ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2016.0170

AN IMPLEMENTATION OF EIS-SVM CLASSIFIER USING RESEARCH ARTICLES FOR TEXT CLASSIFICATION

B. Ramesh[1] and J.G.R. Sathiaseelan[2]
[1,2]*Department of Computer Science, Bishop Heber College, India*
E-mail: [1]ram.73110@gmail.com, [2]jgrsathiaseelan@gmail.com

Abstract

Automatic text classification is a prominent research topic in text mining. The text pre-processing is a major role in text classifier. The efficiency of pre-processing techniques is increasing the performance of text classifier. In this paper, we are implementing ECAS stemmer, Efficient Instance Selection and Pre-computed Kernel Support Vector Machine for text classification using recent research articles. We are using better pre-processing techniques such as ECAS stemmer to find root word, Efficient Instance Selection for dimensionality reduction of text data and Pre-computed Kernel Support Vector Machine for classification of selected instances. In this experiments were performed on 750 research articles with three classes such as engineering article, medical articles and educational articles. The EIS-SVM classifier provides better performance in real-time research articles classification.

Keywords:

Automatic Text Classification, ECAS Stemmer, Efficient Instance Selection and Pre-computed Kernel Support Vector Machine

1. INTRODUCTION

Nowadays, the growth of IT everything should be digitalized. In this reason of, the usage of text information is very huge. Text classification is an innovative research topic in text mining. Automatic text classifier is very useful to organize and retrieve the text documents. Machine learning techniques are very useful to develop text classifier with better efficiency. Support vector machine is popular technique in supervised machine learning techniques. Text mining techniques are very useful to analyze, classify and retrieve necessary information from huge text repository. Text mining[1, 2] techniques are successfully applied to several fields such as clinical text, newspaper, educational text, web text, information retrieval and so on. Text classification process mainly divided into two category such as text pre-processing and classification task. The text pre-processing is used to dimensionality reduction for increase the efficiency of the classifier. In text pre-processing, ECAS stemmer was used to find root word and Efficient Instance Selection (EIS) for reduce dimensionality of data set.

Bag of words approach is used to transform the document into bag of concepts. Levent et al. [3] discussed feature coverage policies for feature selection to text classification. They are analyzed two stage feature selection mainly focused on pruning and keyword selection by varying parameters. Pui cheong et al. [4] discussed feature selection strategies of text classification. They are used scoring, ranking and selection method. Feature selection methods are used to extracting features. Alper et al. [5] introduced a Distinguishing Feature Selector (DFS) method for text classification. DFS is compared with various filtering methods. Also, the performance of DFS is well suited for text classification. García et al. [6] presented detail experimental study which is used to different machine learning techniques such as Naive Bayes, KNN, SVM and Rocchio classification. Finally, they are conclude support vector machine provide better classification accuracy compare to others. In this study, different criteria are followed to evaluate the text classifier performance such as speed, scalability, accuracy, time complexity and flexibility. Bashar et al. [7] proposed novel concept based weighting scheme to index the text with the flavor of indexing scheme. This technique is uses N-gram approach to extract the words. Ya Gao, et al. [8] used kernel and non-linear kernel for text classification with support vector machine. The aim of stemming is, to find the root word, which is useful to dimensionality reduction of dataset. Ruba et al. [9] described detail study on recent stemming techniques. Chaitali et al. [10] conducted experimental study using porter stemming with support vector machine for text classification.

Technical systematic reviews answer certain questions within a very specific area of expertise by selecting and analyzing the recent pertinent literature. In several research domains, huge amount of research papers are published. Hence, the classification of research articles is a challenging issue. The remaining of these papers is organized as follows. Section 2 discuss background study followed by implementation is elaborately explained in section 3. Section 4 is analysis of results and followed by conclusion and future directions is section 5.

2. LITERATURE REVIEW

Automatic text classification has recently observed a growing interest, due to this availability of documents in digital form is increased. Stemming is a pre-processing step in text mining applications as well as a very common requirement of Natural Language processing functions. In fact, it is very important in most of the information retrieval systems. The main purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. Ruba et al. [11] proposed APOST (Advanced Porter Stemming) stemmer for text classification. APOST stemmer is finds and extracts accurate root words in maximum occurrences. Agbele et al. [12] proposed Context Aware Stemming (CAS) algorithm. CAS is improved version of porter stemming algorithm which is provided better stemmed words.

