International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 1, Issue 3, September – October 2012 ISSN 2278-6856

# Review on Comparison between Text Classification Algorithms

**Vaibhav C.Gandhi, Jignesh A.Prajapati**

Parul Institute Of Engineering & Technology,
Gujarat Technological University,
P.O.Limda-391760,Vadodara, India

**Abstract:** ***Text classification is the automated assignment of natural language texts to predefined categories based on their content. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some way such as producing summaries, answering questions or extracting data. Now a day the demand of text classification is increasing tremendously. Keeping this demand into consideration, new and updated techniques are being developed for the purpose of automated text classification. This paper presents an algorithm for text classification. Describe the three algorithms that we compare: the k nearest neighbors classifier, naive Bayes and the Support Vector Machines. Last, we define the settings of our scenario and the data on which we performed our experiments on.***

**Keywords:** text categorization; KNN; SVM; NAÏVE BAYES;.

## 1. INTRODUCTION

The aim of text categorization is to build systems which are able to automatically classify documents into categories.

To build text classification systems, the bag of words representation is the most often used feature space. Its popularity comes from its wide use in the field of information retrieval and from the simplicity of its implementation. Yet, as in the bag of words representation each dimension corresponds to the number of occurrences of the words in a document, the task of classifying text documents into categories is difficult because the size of the feature space is very high. In typical problems, it commonly exceeds tens of thousands of words. Another aspect that hampers this task is the fact that the number of training documents is several orders of magnitude smaller than the size of the feature space.

With the fast development of the internet technology, the research on text categorization has come into a new stage, all kinds of methods have consecutively got developed, including machine learning technique, has becoming the leading modality of text categorization, for example Naïve Bayesian Classification, KNN , Support Vector Machine , Neural Network and Boosting and so on. Some of these algorithms has already been realized and used in practical systems, and made good effects.

## 2. CONDUCTING FAIR CLASSIFIER OMPARISON

Although Naive Bayes and the k nearest neighbors classifier are multi-class classifiers, the SVM are by default binary classifiers. Then, to handle multi-class problems, SVM usually relies on a one versus all strategy where as many binary classifiers as there are classes are trained. For instance, in the case of a classification problem with n-classes, n one versus the rest binary classifiers are trained.

Therefore, when running experiments on complex classification tasks involving more than two-classes, we are actually comparing n SVM classifiers (for n classes) to single multi-class naive Bayes or k nearest neighbors classifier. We consider this unfair. Therefore, we do not limit the generality of the results by studying only one against one classification problems. In addition, to observe and compare the behaviors of the classifiers when experimental conditions are varying, these conditions must be controled precisely. Indeed, the properties of the training set can influence largely the learning abilities of the classifiers, while in multi-class problems, it can be difficult to understand the particular influence of each class on the classifier's behaviors.

Of the classifiers, while in multi-class problems, it can be difficult to understand the particular influence of each class on the classifier's behaviors. Therefore, in our scenario, we focus on problems with only two-classes. First, it enables us to discard the influence of the multi-class aggregating algorithm in the case of SVM and thus, to compare SVM more fairly with naive Bayes and the k nearest neighbor classifier. Second, it also gives the possibility to control more carefully the properties of the training set. In that regard, in order to give to both classes the same chance to be learned as well, we only studied situations where the number of training instances is the same in each class. Last, as binary problems are smaller

than multi-class problems, they are usually easier and faster to learn, thus facilitating the conduction of experiments.

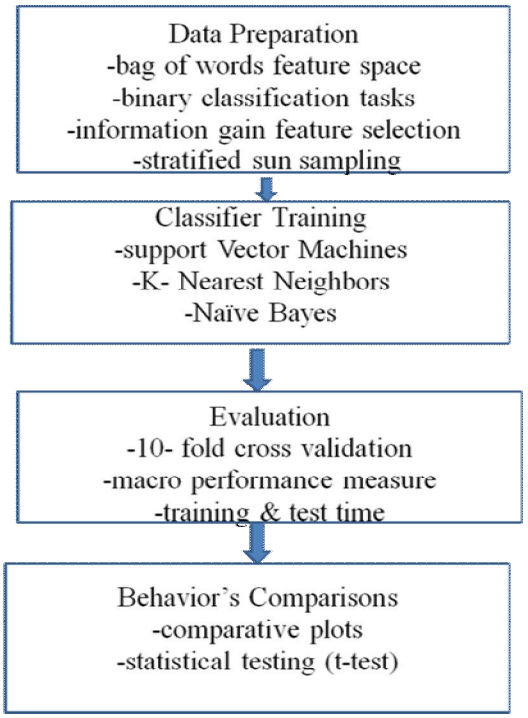


**Figure 1:** A data mining scenario to compare algorithms in the field of text classification.

## 3. CLASSIFICATION OF ALGORITHM

Because of their simplicity and their generally good performance reported in text categorization, we compare the SVM with two well known classifiers, namely the k nearest neighbors classifier and naive Bayes. In the following, we first introduce some general notations and then, we introduce the three classifiers formally.

