An Analysis of Instance Selection Algorithms using Support Vector Machine for Text Classification

B.Ramesh
Ph.D. Research Scholar
Department of Computer Science
Bishop Heber College, Tiruchirappalli, TN, India. E-mail: ram.73110@gmail.com.

Dr.J.G.R.Sathiaseelan
Head, Department of Computer Science
Bishop Heber College, Tiruchirappalli, Tamil Nadu, India. E-mail: jgrsathiaseelan@gmail.com.

Abstract: Automatic text classification is a popular research topic in text mining. Automatic text classification is an eminent field of research in text mining, which is tries to automatically classify the text documents into pre-specified categories. Text mining involves several pre-processing and classification techniques. In this paper, we have analysed several feature selection methods with support vector machine for text classification. Several instance selection methods analysed with different aspects such as accuracy, processing time, reduction rate and selected support vectors. This analysis is mainly focused to techtc-100 dataset with several feature selection strategies. Support Vector Oriented Instance Selection is shown better efficiency in all aspects. This study is very useful to novice text mining researchers.

Keywords: Text Mining, Text Classification, Pre-processing methods, Support Vector Machine.

1. INTRODUCTION

Due to vast usage of text data, text mining has become a challenging research area in data mining. Hence, It’s organizes and classifies the text document which is considered as a challenging issue. 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. Classification of the text documents or information is stated as text classification. In text classification, several steps are involved such as feature selection, feature extraction and classification methods. Text mining most commonly used to classification of text or information, clustering of text, summarization of text information. Text classification is one of the significant part in text mining. Text classification [3] is very useful to blogs, social networks, e-commerce sites and so on. The difficulties of supervised text classification is elaborately discussed in literature and metrics. However, large scale text labeling is a challenging issue. The main objective of text classification is automatically texts documents can be classified into inter related categories. Most commonly used feature selection steps are tokenization, stop word removal, lowercase conversion and stemming.

Classification is the separation or ordering of objects into class. Nowadays, text classification is trendy research topic in the area of information retrieval. Henceforth, there is a huge hike in the field of digital data storage. In this situation, the process of organizing and searching of E- documents is very hard. Therefore, text classification is gained its importance as tools to search, organize and store such huge volume of data. The objective of text classification is to classify the document into pre-specified categories.

Machine learning techniques are very useful to classify the text documents. Nowadays, several text classification techniques are available namely Naïve Bayes, k-Nearest Neighbor, Decision Tree, SVM and so on. Among these techniques Support Vector Machine (SVM) proposed by Vapnik et al. and it is a popular classifier in supervised machine learning techniques. SVM is providing high efficiency. SVM is providing efficient result in text classification. Hence, SVM is included several limitations. SVM is computationally expensive when handling large data set.

Feature selection is one of the critical stage in text mining and knowledge discovery which is used to data reduction. Feature selection in Text is commonly used to tokenization, stop word removal and stemming. Text pre-processing commonly used to TF, IDF. The feature extraction stage generally utilizes the vector space model that makes use of the bag-of-words approach. The feature selection [4] stage, most of the time, employs the filter methods such as document frequency, mutual information, information gain, chi-square, Gini index, wrapper method and distinguishing feature selector. García et al. [5] conducted an experimental study uses different machine learning algorithms such as naive bayes, KNN, SVM and Rocchio classification. Finally, they concluded that support vector machine provided better classification accuracy comparing to others. In this study, different criteria are followed to evaluate the text performance such as speed, scalability, accuracy, time complexity and flexibility.

Instance selection is an another solution of feature selection in text mining which is reducing the text size by filtering out the noisy and redundancy data from given data.
set. Instance selection is used to increasing the speed of SVM. Several studies introduced multiple instance selection methods for increasing mining quality in the literature. Particularly, N Jankowski et al. [6] conducted a comparative study of different instance selection algorithms such as noisy filters, condensation algorithms and prototype selection. Also, they are discussed complexity of different instance selection algorithms. Deqing et al. [7] proposed centroid classifier for text classification and it’s based on the border instance approach. Centroid classifier is consuming less time comparing to SVM. But, it’s not deal with multi label classification.

