Machine Learning Based Efficient Filtered Classifiers for Text Document with Unseen Test Data

War War Cho
Faculty of Information Science, University of Computer Studies (Maubin)
warwarcho@ucsmub.edu.mm

Yu Mon Win Myint
Information Technology Supporting and Maintenance, University of Computer Studies (Taunggyi)
yumonwinmyint247@gmail.com

Ei Ei Moe
Faculty of Computer Science, University of Computer Studies (Maubin)
eieimoe1989@gmail.com

Abstract

Text document classification methods have been rapidly implemented and developed in recent years. The more useful and understand classifiers are needed to get the high accuracy of classification result on the large amount of text document due to the increasing number of complex documents and text documents. The main expected goal of classification on text documents is to exactly categorize the text piece into the various predefined categories. Because of rapidly improvement of using information technology, the automatically classification task on the online text document is become a main role for the discovery and collection information among the large amount documents. Moreover, the various methods of classification and algorithms for classification are improved by previous researchers however these methods and algorithms needed to filter the documents data before classification of these documents. Keeping these requirements, this system implements the idea problem of classifiers on unseen data classification for text documents that are not homogenous. In this proposed system, the filtering step is used before classification step on the text documents to get the better performance of text documents classification. C4.5, Support Vector Machine and Naïve Bayes from the machine learning approaches are used in the classification step of the system, the text documents the structure of the filter is based on dataset (training and testing) then these dataset are processed by the filters without changing behavior of the document structure. Data Discretization is a popular and useful step in the process, since it is easier for classifiers with discrete attributes value rather than continuous attributes value. In the preprocessing step of this system, data discretization is used to discretize the value of numerical attributes in the text document dataset into nominal attributes. Newsgroup data are employed to test the unseen test data and the performance efficiency of classification used filters are analyzed. As the experimental result, the detail results of various classification methods and the classification accuracy result of these classifier are analyzed and describe in this system.

Keywords: C4.5, Naïve Bayes, Support Vector Machine, Discretization, Filtered Classifier;

1. Introduction

Nowadays, the evaluation of using internet has led to emerge the useful and well-developed tool as an important role for people in retrieving significant information from the large amount of Web documents [1]. Web data documents are collected from different sources and locations that are simple related but various categories [2]. However, these documents may appear in some portion of the data which are rarely seen in the data. Applications of text classification include categorizing news topics and contents into topics, collecting Web pages into predefined various categories, predicting user queries, routing support ticket and analyzing customer feedback. The classification of text also known as text tagging or text categorization is the application of categorizing text into predefined categories groups. It is becoming a very important part of business as it allows to easily get insights from data and automate business preprocess. In changes of situation, training data and text data of non-homogeneous are not represent of the distribution under these classification methods.

According to the widespread use of internet, large amount of documents are needed to categorize and information can be extracted from these documents [3]. Through the training data to measure the efficiency of the model, the test data in documents are used. However, these model are a representative up to some part of the model for future [4]. In some literatures, the training data are collected from one place which does not represent the distribution of data in various places to create model of classification for the data of customers into predefined various categories [5]. The important features are involved in testing data which are not detected or infrequently observed in trained dataset but these features are needed as the vital role in the model of classification. Many researchers developed various proposed algorithms with classification methods but these approaches are needed to filter the data before classification.

In general most of the discretization process lead to a loss of information and can't reduce values of discrete attributes. To eliminate noise data from text document, reduce the incorrect data and make for reduction on dimension, this kind of discretization filters are used. When the system use classification filters for these documents, the performance of classifiers are increased and provide the better performance [6]. The main idea in pruning of a classification decision tree is to remove the building a tree that simple fit to unseen text data documents [7]. In this system, discretization method is used as the filtered classifier before text classification on unseen test data. For the classification of this system, C4.5, Naïve Bayes and Support Vector Machine (SVM) from machine learning approaches are used to get the better performance for text documentation.

2. Related Works

In 2003, Ratnaparkhi and Adward developed a statistical approach that trains on noted corpus with part of speech tags to easily extract unseen test data [8]. The algorithm for searching feature set is used to test on training sets and to evaluate the test result on unseen test datasets. The emotional recognition system from speech signals was proposed by Sherif and Yacoub, in 2003.

In their proposed system, classifiers for emotion was used to test the unseen speech form unseen speech signal to ensure independent speakers [9]. Bayesian online perceptron approach and Gaussian methods are developed with information gain and adaptive batch filtering technique for classification of text by Kian Ming and Adam Chai et al. [10]. In 2008, the classification methodology for Webpage using classification methods was proposed combining with Web-content and structure analysis by Michael Chau et al. [11].

The document classification approach was implemented by Dino Isa et al. using powers of the self-organizing map (MOP) utilized as multi-dimensional unsupervised classifier and Naïve Bayes Classifier [12]. The distinguishing was proposed by Alper Kursat Uysal et al. using probabilistic feature selection [13]. For text classification, Kashif aved et al. suggested a two-stage based feature selection approach using a Markov Blanket Filter [14]. Guansong Pang et al. implemented the classifier with cluster centroid for categorization on text by developing the clustering algorithms and k-Nearest Neighbor (kNN) classifier with Rocchio classifier [15].

A framework for text classification are implemented by Bruno Trstenjak et al. using k-NN classifier with TF-IDF method [16]. Sebastian Schmidt et al. suggested a classification approach for preprocessing step in search engines based on domain specific [17]. Deep Learning based model for text documents are developed by Navoni Majumder et al. using deep convolutional neural network with assigning of the documents into the inside layer [18]. However, these approaches do not improve accuracy of classification from unseen data pass through the test dataset. For keeping these restrictions, this proposed system attempted a filtered classifier in the classification of unseen test data on text documents.

