. Joe Maguire Scott Michelson.

[3] Hensinger, Elena, Ilias Flaounas, and Nello Cristianini. "Modelling and predicting news popularity." Pattern Anal- ysis and Applications 16.4 (2013): 623-635.

[4] K. Fernandes, P. Vinagre and P. Cortez. A Proactive In- telligent Decision Support System for Predicting the Pop- ularity of Online News. *Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence*, September, Coimbra, Portugal.

[5] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: A li- brary for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011): 27.

[6] James, Gareth, et al. *An introduction to statistical learning*. New York: springer, 2013.

[7] K. Fernandes, P. Vinagre and P. Cortez. *A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News*. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

Future Scope

As is seen from the result, no algorithm can reach 70% accuracy given the data set we have, even though they are state-of-the-art. To improve accuracy, there is little room in model selection but much room in feature selection. In the pre-processing round, 59 features were extracted from news articles, and our later work is based on these features. However, the content of news articles hasn’t been fully explored. Some features are related to the content, such as LDA topics (feature #39 - #43), which are convenient to use for learning, but reflect only a small portion of information about the content.

In the future, we could directly treat all the words in an article as additional features, and then apply machine learning algorithms like Naive Bayes and SVM. In this way, what the article really talks about is taken into account, and this approach should improve the accuracy of prediction if combined with our current work.