Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification

Aaditya Jain
M.Tech Scholar, Department of Computer Science & Engg., R. N. Modi Engg. College, Rajasthan Technical University, Kota, Rajasthan, India

Jyoti Mandowara
Department of Computer Science & Engg., R. N. Modi Engg. College, Rajasthan Technical University, Kota, Rajasthan, India

ABSTRACT

Text categorization or document classification is one of the major tasks in text data mining and information retrieval. Many efficient classifiers for text classification have been proposed till date. However, the individual classifiers show limited applicability according to their respective domains and scopes. Recent research works evaluated that the combination of classifiers when used for classification showed better performance than the individual ones. Our work provides description about text classification process and related popular classifiers. In this paper, the numbers of approaches dealing with combining text classifiers for improving the efficiency in the field of text classification are also surveyed.

Keywords

Text Classification, Text Classifiers, Combining Text Classifiers, Naïve Bayes, KNN, SVM, and Maximum Entropy.

1. INTRODUCTION

The amount of text available for analysis has increased hugely in recent years due to the social networking, micro blogging and various messaging/ bulletin board systems. Besides these, many articles, news feeds and documents are now available in soft copy. An important step in text classification is to classify the text documents among some known set of classes/ categories. Classifying a document into a specified domain is significant because most of other problems in text mining, such as text summarization, information extraction, discovering semantic relations, etc. only achieve high performance/results in a particular domain.

The task of text mining can be done through two processes: Classification and Clustering. While clustering is an unsupervised learning approach which aims at grouping a set of related data objects into clusters based on some similarity or distance measure between them. Classification is a supervised form of machine learning aiming at identifying the category from a set of categories to which a selected text belongs [1]. It is done on the basis of a predefined training dataset.

2. TEXT CLASSIFICATION PROCESS

Text Classification (TC) may be formalized as the task of approximating the unknown target function $\Phi: D \times C \rightarrow \{T, F\}$, that describes how documents to be classified, according to a supposedly authoritative expert by means of a function(Φ) called the classifier [2]. Here $C = \{c_1 . .. c|C|\}$, is predefined set of categories and D is a (possibly infinite) set of documents. If $\Phi(d_j, c_i) = T$(true), then $d_j$is called a positive example (or a member) of c_i, while if $\Phi(d_j, c_i) = F$ (false) then d_j is called a negative example (or non-member) of c_i.

The classification process in itself is a very detailed process consisting of various stages [2], [3], [4]. Each stage then has a set of methods to choose from depending on the text and the given classification problem.

- The classification process starts by collection of documents in different formats like html, web content etc.
- Converting the documents to clear word format followed by tokenizing and stemming are parts of the preprocessing stage aiming to make the text ready for the classification stage.
- Indexing, also a part of the preprocessing stage, then converts the full text document to a document vector.
- The most important stage is that of feature selection involving selection of subset of features important for the classification based on some predetermined measure. There are various proposed feature selection methods for the same.
- After the feature selection stage, the text is now ready to be applied to a feasible classification algorithm, also known as “classifier” from among the various classifiers proposed till date.
- The last stage of the classification process measures the performance of the classification done.

3. POPULAR CLASSIFIERS

Classifiers are algorithms for performing the task of classification. Various text classifiers have been proposed till date.

3.1 Naïve Bayes Classifiers

Naïve Bayes Classifiers are simple probabilistic classifiers based on the Bayes Theorem [5]. These are highly scalable classifiers involves a family of algorithms based on a common principle assuming that the value of a particular feature is independent of the value of any other feature, given the class variable. In practice, the independence assumption is often violated, but Naive Bayes classifiers still tend to perform very well under this unrealistic assumption and very popular till date.

3.2 Maximum Entropy

Maximum Entropy (ME) modeling is a general and intuitive way for estimating a probability from data and it has been successfully applied in various natural language processing tasks such as language modeling, part-of-speech tagging and text segmentation [6], [7]. The principle underlying ME is that the estimated conditional probability should be as uniform as possible, that is, it should have the maximum entropy. The main advantage of ME modeling for the classification task is that it offers a framework for specifying any arbitrary relevant information. But the algorithm which is used to find the solution can be computationally expensive if complexity of the problem is high.

3.3 Rocchio’s Algorithm

Rocchio’s classifier is a feedback approach developed using the Vector Space Model [8]. The algorithm is based on the assumption that most users have a general conception that documents should be denoted as relevant or non-relevant. Therefore, the user's search query is revised to include an arbitrary percentage of relevant and non-relevant documents as a means of increasing the search engine's recall, and possibly the precision as well, relevance being judged to allow the documents to enter a query. It is easy to implement and efficient in computation but poor performance when proportion of relevant documents in the whole corpus is low.