II. BACKGROUND STUDY

Feature Selection

An effective text mining is predicted with sophisticated pre-processing techniques. Text pre-processing is one of most important steps in Text mining. Text pre-processing is necessary to filter the information using relevant condition depends upon the text. Pre-processing consists of several stages such as data cleaning, data integration, description and summarization, attribute transformation and data reduction. Natural Language Processing is one of task in frequently used in text pre-processing. After, pre-processing the information posted to classification methods. In text mining, text pre-processing usually follow the steps such as tokenization, stop word removal, lower case conversion and stemming.

Tokenization is the task to segment the sentence into words, phrases or other meaningful parts which are stated as tokens. Stop word is a process for removing irrelevant words such as articles, preposition, and conjunction and so on. Stemming is used to find the root of derived word in a document. S.P.Ruba et al. [8] detailed discuss several stemming algorithms and issues of stemming algorithms. Therefore, pre-processing task is commonly used to extract the features and feature selection.

Concept of instance selection

The major objective of instance selection algorithm is to increase efficiency of NN classifier. The NN algorithm is one of the most popular algorithms in machine learning. This algorithm calculates Euclidean distance between an instance to be classified and each training-neighboring instances. Nowadays, several instance selection algorithms are proposed by several authors. Edited Nearest Neighbor (ENN) algorithm was created in 1972 by Wilson. There are Edited Nearest Neighbor, RENN is revised version of ENN. Iterative Case Filtering (ICF) introduced by Brighton and Mellish in 2002. Decremental Reduction Optimization Procedure 1 was proposed by Wilson and Martinez in 2000. DROP2, DROP3 are updated version of DROP1. IB3 is an incremental algorithm and it’s described by Aha et al. Instance selection for text classification can be defined as follows. Consider a k class data set S, composed of training set TR, and test set TE. Let \( x_i \) be the \( t \)th instance in TR, where \( x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}) \) which contain \( n \) different features. Let \( SSC TR \) be the sub set of selected instances, which is resulted from instance selection techniques.

Jingnian Chen et al. [9] divided Multi Class Instance Selection (MCIS) for speed up SVM. Analyzing performance of MCIS with several data sets including gene expression and text data sets. B.Ramesh et al. [10] proposed Improved Multi Class Instance Selection (IMCIS) for efficient SVM and its performance is assessed with different gene expression data sets. Four popular instance selection algorithms, ENN, IB3, ICF, SVOIS are discussed below. Rui et al. [11] discussed several multi instance learning algorithms with different datasets. C.F.Tsai et al [12] proposed Support Vector Oriented Instance Selection (SVOIS) for text classification. The performance of SVOIS is analyzed with SVM and K-NN classifier, which is produced more accuracy in text classification. However, SVOIS is not extended to non-linear criteria. Instance Selection algorithms are earn special interest of its adaptability. Chih-Fong Tsai et al. [13] suggested Efficient Genetic Algorithm (EGA) for text classification. The EGA is borrowed idea from genetic and instance selection algorithms. Special features of EGA are reasonable convergence, non-linear adaptability and great migration. The performance of EGA is assessed with techtc-100 dataset. Chih fong et al. [14] proposed instance selection based Biological Genetic Algorithm (BGA) for text classification. However, the performance of BGA is not sufficient and it’s developed with lower cost.

Support Vector Machine

Text classification is a well-studied problem in supervised machine learning techniques. A several variety of algorithms have been proposed for text classification. The classical SVM is modified and combined with several algorithms. That is called Traditional SVM. Among these traditional SVM is providing high performance. However, SVM suffers several drawbacks namely high memory, time complexity of training stage in large scale classification task. B.Ramesh et al. [15] conducted extensive study of several kernel methods using SVM. Pre-computed kernel shown better performance comparing to other kernel methods.
III. ANALYSIS OF ALGORITHMS

Instance selection algorithms are increasing the efficiency of support vector machine. Support vector machine is a well suited technique to text classification. In this paper, we are theoretically analyzing performance of some instance selection algorithms such as SVOIS, EGA, BGA, IB3, DROP3, ICF and ENN. This analysis is used to tech-tc 100 dataset. They are several aspects considered for this analysis such as issues of algorithm, accuracy, time, reduction rate and selected instances. We are taken average results of classification accuracy, time consumption, selected instances and reduction rate of Tech-Tc 100 dataset.

