Technology Review: Comparison of sentiment classification methods: Naive Bayes and Support Vector Machine(SVM)

Jianjia Zhang
jianjia2@illinois.edu, Master of Computer Science

Nov 10 2020

1 Introduction

The goal of the sentiment classification is to extract the subjective information conveyed by the word in the form of text like movie reviews, product reviews, twitter. There are a lot of methods for sentiment analysis currently. However, there aren't many articles related to the comparison of Naive Bayes and SVM. In this technology review, the comparison of sentiment classification for Naive Bayes and SVM is proposed.

2 Naive Bayes model for sentiment classification

Naive Bayes is a simple and effective classification method. The Naive Bayes model is a good machine learning method for text classification and has high accuracy for many other areas[1]. In reality, the independent assumption usually doesn't hold for text document data. In contrast, words are conditionally dependent on each other. Although this learning method is based on unrealistic independent assumption, its simplicity just makes the Naive Bayes model useful for many domains. Studies have found that Naive Bayes classifier has high accuracy and can be implemented effectively to large-scale databases[2].

2.1 Naive Bayes Probabilistic Model

According to Narasimha [3], given the problem instance feature variables vector $(F_1, ..., F_n)$, the probability model of Naive Bayes is:

$$p(C|F_1, ..., F_n) \tag{1}$$

The variable C is a dependent variable with a small number of classes. And the value of $p(C|F_1, ..., F_n)$ is the probability of C given the feature vector $(F_1, ..., F_n)$. Naive Bayes classifier is suitable for problems that asking for an optimal classifier. Since Naive Bayes classifier is relying on the probability distributions of data, it might be onerous to get the probability distribution when applying the data.

3 Support Vector Machine classification

In the 1990s, Support Vector Machine was applied commonly. The Support Vector Machine transforms the nonlinear problem into a linear problem. In this way, the SVM could be used as a linear learning machine[4]. SVM has been recognized as effective and useful text classification methods.

3.1 Support Vector Machine Model

Generally, the SVM is used to separate data into two classes by a function. For a set of training data:

$$\{(x^1, y^1, ..., (x^m, y^m)\}, x \in R^n, y \in \{1, -1\} \tag{2}$$

SVM methods generally work well for traditional classification problems. However, the are some critical issues like the determination of the most appropriate feature subspace[5].

4 Comparison of Naive Bayes and SVM

For accuracy, according to Tomas[6], Naive Bayes Classification could achieve a result that is around 1% higher than Support Vector Machine. Tomas concluded that the Support Vector Machine Classification is better than Naive Bayes

Classification. Also, many researchers choose Support Vector Machine Classification for classifying text data.
Sundus proposed that with external enriching the Naive Bayes Classification would be a better choice than Support Vector Machine Classification[7].

Figure 1: Comparison of Developed Classifiers (SVM and Naive-Bayes)[8]

Classifier / Technique	**Dataset (Training and Testing)**	**Accuracy**	**Class**	**Precision**	**Recall**	**F-Measure**
Subjective tweet classification + Polarity Classifier / SVM	Emoticons + Manual Labeling	0.800	**Positive**	0.839	0.873	0.856
			Negative	0.715	0.657	0.685
			Weighted Average	0.799	0.802	0.800
Subjective tweet classification + Polarity Classifier / Naive-Bayes	Emoticons + Manual Labeling	0.777	**Positive**	0.91	0.742	0.817
			Negative	0.616	0.849	0.714
			Weighted Average	0.813	0.777	0.783

According to figure 1, the highest accuracy comes from Support Vector Machine Classification. For sentiment analysis, Support Vector Machine Classification would be better than Naive Bayes Classification[8].

Conclusion

Overall, Naive Bayes Classification is simple and effective. Support Vector Machine Classification is a little more complicated classification method. And for sentiment analysis, the Support Vector Machine Classification would be better compared to Naive Bayes Classification.

References

[1] Qiang Ye, Ziqiong Zhang, and Rob Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, 36(3):6527–6535, 2009.

[2] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, 18(11):1457–1466, 2006.

[3] M Narasimha Murty and V Susheela Devi. *Pattern recognition: An algorithmic approach*. Springer Science & Business Media, 2011.

[4] Fabrice Colas and Pavel Brazdil. Comparison of svm and some older classification algorithms in text classification tasks. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pages 169–178. Springer, 2006.

[5] Yakoub Bazi and Farid Melgani. Toward an optimal svm classification system for hyperspectral remote sensing images. *IEEE Transactions on geoscience and remote sensing*, 44(11):3374–3385, 2006.

[6] Tomas Pranckevičius and Virginijus Marcinkevičius. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221, 2017.

[7] Sundus Hassan, Muhammad Rafi, and Muhammad Shahid Shaikh. Comparing svm and naïve bayes classifiers for text categorization with wikitology as knowledge enrichment. In *2011 IEEE 14th International Multitopic Conference*, pages 31–34. IEEE, 2011.

[8] André Luiz Firmino Alves, Claudio De Souza Baptista, Anderson Almeida Firmino, Maxwell Guimarães de Oliveira, and Anselmo Cardoso de Paiva. A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: a case study with the 2013 fifa confederations cup. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, pages 123–130, 2014.