Volume 6, Issue 4, April 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparison of Performance of Various Methods Used for Text Classification: A Survey

Debika Mishra, Subhalaxmi Das
Department of CSE, College of Engineering and Technology, Bhubaneswar,
Odisha, India

Abstract— This document portrays a brief comparative analysis of different algorithms used for Text classification. With the increase in volume of information, it has become quite necessary to devise efficient methods for Text classification. This paper takes into account three algorithms namely SVM, BPNN and TESC and considers the accuracy of classification as a performance measure.

Keywords— Classification, semi-supervised, SVM, BPNN, TESC, clustering

I. INTRODUCTION

Due to the facilities provided by the World Wide Web and rapid advances in storage media, storing capacities; sharing and gathering information has become quite convenient and easy. This has resulted in a voluminous increase in information available publicly. And these comprise of all the categories, for example news, movie reviews, weather reports, historical articles etc. But many of these information is routinely available in unstructured and text format, and copious in volume.

Now in order to make this information of use to people, they need to be categorised properly. Various classification techniques have been proposed over time to enable text classification. Semi-supervised classification is being stressed upon in this survey which is an interesting technique of classification. Only pre-classified data is used to train traditional classifiers. But the problem lies in the fact that pre-labelled data is quite tedious to obtain, and are generally not cheap, and also more time is required to get them, as they need experienced manual effort. Whereas raw data that has not been classified is quite easy to get but the problem lies in the fact that not many techniques have been proposed to use the same. Semi-supervised learning attempts to resolve this issue by putting to use such large amount of raw data, along with the pre-classified data, to obtain improved classifiers. The advantage of semi-supervised learning is that very little manual effort is needed and quite accurate results are obtained, which is precisely why it has become a topic of interest in practice as well as theory

II. CLASSIFICATION OF TEXT (USING SSC)

Text categorization or text classification is the method by which categories are labelled to documents which can be anything from web pages, news, movies, etc. and categories can be based on the way of writing, subject, organisation of words etc. But in spite of the type of approach used for classification, the process always begins with a training set $\boldsymbol{T} = (t_1,\ldots,t_n)$ of data that is pre-classified with a label say $L \in C$ (e.g. news, movies). The work after that is to find a categorization method as in Eq. (1) which can classify new data to the correct label/area.

$$f: D \rightarrow C \; f(d) = L \quad (1)$$

The performance can be appraised by accuracy that is calculated by measuring the fraction of documents that have been labelled correctly to that of the total number of documents. This can be facilitated by keeping aside a random portion pre-classified documents and not using them for training. Then the classification model can be used to classify these test sets and the estimated labels can be compared to true ones. The above mentioned process is a fundamental method for performance evaluation[1].

A. Semi-supervised clustering (SSC)

This method is used to classify raw data by taking help of the knowledge gained from pre-labelled data (that is used to train the system). This data that is already classified helps in training of the algorithm and it has been proved to lead the clustering algorithms in the direction of better clustering results.

Owing to its great success since last few years, this method is gaining significant attention from researchers worldwide. A semi-supervised clustering method requires the presence of two information i.e. the similarity measure that is to be used for the purpose of clustering and some constraints like (cannot-link or must-link). Significant advantage can be obtained from this method, only if both the above mentioned information are not opposite to each other. But the disadvantage of this method as compared to traditional clustering is that very few methods have been developed till date. The major difference between the above two types of clustering is the way the above mentioned information is mixed together to classify data i.e. either by adapting similarity measures or by modifying the search for appropriate clusters.

III. SCHEMES FOR SSC

Presently there exist two schemes for SSC : **a-priori** scheme, and the **interactive** scheme. Their difference lies in the way the side data that is required for processing is collected in each scheme. In the former one, all side information is given once before the SSC algorithm is executed while in the latter the side information is collected iteratively by interacting with the supervisor.

A. A Priori Scheme

In the a priori scheme (shown in Fig. 1), the SSC algorithm reads all the side information once and uses these information to improve the clustering performance.

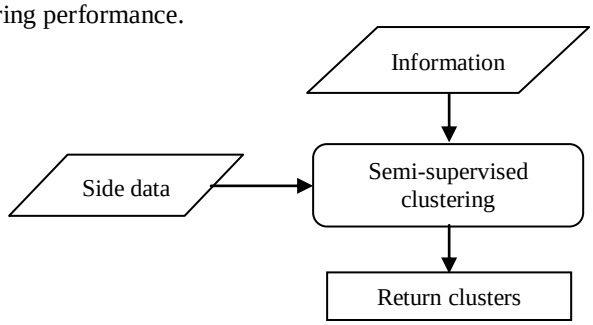

Figure 1. A priori Scheme

Existing methods in this scheme for semi-supervised clustering are divided into two approaches i.e *constraint-based* and *metric-based* methods.

