Twitter News Classification: Theoretical and Practical comparison of SVM against Naive Bayes Algorithms

Inoshika Dilrukshi[#1], Kasun De Zoysa[#2]

University of Colombo, School of Computing, Colombo, Sri Lanka

[1]inoshi.fernando@gmail.com [2]kasun@ucsc.lk

Keywords— **Text classification, SVM, Naive Bayes Classification**

INTRODUCTION

With the development of web technology, researchers had started using blog data in many research aspects and twitter messages is one of them. However, these data are un-organized and thus, it should be organized before gather the information. Classification is one way of organizing the twitter messages. SVM and Naive Bayes classifiers are the most popular classification methods which are often use for text classification. Theoretically, it proves that Naive Bayes performs more faster than any other classifiers with less error. However, this depends on how the situation achieves the Naive Bayes assumptions as naive Bayes assumes that the features are independent. This paper presents a practical experiment to choose a high perform classification method and the theoretical reasons for the high perfonned classification.

A. Data preparation and training

The classification was applied for the short messages of twitter micro blog. The features were extracted from the collected short messages. The bag-of-words are used as features. Using this feature set, the count (frequency) of each word was used to create the dataset. To reduce the dimension, the noise words and the common words were removed from the feature set. The low frequency words can consider as noise words and according to Ziph's law , high frequent data can be consider as common data. Thus, noise words and most common words are removed from the feature set . The prepared data contains 126 of inputs and 1400 records. As, there can be multi groups for one short message, the test was conduct as binary classification problem. 10 fold cross validation was used to reduce the effect of bias. the feature set. The low frequency words can eonsider as noise words and according to Ziph's law, high frequent data can be consider as common data. Thus, noise words and most common words are removed from the feature set. The prepared data contains 126 of inputs and 1400 records. As, there can be multi groups for one short message, the test was conduct as binary classification problem. 10 fold cross validation was used to reduce the effect of bias.

B. Comparison of SVM and Naive Bayes classifier

The comparison was done by measuring the effectiveness of the two classification methods. The effectiveness will be measure using precision and recall values. It is important to calculate a single measurement instead of 2 values. Thus, the weighted harmonic mean, F_β was calculated. For the current situation, the main focus will be on the fraction of retrieved short messages that are relevant. Thus $F_{0.5}$ will be used to measure the effectiveness. Table I shows the $F_{0.5}$. It clearly shows that SVM performs well than Naive Bayes classification. However, this result can be prove using the SVM algorithm. The properties of text messages can be explain as follows. SVM algorithm was capable in dealing with all these features.

TABLE I
AVERAGE F-MEASURES FOR TWO CLASSIFICATIONS

Group	SVM	Naive Bayes
Economic-Business	0.836	0.653
War-terrorist-crime	0.804	0.639
Health	0.976	0.694
Sports	0.881	0.673
Development-government	0.694	0.388
Politics	0.787	0.583
Accident	0.879	0.684
Entertainment	0.978	0.699
Disaster-Climate	0.899	0.647
Education	0.923	0.702
Society	0.819	0.628
International	0.786	0.574

1) Few irrelevant features: In text, there are very few amounts of irrelevant features. Thus, the dimension cannot reduce for an acceptable level.

2) High dimensional input space: Because of few irrelevant features, still it remains large amount of features, even after removing the irrelevant features. This will cause to have over fitting. When using SVM, the high dimensional space do not need to be deal with directly. To obtain the hyper plane, SVM do not deals with all data. It only consider on support vectors. This avoids over fitting.

3) Document vector are sparse: As there are large amount of features (words) and a short message relatively contents less amount of words, in the dataset, there are only few entries which are nor zero (spare vectors).

4) Most text categorization problems are linearly separable : The most text categorization problems are linearly separable. Thus, the objective of SVM is perfectly match with text categorization problems.

5) Words of the short messages are not independent: The words are depend on each other. Thus, the main assumption of Naive Bayes, the features should be independent, will be violated.