Jingnian Chen et al. [13] offered Multi Class Instance Selection (MCIS) method for choose instances nearest boundary. MCIS has been used to speed up support vector machine. On the other hand, using this method selection of instances will be less. Chih-Fong Tsai et al. [14] introduced Support Vector Oriented Instance Selection (SVOIS) for instance selection to text classification. SVOIS provided best performance compared with

several states of art instance selection algorithms such as ENN, IB3, ICF and DROP3. Hence, this method is not enhanced to non-linear regression plane and large scale experiments. Ramesh et al. [15] discussed several kernel functions with support vector machine using micro array data sets. The pre-computed kernel support vector machine is providing better results comparing to other kernel functions.

3. METHODOLOGY

An effective text mining is predicted with suitable pre-processing techniques. Text pre-processing is necessary to dimensionality reduction of dataset. The text pre-processing consists several steps such as tokenization, stop word removal and stemming. Tokenization is a process, which is a document converted into terms and words. After tokenization, stop word removal is removing the meaningless words from a document such as articles, prepositions and conjunctions. The several techniques are available for tokenization and stop word removal.

Stemming is a process to find the root word. They are several techniques are available for stemming such as lovins, porter and YASS stemmer. The existing stemmer includes several drawbacks. The ECAS stemmer shown better efficiency to find the root word. The proposed classifier uses ECAS stemmer for stemming process. ECAS stemmer consists less morphological rules and using Extract From Table (EFT) method. Due to this reason, easily identify root word from given word with less time consumption.

Term Frequency (TF) and Inverse Document Frequency (IDF) is numeric statistical value which is very useful to importance of word in a document or a collection and corpus. It is popular word weighting technique, which is very useful to search engines and document ranking system. Term frequency and inverse document frequency is very useful to pre-processing task in text mining. Term frequency is proportionate between number of times occurred in a word and total number of words in a document. Inverse document frequency is logarithmic value of proportion between to total number of documents and number of document occurred in a particular word.

The terms are occur computer network, data mining, computing, engineering, electric, electronic and so on the research articles are classified to engineering articles. And the terms are occurring cancer, diabetics, bio-medical, genetics, medicine and so on the research articles are classified to medical articles.

$$\text{TF}=\frac{\text{Number of times occured in word}}{\text{Total number of words in a document}} \quad (1)$$

$$\text{IDF}=\log_e \frac{\text{Number of times occured in word}}{\text{Total number of words in a document}} \quad (2)$$

The major role of instance selection techniques is to reduce the complexity of training SVM. Recently proposed multi class instance selection is used to Gaussian kernel. The performance of MCIS is not sufficient. Hence, the advanced MCIS is used pre-computed kernel support vector machine and its performance is better. The main objective of the whole process the text document convert into sparse vector.

The pre-processed document (selected instances) posted to support vector machine for research articles classification. The proposed classifier has been tested with standard support vector machine formulation as Eq.(3). Given a training set of instance-label pairs (x_i, y_i) $i = 1,2,\ldots,n$; where $x_i \in \Re^n$ and $y_i \in \{1,-1\}$, the following standard support vector machine with pre-computed kernel optimization formulation in Eq.(3).

$$\min_{w,b,\varepsilon} \frac{1}{2} y\|w\|^2 + C\sum\nolimits_i \varepsilon_i \quad (3)$$

Show that, $y_i((w, \phi(x_i)) + b) \geq 1 - \varepsilon_i$

$\varepsilon_i \geq 0$.

Selected vectors (x_i, y_i) $i = 1,2,\ldots,n$; are mapped into a higher dimensional space by the function ϕ. PKSVM finds a non-linear separating optimal hyper plane with the maximal margin in this higher dimensional space.

4. EXPERIMENTAL ANALYSIS

In this work, we collect technical papers from annual conference proceedings to determine the categories of the paper. The total number of papers in the collection is 750, while the total number of classified categories is two such as engineering and medical articles. The articles collection consists 586 engineering articles and 164 medical articles.

All the experiments are performed with MATLAB with edited Libstring interface. They are two types of classification are followed such as only title and title with keyword. The below results are found when the range is 0.082 and margin value is 0.039. The Table.1 shows the performance of the classifier.