In **constraint-based approaches**, the clustering algorithm itself is modified so that user-provided labels or constraints are used to get a more appropriate clustering. This can be done in several ways, such as :

- by performing a transitive closure of the constraints and using them to initialize clusters,
- by including in the cost function a penalty for lack of compliance with the specified constraints, or
- by requiring constraints to be satisfied during cluster assignment in the clustering process

In **metric-based approaches**, an existing clustering algorithm that uses a particular distortion measure is employed; however, the measure is first trained to satisfy the labels or constraints in the supervised data. Several similarity measures were employed for similarity-adapting semi-supervised clustering: the Jensen-Shannon divergence trained with gradient descent, the Euclidean distance modified by a shortest-path algorithm or Mahalanobis distances adjusted by convex optimization .Among the clustering algorithms using such adapted similarity measures we can mention hierarchical single-link or complete link clustering and k-means.

B. Interactive Scheme

In this scheme (as demonstrated in Fig. 2), a query is presented to the supervisor with the clustering result. The supervisor can be an oracle system or a user. This supervisor analyses the provided result and returns feedback to the algorithm. This algorithm then studies this input from the supervisor and bias the algorithm based on it. This interaction is to be halted only when some condition of convergence has been reached.

Based on the algorithm used and the role of the user, the feedback can be obtained in following two ways:

- If the supervisor plays the active role, then he/she actively provides the constraints to the SSC algorithm.
  In the case that the SSC algorithm is the active role, the SSC algorithm will pose queries to the supervisor, and the supervisor is supposed to answer these queries.

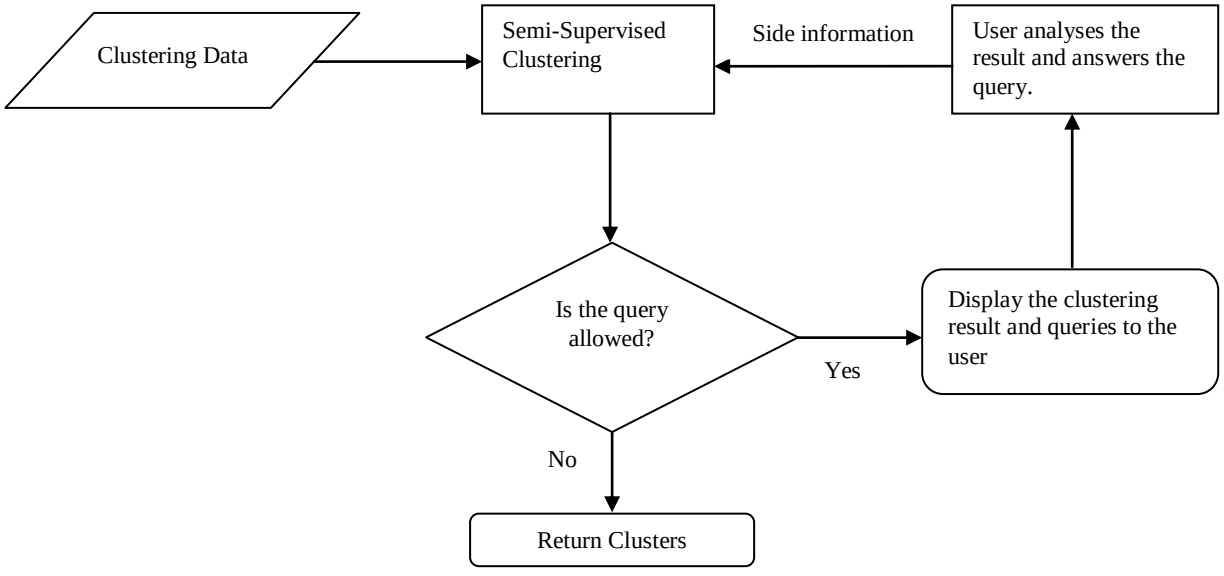

Figure 2. Interactive Scheme

It has been seen that the first method is outperformed by the second one. This is so because in the first method, the user needs to have knowledge of the best constraints to be provided to the algorithm. But in the second one, the tedious work falls on the algorithm. Moreover it is for the best if the algorithm rather than passively receiving feedback puts query to user regarding the things that are not clear.

The algorithms in the first approach will be referred as the passive SSC algorithms, while the ones in the second approach will be called the active SSC algorithms.