3.4 K- Nearest Neighbours

K-Nearest Neighbor Classification by is a non-parametric method for classification and is among the simplest of the classification algorithms [9]. It is a method of lazy learning because the function is locally approximated and all computation is postponed until classification, the output of the classification being a class membership. Classification of an object is done seeing the commonality among its k-nearest neighbors where k is a positive integer. If k = 1, then the object is simply assigned to the class of that single nearest neighbor. Advantage includes easy implementation and non parametric properties but classification process takes long time to conclude.

3.5 Support Vector Machine

A Support Vector Machine (SVM) is a supervised text classification algorithm [10]. The SVMs are different from other classification methods because they need both positive and negative training set, to seek for the decision surface that best separates the positive from the negative data in the n-dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vectors. SVMs aim to find linear separators to solve text classification problems. When comparing with Naive Bayes, SVMs show outperformed classification effectiveness. Drawbacks of SVMs are that it can be applied only to binary classification and problems arise in representing documents into numerical vectors.

4. NEED OF COMBINING TEXT CLASSIFIERS

All the popular classifiers explain in the previous section have been proved great in executing their classification tasks, but overall no single classifier can be stated as the best or a standard model for text classification. This is probably due to the fact that each classifier has its particular application or working domain or are they limited to a particular kind of data for text classification.

Recent work in this direction was the idea of combining various classifiers or performing hybrid approach. This approach focuses only on the individual advantages of the classifiers discarding their shortcomings and testing if they worked beneficially or not. Evaluation studies of all the related work were in favor of the combination of classifiers.

In short, combining classifiers takes into account only the gains of using those classifiers that match the needs of the classification diminishing each classifier’s individual limitations.

5. APPROACHES FOR COMBINING TEXT CLASSIFIERS

Text Classification can be further classified into discriminative techniques, like SVMs [10], Decision trees [11] and generative or probabilistic techniques related to the aspect model like Naïve Bayes [5] and Maximum Entropy [6]. When considering combination of classifiers, classifiers belonging to the same paradigm can be combined.

Larkey and Croft in [12] propose the combination of three classifiers, KNN, Relevance feedback and Bayesian’s independence classifiers to be used in the medical domain for automatic assignment of ICD9 codes. The task was done first with individual classifiers and then with combined to check the effectiveness of both the approaches and the hybrid approach was concluded better. The performance of the classifiers were measured based on document ranks. This is an example where classifiers are used for document ranking. The approach is of using weighted linear combination.

Bennett et al. in [13] proposed a probabilistic method to combine the classifiers such that the contribution of a classifier depends on its reliability. The reliability is measured through reliability indicators which are linked to the regions where a classifier might perform relatively good or poor. Instead of the rank of document, the indicators are based on performance of the classifier itself thus making the proposal more generalized.

Grilheres et al. in [14] published detailed study of effect of combining classifiers to classify multimedia documents into heterogeneous classes. Various combinations are applied to a five thousand web pages document of the European Research Project Net Protect II and experiment results prove that with a prior knowledge on classifiers, better filtering performances are possible. The approaches used for combining are both voting-based and logic-based.

Besides the conventional style of linear or voting based combination a new technique based on Dampster-Shafer theory was proposed by **Sarinapakrn et al.** in [15]. Their main aim is fusion of sub-classifiers since the application is towards multi-label classification.

Dino Isa et al. in their two successive papers [16] and [17] have proposed a novel idea as to how meta-outputs of a Naïve Bayes technique can be used with SVM and Self-organizing maps (SOM) respectively. Bayes formula is used to convert the text document into a vector space where the values denote the probabilities of documents towards any class depending on the features contained. This is called the vectorisation phase of the classifier. It is common to both the classifiers. SVM is then applied on this vector space model for final classification output. The proposal had improved classification accuracy compared to the pure naive Bayes classification approach. In [17] the probability distributions obtained by Bayes technique are followed by an indexing step done through SOM to retrieve the best match cases. SOM is similar to clustering of documents based on a similarity measure between the documents like Euclidean distance.

Miao et al. [18] considered very different combination of classifiers, namely KNN and Rocchio methods. A variable precision rough set is used to partition the feature space to lower and upper bounds of each class. Each subspace is classified through Rocchio technique. But it fails when the arriving document is in boundary region, here kNN is used. This presents a new style of combining classification methods to overcome each others' drawbacks.

A more recent research by **K. Fragos et al.** in [19] also concludes in favor of combining different approaches for text classification. The methods that authors have combined belong to same paradigm – probabilistic. Naïve Bayes and Maximum entropy classifiers are chosen to test on the applications where the individual performance is good. The merging operators are used above the individual results. Maximum and Harmonic mean operators have been used and the performance of combination is better than the individual classifiers.