Table.1. Classifier Performance

Type	TP	FP	FN	TN
Title	562	115	27	46
Title with keyword	525	145	41	39

The Table.2 shows the result analysis of the classifier. The classifier is classified into two categories such as engineering articles and medical articles. The classifier classified out of 750 research articles are 562 and 525 is title and title with keyword respectively correctly classified.

Table.2. Result Analysis of the Classifier

Category	Classification Accuracy (%)	Selected Support Vectors	Execution Time (sec)
RAT	85	56	0.46
RAT with keyword	74.9	49	0.58

The Fig.2. shows the comparison of classifier for two types of classification. According to the representation of Fig.2, the classifier correct prediction is 74.9% in title and 70% in title with keyword. Hence, the classifier prediction is true in maximum occurrences. In this reason of, the classifier performance is better in real time datasets.

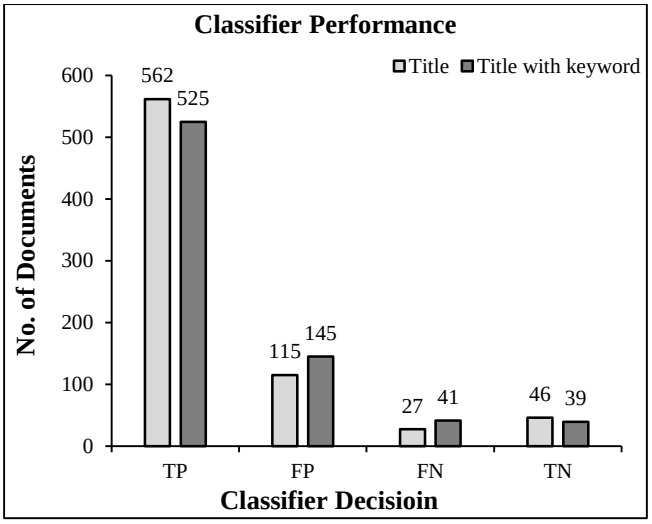

Fig.2. Comparison of classification type

The Fig.3 show the representation of selected instances and Fig.4 shows classifier classification representation. Depending upon the margin value the classifier is classified into two categories.

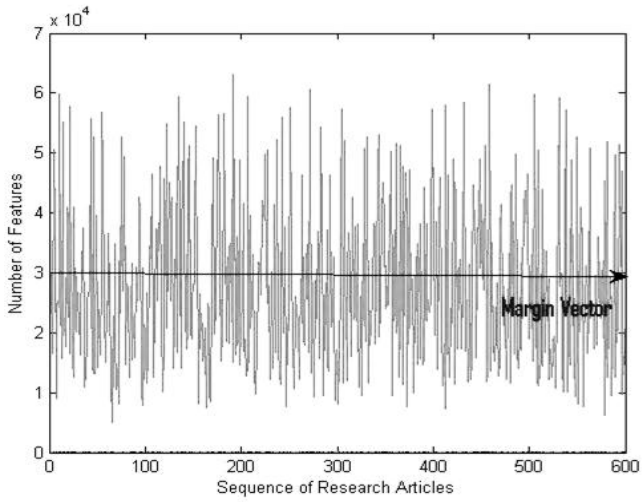

Fig.3. Representation of selected instances

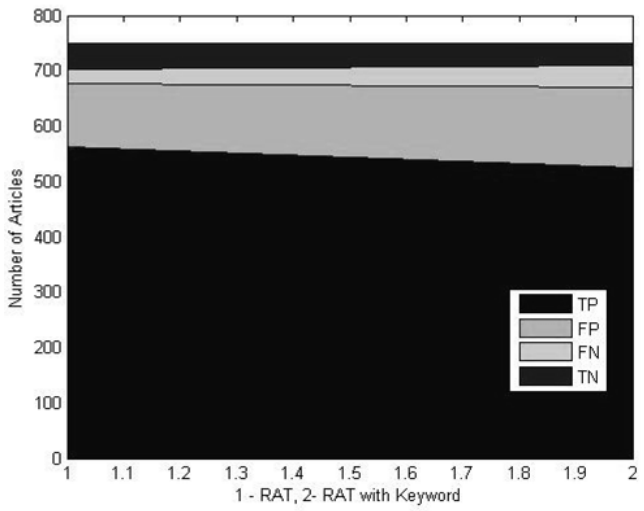

Fig.4. Representation of classification

5. CONCLUSION

Text mining is powerful method to extract knowledge from huge text repository. Automatic text classification is significant research topic in text mining. Stemming is to find root word and instance selection is reduce the dimensionality reduction of dataset. In this work, we implement ECAS stemmer and Advanced Multi Class Instance with pre-computed kernel based support vector machine using real-time research articles dataset. The classifier providing better performance in real-time research articles classification. The classifier shown 75% accuracy in research article classification. Hence, ECAS stemmer and EIS shown better result in real-time research articles classification. In future, the classifier is extended more than two classes.