Consider a database of instances xi and class membership yi, i = 1, ...,N and d the dimension of the feature space, i.e. the dimension of xi. Denote by a function " the mapping in the database between each instance and its class membership such that yi = !(xi). Considering only binary classification problems, this mapping takes values in C = {−1, +1}. Then, a classification algorithm can learn this mapping by training and we denote the estimated classification function by ˆ!(xi).

### 3.1 K - Nearest Neighbors.

Given a test point x! and a predefined similarity metric (sim) that can order the training points by their similarity to x!, a k nearest neighbor classification rule will assign to x! the class having the highest similarity score. These scores are calculated by summing up the similarities of the k nearest neighbors in each class. The classification rule compares these scores and return the class having the highest, it is defined as

$$\hat{\Phi}(\boldsymbol{x}') = argmax_{y' \in \mathcal{C}} P(y') \prod_{j=1}^{d} P(x'_j|y').$$

with K the number of nearest neighbors

$$\delta(y', \Phi(\boldsymbol{x}_k)) = 1 \text{ if } \Phi(\boldsymbol{x}_k) = y', \ 0$$

otherwise.

### 3.2 Naive Bayes

For y ∈ C , let P (y ) be the prior probability of each class. For xij (feature j x.j conditionally to y . Then, given a test point x whose feature values are (x1 , ..., xd ), the naïve Bayes classification function is expressed by

$$\hat{\Phi}(\boldsymbol{x}') = argmax_{y' \in \mathcal{C}} P(y') \prod_{j=1}^{a} P(x'_j|y')$$

### 3.3 Support Vector Machine

The SVM are based on statistical learning theory [Vap95]. Its theoretical foundations together with the results obtained in various fields makes it a popular algorithm in machine learning.

The SVM classification function of a test point x! is given by

$$\hat{\Phi}(\boldsymbol{x}') = sign(\langle \mathbf{w}.\boldsymbol{x}' \rangle + b)$$

With w and the scalar b, the coordinates of the separating hyper plane and the bias to the origin. The particularity of this hyper plane w is that it is the one

Separating the points of the two classes with the maximum distance when these are linearly separable. This concept of maximum separating distance is formalized by the geometrical margin which is defined as

$$\gamma = \frac{1}{2||\mathbf{w}||_2}$$

Therefore, the SVM problem resides in searching the maximum of ! or, alternatively, the minimum of ‖w‖2 given the constraints. To identify this w, an optimization problem must be solved. Its primal form is expressed by

$$minimize_{\mathbf{w},b} \qquad \frac{1}{2}\langle \mathbf{w}.\mathbf{w} \rangle,$$

$$subject\ to \qquad y_i(\langle \mathbf{w}.\boldsymbol{x}_i \rangle + b) \geqslant 1, i = 1, ..., N,$$

In order to limit values of Lagrange coefficients "i, an upper bound C is introduced so that each training instance has a maximum contribution when classes are not linearly separable. This type of SVM is referred to as soft-margin SVM.

Concerning the kernel function, even though problems may not always be separable in text classification, a linear kernel is commonly regarded as yielding the same performance as non-linear kernels in our text classification domain [Yan99b]. For this reason, we only considered linear kernels in our scenario.

Next, recall that the hyper plane w is defined by

$$\mathbf{w} = \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{x}_i .$$

Then, upper-bounding the Lagrange multipliers gives the constraints.

$$0 \leqslant \alpha_i < C.$$

Observe that the norm of the hyper plane tends to vanish as C goes to zero

$$\lim_{C \to 0} ||\mathbf{w}||_2 = 0.$$

This implies that the geometrical margin goes to infinity (||w||2 > 0)

$$\lim_{C \to 0} \gamma = +\infty.$$

Consequently, lowering C to very small values will eventually lead to a SVM solution where all training instances are within the margin. We further discuss the quality of this SVM solution in coming sections.

## 4. IMPLEMENTATION OF ALGORITHM

For naive Bayes and the k nearest neighbors classifier, we used the libbow library [McC96]. With respect to SVM, we used both the Platt's SMO algorithm in libbow and the libsvm implementation of [Cha01]. Therefore, in this thesis,
our conclusions on SVM do not relate to a particular implementation because we could reproduce them for two different implementations.

## 5. CLASSIFICATION OF ALGORITHMS & EVALUATION

The text categorization can be roughly classified as two classes, one is statistical based, e.g. Nave Bayes, the maximum Shannon entropy model, KNN, Support Vector Machine and so on; the second is knowledge based classification method, e.g. Productive rules, neural network etc.