The performance evaluation of nonlinear SVOIS is used to TechTc-100 data set. It includes 100 different two class datasets which is included largest and smallest datasets contains 165 and 125 documents respectively. In addition, it ranges from pairs of easy categories, such as Games, Video Games, Shooter and Recreation, Autos, Makes and Models, to medium difficulty ones with Arts, Music, Bands and Artists vs. Arts, Celebrities, to hard ones such as Regional, North America, United States, Virginia/ Richmond/ Business and Economy vs. Regional North America, United_States, Florida, Fort_Myers, Business_and_Economy. Specifically, the feature dimensionality of each pair ranges from 12,813 to 29,260. Table I shows the issues of instance selection algorithms.

A. Issues

Nowadays, instance selection algorithms are performed very well in text classification with support vector machine. However, instance selection algorithms face some drawbacks. Table-I shows the issues of instance selection methods.

<table>
<thead>
<tr>
<th>ALGORITHMS</th>
<th>ISSUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVOIS</td>
<td>Less reduction rate and it’s not extended to non-linear case.</td>
</tr>
<tr>
<td>EGA</td>
<td>Scalability.</td>
</tr>
<tr>
<td>BGA</td>
<td>Number of selected instances is less.</td>
</tr>
<tr>
<td>IB3</td>
<td>Learning performance is highly sensitive to the number of irrelevant attributes.</td>
</tr>
<tr>
<td>DROP3</td>
<td>More time consumption.</td>
</tr>
<tr>
<td>ICF</td>
<td>Providing less accuracy.</td>
</tr>
<tr>
<td>ENN</td>
<td>Reduction rate is minimum.</td>
</tr>
</tbody>
</table>

B. Accuracy and Reduction Rate

Table II shows accuracy and reduction rate of several instance selection algorithms. SVOIS is provided highest classification accuracy is 77.26% comparing to another all algorithms. BGA is provided better reduction rate is 84.25% comparing to another all algorithms. Hence, it’s provided less classification accuracy. SVOIS is provided less reduction rate. Fig.1 explains comparison of average accuracy and reduction rate.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy (%)</th>
<th>Reduction Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVOIS</td>
<td>77.26</td>
<td>2.6</td>
</tr>
<tr>
<td>EGA</td>
<td>66.25</td>
<td>50</td>
</tr>
<tr>
<td>BGA</td>
<td>76.82</td>
<td>84.25</td>
</tr>
<tr>
<td>IB3</td>
<td>73.18</td>
<td>50.17</td>
</tr>
<tr>
<td>DROP3</td>
<td>69.14</td>
<td>40.99</td>
</tr>
<tr>
<td>ICF</td>
<td>63.03</td>
<td>40.32</td>
</tr>
<tr>
<td>ENN</td>
<td>69.09</td>
<td>40.1</td>
</tr>
</tbody>
</table>

C. Selected Instances and Time Consumption

Table III shows selected instances and time consumption of all listed instance algorithms. SVOIS is selected more instances with less time consumption. SVOIS is selected highest 109 instances with less time in 2.43 seconds. DROP3 is consumed more time 155.5 seconds and selected less 66 instances. SVOIS performance is better to comparing another all instance selection algorithms. Fig.2 shows the comparison of average selected instances and time consumption of all listed instance selection algorithms.
IV. CONCLUSION

Automatic text classification has recently witnessed a booming interest research topic in text mining, due to the increased availability of documents in digital form and organizing the digital text documents. Support Vector Machine is most popular classifier in supervised machine learning techniques. SVM produced best results in text classification. Instance selection algorithms are increasing efficiency of SVM. Based on the above analysis SVOIS provided better classification accuracy comparing to another instance selection algorithms. SVOIS is selected more instances in 109 instances with 2.43 seconds. Also, SVOIS is consuming less time, more classification accuracy and maximum number of selected instances. In future, SVOIS performance is analyzed with huge datasets which is extended to non-linear case using advanced kernel methods.

REFERENCES


