SVM Based GA Text Classification

[1] Geetanjali kukade, [2] Dharmendra Sharma
[1]M.E. Student, [2]Assistant Professor
[1]Computer Science and Engineering, SDBCE, Indore, M.P, India
[1]Computer Science and Engineering, SDBCE, Indore, M.P, India
[1]kukade.geetanjali@gmail.com, [2]dharmendrasharma@sdbce.ac.in

***Abstract*— This paper proposed the use of Support Vector Machines (SVMs) for learning text classifier from examples. It analyzes the particular properties of learning with text data and ideates. We proposed text classification which is based on genetic algorithm. There are other two open problems in text mining: polysemy and synonymy. Polysemy refers to the fact that a word can have multiple meanings. Distinguishing between different meanings of a word (called word sense disambiguation) is not easy. In order to explicitly capture the optimality of word clusters in an information theoretic framework we will apply the genetic algorithm SVM parameters modeling to social networking sites for text classification and our investigational outcome will illustrate that SVM based GA significantly better than Naïve Bayes and SVM, even when the data is noisy or partially labeled. Our work includes a proportional revision on classification of the data set with added machine learning algorithms such as support vector machines, genetic algorithm parameters. Since character genetic algorithm SVM parameters to be effective in text categorization, we plan to explore their competence in information retrieval tasks for agglutinative languages.**

***Index Terms*— SVM based GA, Text classification, information retrieval.**

I. INTRODUCTION

Text classification has extended been a significant research topic in information retrieval (IR) related areas. In the literature, the many researchers projected model has been extensively used to categorize a document in text classification and many other applications. However, which ignores the relationships between terms, offers a rather poor document illustration. Some preceding research has exposed that incorporating language models into the Machine Learning Algorithms can get better the performance of text classification. Although the widely used vector space model (VSM) can utilize the relationships between words to some extent, they cannot model the long-distance dependencies of words. In this research, we study the term genetic algorithm SVM parameters modeling approach within the translation framework for Text classification. This paper proposes new model call the term genetic algorithm svm parameters modeling to incorporate term associations into the document.

II. RELATED WORK

Wherever YunFei Yi in at al [1] proposed a method to compute the co occurrence based term association (CBTA) to address the problem of keyword based method. The idea of term association is original proposed as an information retrieval method for query expansion. It also works well when apply it to text classification. Term associations are based on the hypothesis that terms have relationships if they co-occur often in documents. The associated terms then have been used for query expansion and text classification.

Lei Shi in at al [2] proposed a new approach to CLTC, which trains a classification model in the source language and ports the model to the target language, with the translation knowledge learned using the EM algorithm. , they have method takes into account different possible translations for model features. The translated model serves as an initial classifier for a semi-supervised process, by which the model is further adjusted to fit the distribution of the target language. They have method does not require any labeled data in the target language, nor a machine translation system. Instead, the only requirement is a reasonable amount of unlabeled data in the target language, which is often easy to obtain.

Sergio Di Martino in al [3] in this paper we have presented an approach that exploits a Genetic Algorithm (GA) to configure Support Vector Machines (SVMs) for predicting fault-prone software components, proposed approach has been assess by an experimental analysis meant to verify the effectiveness of the approach to configure SVMs. As datasets, we have employed data on jEdit, a well-known text editor written in Java,

Jianguo Zhou in at al [4] this paper applies a. classifier, hybridizing rough set approach and improved support vector machine(SVM) using genetic optimization algorithm (GA), to the study of credit risk assessment in commercial banks. We can get reduced information table, which implies that the number of evaluation criteria, such as financial ratios and qualitative variables is reduced with no information loss through rough set approach.

Wei Xu in at al [5] in this paper, a novel support vector machine (SVM) based ensemble model is proposed for credit risk assessment. In the proposed method, principles component analysis (PCA) is firstly employed for credit feature selection. Secondly, SVMs with different kernels are trained by using genetic algorithm (GA) to optimize the parameters, and the corresponding assessment results are obtained. Thirdly, all results produced by different SVMs are combined by several ensemble strategies.

Alec Go in at al [6] Introduce a novel approach for automatically classifying the sentiment of Twitter messages. These messages are classier as either positive or negative with respect to query term. This is useful for consumers who want to re- search the sentiment of products before purchase, or com- ponies that want to monitor the public sentiment of their brands.

Bernard J. Jansen in al [7] in this paper, we investigating micro-blogging as a form of online word of mouth branding. We analyzed micro-blog postings containing branding comments, sentiments, and opinions. We investigated the overall structure of these micro-blog postings, types of expressions, and sentiment fluctuations. We discuss the implications for organizations in using micro-blogging as part of their overall marketing strategy and branding campaigns.