S. Keretna et al. [20] have worked on recognizing named entities from a medical dataset containing informal and unstructured text. For this, they combine the individual results of Conditional Random Field (CRF) classifiers and Maximum Entropy (ME) classifiers on the medical text; each classifier trained using a different set of features. CRF concentrates on the contextual features and ME concentrates on the linguistic features of each word. The combined results were better than the individual results of both the classifiers based on Recall rate performance measure.

S. Ramasundaram et al. [21] aimed to improve the N-grams classification algorithmby applying Simulated Annealing (SA) search technique to the classifier. The hybrid classifier NGramsSA brought about an improvisation to the original NGrams classifier while inheriting all the advantages of N-grams approach. Feature reduction using χ^2 method is used but its multivariate value among the n-grams affects the performance of the classifier.

6. CONCLUSION

Combining classifiers has become a promising research area now a day. This paper gives an insight about the various methodologies that can be used for combining classifiers. Some of the relevant works for each method of combination are also discussed. And it is clear that the results obtained from the combination of classifiers are much better than the results obtained by the same classifiers individually. This further gives a boost to the research headed in providing faster and more efficient text classification process through classifier combinations.

7. REFERENCES

[1] Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar, "Foundations of Machine Learning", The MIT Press, ISBN 9780262018258, 2012.

[2] Vandana Korde, C Namrata Mahender, "Text Classification and Classifiers: A Survey", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[3] Pratiksha Y. Pawar and S. H. Gawande, "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.

[4] Kjersti Aas and Line Eikvil, "Text Categorization: A Survey" Report No. 941. ISBN 82-539-0425-8., June, 1999.

[5] D. Lewis, "Naive Bayes at Forty: The Independence Assumption in Information Retrieval", Proceedings of the 10th European Conference on Machine Learning(ECML-98), 1998.

[6] Adam L. Berger, Vincent J. Della Pietra and Stephen A. Della Pietra "A maximum entropy approach to natural language processing" Computational Linguistics, Volume 22 Issue 1, pp. 39-71, March 1996.

[7] Kamal Nigam, John Lafferty and Andrew McCulllum, "Using Maximum Entropy for Text Classification", IJCAI-99, Workshop on Machine learning for Information Filtering, pp. 61-67, 1999.

[8] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "An Introduction to Information Retrieval", Cambridge University Press, page 181, 2009.

[9] Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, "KNN Model-Based Approach in Classification", Proceedings of the ODBASE, pp- 986 – 996, 2003.

[10] A. Basu, C. Waters and M.Shepherd, "Support Vector Machines for Text Categorization", Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003.

[11] Manish Mehta and Rakesh Agrawal "SLIQ: A Fast Scalable Classifier for Data Mining" 1996.

[12] L.S. Larkey. and W. B. Croft, "Combining classifiers in text categorization",Proc. SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval (Zurich, CH, 1996), pp. 289–297 1996.

[13] Paul N. Bennett, Susan T. Dumais, Eric Horvitz, "Probabilistic Combination of Text Classifiers Using Reliability Indicators: Models and Results", Proceedings of 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, August 2002. ACM Press.

[14] B. Grilheres, S. Brunessaux, and P. Leray, "Combining classifiers for harmful document filtering", RIAO '04 Coupling approaches, coupling media and coupling languages for information retrieval, Pages 173-185, 2004.

[15] Kanoksri Sarinnapakorn and Miroslav Kubat, "Combining Subclassifiers in Text Categorization: A

DST-Based Solution and a Case Study", IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 12, December 2007.

[16] Dino Isa, Lam Hong lee, V. P Kallimani, and R. Raj Kumar, "Text Documents Preprocessing with the Bayes Formula for Classification using the Support vector machine", IEEE Transactions of Knowledge and Data Engineering, vol.20, no. 9, pp.1264-1272, September 2008.

[17] Dino Isa, V. P Kallimani and Lam Hong lee, "Using Self Organizing Map for Clustering of Text Documents", Expert System with Applications, vol. 36, no. 5, pp. 9584-9591, July, 2009.

[18] Duoqian Miao , QiguoDuan, Hongyun Zhang, and Na Jiao, "Rough set based hybrid algorithm for text classification", Journal of Expert Systems with Applications, vol. 36, no. 5, pp. 9168-9174, July 2009.

[19] K. Fragos, P.Belsis, and C. Skourlas, "Combining Probabilistic Classifiers for Text Classification",Procedia - Social and Behavioral Sciences, Volume 147 Pages 307–312, 3rd International Conference on Integrated Information(IC-ININFO), doi: 10.1016 /j.sbspro .2014.07. 098 , 2014.

[20] S. Keretna, C. P. Lim and D. Creighton, "Classification Ensemble to Improve Medical Named Entity Recognition", 2014 IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, USA, 2014.

[21] S.Ramasundaram, "NGramsSA Algorithm for Text Categorization", International Journal of Information Technology & Computer Science (IJITCS), Volume 13, Issue No : 1, pp.36-44, 2014.