REFERENCES

[1] Ning Zhonng, Yuefeng Li and Sheng Tang Wu, "Effective Pattern Discovery for Text Mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 1, pp. 30-44, 2012.

[2] Hussein Hashimi, Alaaeldin Hafez and Hassan Mathkour, "Selection Criteria for Text Mining Approaches", *Computers in Human Behavior*, Vol. 51, pp. 729-733, 2015.

[3] Levent Ozgur and Tunga Gungor, "Two-Stage Feature Selection for Text Classification", *Proceedings of 30th International Symposium on Computer and Information Sciences*, pp. 329-337, 2015.

[4] Pui Cheong Gabriel Fung, Fred Morstatter, and Huan Liu, "Feature Selection Strategy in Text Classification", *Advances in Knowledge Discovery and Data Mining*, Vol. 6634, pp. 26-37, 2011.

[5] Alper Kursat Uysal and Serkan Gunal, "A Novel Probabilistic Feature Selection Method for Text Classification", *Knowledge Based Systems*, Vol. 36, pp. 226-235, 2012.

[6] J.J. Garcia Adeva, J.M. Pikatza Atxa, M. Ubeda Carrillo and E. Ansuategi Zengotitabengoa, "Automatic Text Classification to Support Systematic Reviews in Medicine", *Expert Systems with Applications*, Vol. 41, No. 4, pp. 1498-1508, 2014.

[7] Bashar Tahayna, Ramesh Kumar Ayyasamy, Saadat Alhashmi and Siew Eu Gene, "A Novel Weighting Scheme for Efficient Document Indexing and Classification", *Proceedings of International Symposium on Information Technology*, Vol. 2, pp. 783-788, 2010.

[8] Ya Gao and Shiliang Sun, "An Empirical Evaluation of Linear and Nonlinear Kernels for Text Classification Using Support Vector Machines", *Proceedings of Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, Vol. 4, pp. 1502-1505, 2010.

[9] S.P. Ruba Rani, B. Ramesh, M. Anusha and J.G.R. Sathiaseelan, "Evaluation of Stemming Techniques for Text Classification", *International Journal of Computer Science and Mobile Computing*, Vol. 4, No. 3, pp. 165-171, 2015.

[10] Chaitali G. Patil and Sandip S. Patil, "Use of Porter Stemming Algorithm And SVM for Emotion Extraction from News Headlines", *International Journal of Electronics, Communication and Soft Computing and Engineering*, Vol. 2, No. 7, pp. 9-13, 2013.

[11] S.P. Ruba Rani, B. Ramesh and J.G.R. Sathiaseelan, "An Increasing Efficiency of Pre-processing using APOST Stemmer Algorithm for Information Retrieval", *International Journal of Emerging Technologies and Innovative Research*, Vol. 2, No. 7, pp. 3219-3223, 2015.

[12] K.K. Agbele, A.O. Adesina, N.A. Azeez and A.P. Abidoye, "Context-Aware Stemming Algorithm for Semantically Related Root Words", *African Journal of Computing and ICT*, Vol. 5, No. 4, pp. 33-41, 2012.

[13] Jingnian Chen, Caiming Zhang, Xiaoping Xue and Cheng Lin Liu, "Fast Instance Selection for Speeding Up Support Vector Machines", *Knowledge-Based Systems*, Vol. 45, pp. 1-7, 2013.

[14] Chih-Fong Tsai and Che-Wei Chang. "SVOIS: Support Vector Oriented Instance Selection for Text Classification", *Information Systems*, Vol. 38, No. 8, pp. 1070-1083, 2013.

[15] B. Ramesh and J.G.R. Sathiaseelan, "Support Vector Machine using Efficient Instant Selection for Micro Array Data Sets", *Proceedings of IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-4, 2014.