Jae H. Min in at al [8] this paper applies support vector machines (SVMs) to the bankruptcy prediction problem in an attempt to suggest a new model with better explanatory power and stability. To serve this purpose, we use a grid-search technique using 5-fold cross-validation to find out the optimal parameter values of kernel function of SVM. In addition, to evaluate the prediction accuracy of SVM, we compare its performance with those of multiple discriminant analysis (MDA), logistic regression analysis. Times are specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts. True Type 1 or Open Type fonts are required. Please embed all fonts, in particular symbol fonts, as well, for math, etc.

III. PROPOSED METHODOLOGY

Before Subsequent to feature selection and transformation the documents can be effortlessly represent in a form that can be used by a SVM based GA algorithm. Numerous text classifiers have been proposing in the literature using machine learning techniques, probabilistic models. They often differ in the approach adopted: decision trees, naive-Bayes, rule induction, neural networks, nearest neighbors, and lately, support vector machines. Though many approaches have been proposed but automated text classification is still a major area of research mainly because the efficiency of current automated text classifiers is not faultless and still needs development. Naive Bayes is often used in text classification applications and research because of its simplicity and effectiveness. However, its performance is often degraded because it does not model text well. Schneider addressed the problems and show that they can be solved by some simple corrections. Klopotek and Woch presented results of empirical evaluation of a Bayesian multinet classifier based on a new method of learning very large tree-like Bayesian networks. The study suggests that tree-like Bayesian networks are able to handle a text classification task in one hundred thousand variables with sufficient speed and accuracy.

Johnson et al. describe a fast decision tree building algorithm that takes advantage of the sparsity of text data, and a rule simplification method that converts the decision tree into a logically equivalent rule set.

Support vector machines (SVM), when applied to text classification provide excellent precision, but poor recall. One means of customizing SVM is to improve recall, is to adjust the threshold associated with an SVM. To propose an automatic process for adjusting the thresholds of SVM based GA with better results.

The level of difficulty of text classification tasks naturally varies. As the number of distinct classes increases, so does the difficulty, and therefore the size of the training set needed. In any multi-class text classification task, inevitably some classes will be more difficult than others to classify. Reasons for this may be:

1. Very few positive training examples for the class, and/or
2. Lack of good predictive features for that class.

When training a binary classifier per category in text categorization, we use all the documents in the training corpus that belong to that category as relevant training data and all the documents in the training corpus that belong to all the other categories as non-relevant training data. It is often the case that there is an overwhelming number of non relevant training documents especially when there is a large collection of categories with each assigned to a small number of documents, which is typically an "imbalanced data problem".

This problem presents a particular challenge to classification algorithms, which can achieve high accuracy by simply classifying every example as negative. To overcome this problem, cost sensitive learning is needed.

Recently in the area of Machine Learning the concept of combining classifiers is projected as a new direction for the development of the performance of individual classifiers. Numerous methods have been suggested for the creation of ensemble of classifiers. Mechanisms that are used to build ensemble of classifiers include.

We explore dissimilar technique of automatic text classification that have been widely used and have been exposed to be competitive techniques in the machine learning and information retrieval literature. We provide an approach for the information retrieval**.** We create a svm based GA classifier based upon the training examples. This classifier is then used to classify the text by providing the probability that each piece of knowledge belongs to which class. These pieces of knowledge are then treated as training examples and are added into the original set to create a new classifier. This method had been used with unlabeled examples, with the implicit assumption that each unlabeled example definitely fits into one of the classes of the problem. Our observation is that information retrieval from the same domain can be helpful in this approach as well. This is the case even when the pieces do not clearly correspond to any one of the classes that are used to label the training and test sets. This system present new and different uses of background knowledge. We incorporate background knowledge into a nearest neighbor classification algorithm. Our propose approach finds those training examples that are closest in proximity (using information retrieval metrics)

to a new test example. These close neighbors then vote to determine the class of the test example. With the addition of text information retrieval, we once again find those training examples that are closest to the test example. However, we do so by finding the pieces that are closest to the test example, and comparing those pieces to the training corpus. In this way, a training and test example are not compared directly but are considered close if there exist an information retrieval piece of knowledge that is close to both the training and test example. Many such close training examples are found, and they vote for the final classification result. To accomplish this indirect comparison we use the text database. Test examples are expressed in this same reduced space and are then compared to the training examples. Those training examples that is closest to the test example. In this space vote to obtain the final results of classification. This result used in the reexpression of both training and test data in a space that was created using the information contain. Since the purpose of Latent Semantic Indexing is to find relationships between words in the corpus, the inclusion of background knowledge allows us to model relationship that cannot be found in the training set alone. Statistical methods routinely used for textual analyses of all kinds Machine translation, part-of-speech tagging, information extraction, question-answering, text categorization, etc. Not reported in the statistical literature Text Categorization Automatic assignment of documents with respect to manually defined set of categories Applications automated indexing, spam filtering, content filters, medical coding, essay grading.