The statistical based classification method due to its simple mathematical computation, not demanding complex

The statistical based classification method due to its simple mathematical computation, not demanding complex semantic knowledge and domain knowledge, has got good effect in practical applications, and becoming popular text categorization method recently. While knowledge based text categorization system can he applied to a specific area, and need the knowledge base of the area as support. Since **its** problems in extraction, modification, maintaining of the knowledge and self-learning, its applicable area is restricted. In addition, there are some other classification methods, such as Boosting Lr: algorithm, it **is** a kind of voting based classification method, the idea **is:** for the task requiring experts knowledge, the effective composition of independent s experts' judgments are better than one. In text categorization, this method uses S different classifiers $\phi_1, \cdots, \phi_s$ (also called weak assumptions) to judge if a text *di* belong to category C;, then merge those judgments for the text category suitably.

To compare these methods in above algorithms in most cases, support vector machine (SVM) and K nearest neighbor (KNN) have better effect, neural network is after them, Naive Bayes **is** the last. And its evaluation index is the break even point (i. e. the classification correct rate got at the point where the precision equals the recall rate). The test document collection is Reuters-21578. The comparison data is as in table l.

Since more than 40 years research, text classification technique has become comparatively mature in some aspects,

Now it has being widely used in library categorization, web sites navigation guidance, and content based email classification and inspection and so on. It can not make semantic understanding of information, so that it has limited effect in processing synonymous, multi semantle meaning, phrase, partial text and text document, and can not get higher precision; on the other hand, even knowledge based classification method can better process definite, big article knowledge, and understand information in semantic sense, but it can not express the uncertain. small article knowledge, bad flexibility and difficulties in constructing knowledge base has strong domain relation, and not easy to transplant. Therefore,

combining these two techniques is inevitable trend. The combination of these two will represent the developing direction of natural language processing based text categorization. Moreover, the better classification effect of boosting method suggests us combining several different methods to do process, e.g. combining support vector machine method with rule-based method, which means to take from other's strong points to offset one's weakness, and comes to the multi-model process technique, the overall performance will be higher than any one of them, so as to increase the precision and efficiency. of course we need to see that it is rather difficult to increase classification effect only depend on improving classification algorithm.

**Table 1:** Comparison data flow different classification algorithms

| Author | Naive Bayes | KNN | SVM | Neural Network |
|---|---|---|---|---|
| Yang[31] | 71.5 | 85.0 | 85.9 | 82.0 |
| Weiss[32] | 73.4 | 86.3 | 86.3 | — |
| Jochims[6] | 72.0 | 82.3 | 86.0 | — |

note: the "—" represents there is not the test results.

The fact has shown that besides the above difficulties, classification system, as a complex system, also has other factors to bring into rather big affect on the classification performance, for example, the selection of document collection sets, the treatment of the feature terms and feedbacks of the classification system, and so on. Therefore, text categorization needs considerations of multi-factor effects on many text processing aspects, to use synthesized method to improve and increase classification performance.

## 6. CONCLUSION

We investigate the problem of automatically classifying text documents into categories which relies on standard machine learning algorithms. These algorithms, given a set of training examples, can learn a classification rule in order to further categorize new text documents automatically. Among the algorithms suggested for use in text classification, the most prominent one is Support Vector Machines and repeatedly, it was shown to outperform other techniques. Yet, we consider that some of the previous comparative experiments of algorithms were not fairly conducted. In fact, other studies have shown that in some situations, other algorithms like naive Bayes or the k nearest neighbors classifier give better results than SVM. Therefore, we first introduced the problem of classifying text documents into categories. Next, with respect to previous comparative studies, we discussed fairness issues when comparing algorithms. Then, given this focus, we described our data mining scenario that aims to compare as fairly as possible classification algorithms. It will help us to better understand the problem of classifying text documents into categories.

## REFERENCES

[1] AN IMPROVED KNN TEXT LASSIFICATION ALGORITHM BASED ON DENSITY- KanshengShi1, Lemin Li2, Haitao Liu1, Jie He3, Naitong Zhang4, Wentao Song1, Proceedings of IEEE CCIS2011

[2] Is Naïve Bayes a Good Classifier for Document Classification

[3] S.L. Ting, W.H. Ip, Albert H.C. Tsang, International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011

[4] Pascal Soucy & Guy Mineau , A Simple KNN algorithm for Text Categorization, IEEE Paper 2002.

[5] April Kontostathis & Richard Liston , The Development of Text-Mining Tools and Algorithms, THESIS. April 2006

[6] Comparison of Text Categorization Algorithms, WUJNS Journal , Vol. 9 No. 5 2004 798-804, 2004

[7] Gülen Toker, Öznur Kırmemiş , TEXT CATEGORIZATION USING k-NEAREST NEIGHBOR CLASSIFICATION

[8] S.Niharika, V.Sneha Latha, D.R.Lavanya. A Survey Of Text Categorization , International Journal of Computer Trends and Technology- volume3Issue1- 2012,

[9] Vidhya. K. A1 & G. Aghila2 , Text Mining Process, Techniques and Tools : an Overview, International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 613-622