IV. ALGORITHMS FOR TEXT CLASSIFICATION

In recent times, researchers have developed various kinds of approaches for Text classification. Four of which have been explained briefly in this section.

A. Support Vector Machine(SVM)

Statistical learning theory is the fundamental base of the SVM model. It first came to light in 1995 (Mulier,1999). The classification model using this theory was first developed by Chervonenkis and Vapnik.[2] SVM was designed to always give globally optimal solution. That was done on the basis of the Kernel and the VC theory (Taylor and Cristianini, 2000), where SVM was proposed which can solve a quadratic programming problem that is linearly constrained. The following optimization problem is used for the purpose of training a SVM for non-separable case.

$$\boldsymbol{min_{\omega}\ \frac{1}{2}\ \|\omega\|^2 + C\sum_i \xi_i} \qquad \textbf{(2)}$$

with constraints

$$\boldsymbol{y_i(x_i w + b) \geq 1 - \xi_i} \qquad \boldsymbol{\xi_i \geq 0\ \forall i}$$

B. Back Propagation Neural Network (BPNN)

Rumelhart, Hilton and Williams (1986) Back Propagation method was presented by Williams, Hilton and Rumelhart (1986). It was used for the purpose of weight updation of a network having multiple layers through the process of supervised learning. The learning is through a process of iteration in which a set of training instances are processed, then the predicted value of the network is compared with the known value that is expected to be the target. With every training instance, modification of weight takes place so as to bring the MSE i.e. the error between the predicted and the target value to a minimum. This algorithm carries on the iteration process in two ways: forward i.e. Input towards the output layer, and backward is the reverse i.e. from output to input. Both of these processes are quite the same; the only difference is the way the MSE i.e the error is sent through the network so as to modify the weights. For each training instance, there is always a target value. Details of this algorithm, readers can be found in Han and Kamberl (2006).[2]

C. Text Classification using semi-supervised clustering (TESC)

The fundamental idea behind this method is to classify one type of texts from the other. Therefore, clustering is put to use for the purpose of identification of components in the collection of texts. In this method, previously classified texts are used to get the outline of the clusters and the raw data is made to adapt to the centroid of these clusters. The category of the texts in a particular cluster denotes the category of the said cluster. Whenever any raw data is in question, the similarity of the same is compared with the existing clusters and then the most near ones to it is assigned to it i.e. the raw data is labelled with the label of that particular cluster. This process contains two steps: one is the separation of labelled and unlabelled data and the second is the method of predicting the label of each raw data [1]

V. COMPARISON OF THE ABOVE ALGORITHMS

All the algorithms as mentioned above have been shown to work well for text classification but TESC outperforms the other two methods. Reuters-21578 distribution 1.0 has been used for this by (Yoshida, Tang and Zhang), which is available online[1]. The accuracy of text categorization have been measured in the different retaining ratios i.e. when the said ratio is 0.2, it means that 20% of the pre-classified documents have been kept separate for the training purpose. The following results were obtained in the best and the worst case scenario.

Table 1

Description of the algorithm	TESC: Text classification using semi supervised clustering	SVM: Support Vector Machine	BPNN: Back Propagation Neural Network
Performance during best case (retaining ratio=0.3)	Accuracy=0.8943	Accuracy=0.7690	Accuracy=0.7702
Performance during worst case (retaining ratio=0.8)	Accuracy=0.8881	Accuracy=0.8733	Accuracy=0.8723

The accuracy is measured in terms of ratio i.e. the fraction of documents that have been correctly classified to the total number of documents.

VI. CONCLUSION

This paper presents a comparison of algorithms used for Text Classification. TESC has been shown to outperform the SVM and BPNN algorithms in both the best case as well as worst case scenario. But, this algorithm can be developed more to classify data in multidimensional space as well.

REFERENCES

[1] W. Zhang et al. , TESC: an approach to text classification using semi supervised clustering ,Knowledge-Based Systems 75 (2015) 152–160.

[2] W. Zhang, X. Tang, T. Yoshida, Text classification toward a scientific forum, J. Syst. Sci. Syst. Eng. 16 (3) (2007) 356–369.

[3] I.W. Tsang, J.T. Kwok, P. Cheung, Core vector machines: fast SVM training on very large data sets, J. Mach. Learn. Res. 6 (2005) 363–392.

[4] W. Zhang, T. Yoshida, X. Tang, Text classification based on multi-word with support vector machine, Knowl.-Based Syst. 21 (8) (2008) 879–886.

[5] X. Zhu, Semi-Supervised Learning Literature Survey, Technical Report No. 1530, Computer Sciences, University of Wisconsin-Madison, 2006