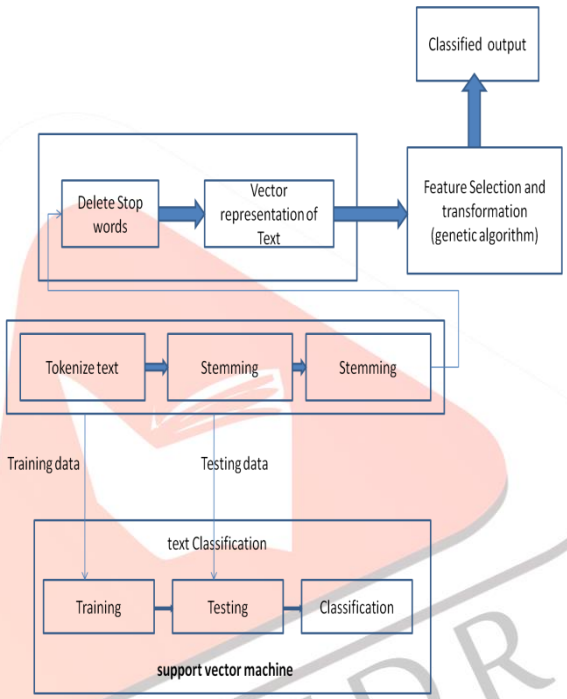

Figure 1: Text classification system architecture

IV. CONCLUSION

We proposed a novel technique for text classification using machine learning to address three common challenges in real-world classification applications, i.e. how to use domain knowledge, how to resist noisy samples and how to use text unlabeled data. These two learners enhance each other through an iterative process to improve the final classification performance. We will evaluate its effectiveness in real applications; we applied novel technique text classification using machine learning to on-line blog and social network classification.

REFERENCES

[1] YunFei Yi, Lijun Liu, Cheng Hua Li, Wei Song," Machine Learning Algorithms with Co-occurrence based Term Association for Text Mining" Fourth International Conference on Computational Intelligence and Communication Networks-2012.

[2] Lei Shi, Rada Mihalcea, Mingjun Tian," Cross Language Text Classification by Model Translation and Semi-Supervised Learning" Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1057–1067, MIT, Massachusetts, USA, 9-11 October 2010. c 2010 Association for Computational Linguistics.

[3] Sergio Di Martino, Filomena Ferrucci, A Genetic Algorithm to Configure Support Vector Machines for Predicting Fault-Prone Components D. Caivano et al. (Eds.): PROFES 2011, LNCS 6759, pp. 247–261, 2011.© Springer-Verlag Berlin Heidelberg 2011.

[4] Jianguo Zhou Credit Risk Assessment Using Rough Set Theory and GA-Based SVM Grid and Pervasive Computing Workshops, 2008. GPC Workshops '08. The 3rd International Conference.

[5] Wei Xu , A Support Vector Machine Based Method for Credit Risk Assessment e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference on.

[6] Alec Go, Twitter Sentiment Classification using Distant Supervision, The list of keywords is linked oÆ of http://twitrratr.com/. We have no association with Twittratr.

[7] Bernard J. Jansen CHI 2009, Micro-blogging as Online Word of Mouth Branding April 4–9, 2009, Boston, Massachusetts, USA. ACM 978-1-60558-247-4/09/04.
[8] Jae H. Min, Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters at ELESIVR .
[9] Samanta, B. (2004), "Gear fault detection using artificial neural networks and supportvector machines with genetic algorithms", Mechanical Systems and Signal Processing, 18(3), 625-644.
[10]G. Mishne. Experiments with mood classi¯cation in blog posts. In 1st Workshop on Stylistic Analysis of Text For Information Access, 2005.
[11]B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1{135, 2008.
[12]Zhang, X.R., and Liu, F.: 'A patten classification method based on GA and SVM', 2002 6th International Conference on Signal Processing Proceedings, Vols I and Ii, 2002, pp. 110-113
[13] Liu, J.J., Cutler, G., Li, W.X., Pan, Z., Peng, S.H., Hoey, T., Chen, L.B., and Ling, X.F.B.: 'Multiclass cancer classification and biomarker discovery using GA-based algorithms', Bioinformatics, 2005, 21, (11), pp. 2691-2697