3. Proposed System Design

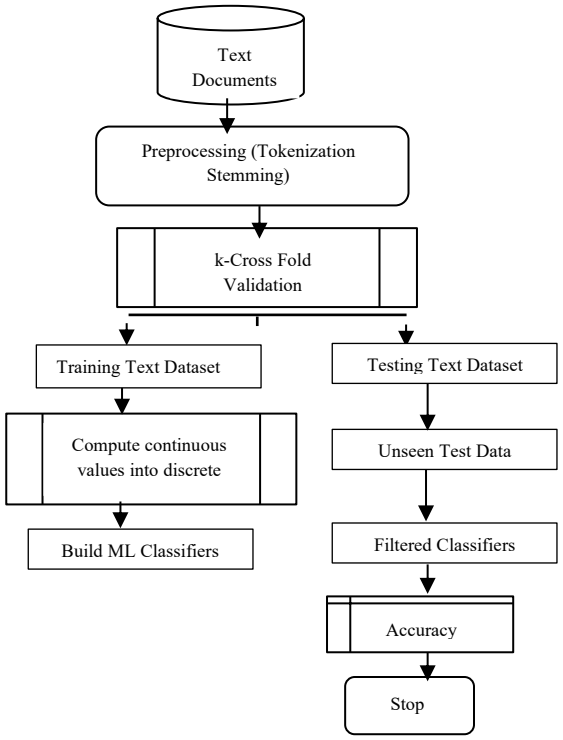

Figure 1. Proposed System design

The flow of the proposed system design is shown in Figure 1. In the proposed system approach, effective classifiers such as C4.5, Naïve Bayes and SVM are used to classify the text documents which has accepted through an arbitrary filter and then analyzed the results of classification. Moreover, to classify the text documents, the classifiers' structure of the filter are based on training and testing dataset that are processed by the filter without changing behavior of structural dataset. Filter for text document classification, the text attributes set is filtered with a filter method to produce the most effective subset before creating the model.

Additionally, this proposed system used Discretization method for a filter, which discretize a range of numerical attributes in the text document data set into nominal attributes. The popular classifiers are used to classify for text documents classification and to

reduce the classifier accuracy on unseen test data instances in the text document.

4. Experimental Results

In this system, the experimental analysis results are evaluated using the effectiveness of filtered on large amount of text document datasets. This study is improved by applying the proposed methods for several categories of text applications, and this approach includes the discretization method as a filter for classification, which discretize numerical attributes in the text document dataset into nominal attributes and for classification with better performance.

4.1. Datasets of the system

To evaluate the performance result of proposed approach, Newsgroups dataset are collected and used as the dataset. These Newsgroups dataset is collected around 10000 documents dataset, which are spilt into across 20 different newsgroups categories. This process becomes the most popular dataset in the text document for real-world applications of machine learning approaches, as classification of documents and clustering for text documents.

4.2. Performance Analysis

In order to evaluate the accuracy of classifier based text document classification method proposed in the study, the value of precision, recall, accuracy of the system and f-measure values were calculated mostly used in the many literature. In most calculation of performance in the classification system, the measurement parameters that shown in table 1 as confusion matrix are used in the data extraction step.

Table 1. Confusion Matrix

	Base Prediction of Correct	Base Prediction of Incorrect
Algorithm Prediction of Correct	True Positive (TP)	False Positive(FP)
Algorithm Prediction of Incorrect	False Negative(FN)	True Negative(TN)

As shown in Table 2, the confusion matrix provides the meaning of TP, FP, TN and FN. These values are used to calculate like precision, recall, accuracy and F-Measure. Accuracy of the system is computed as in the following equation (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

This measure shows the ability of a classification technique to differentiate a sentiment with non-sentiment in given dataset. It uses TP, FP, TN and FN provided in the confusion matrix. Precision is the measure that used to know the ability of classification methods to know the fraction of correctly classified document among all the considered documents. The Precision is computed as in equation (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

In the same fashion, Recall is the measure to know fraction of correctly classified instances among all the instances that are supposed to be in correctly labelled instances. It is measured as in equation (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

There is another measure which is based on the precision and recall. It is harmonic mean of these two measures. It is called F-Measure which is computed as in equation (4).

$$FMeasure = 2 \times ((Precision \times Recall)/(Precision + Recall)) \quad (4)$$

In statistics, the mean squared error (MSE) of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. The value of Root Mean Square Error can be calculated by using the following two equations.

$$Mean\ Square\ Error\ (MES) = \sum \frac{(f(x_i)-y_i)^2}{N} \quad (5)$$

$$Root\ Mean\ Square\ Error = \sqrt{MES} \quad (6)$$

Here, TP represents the correct document classifier type. TN describes the normal document data classified correctly as normal, a FN denotes where a document was classified as normal dataflow and a FP shows that a normal case was classified as an unseen data in test dataset. This system used the F-measure to calculate the harmonic grouping of precision & recall values for text documents. Accuracy metric is a root mean square error (RMSE) used to estimate the predictive accuracy of a classifier model.

The classification accuracy, F-measure and RMSE values are analyzed without and with Filtered Classifier on datasets shown in the Table 2 and Table 3. Proposed system's approaches provide the better performance on unseen data in test datasets where document data distribution is not homogeneous.

Table 2. Analysis of Accuracy Results before using Filtered Classifiers

Classifiers	F-Measure	RMS	Accuracy
C 4.5	0.87	0.40	87.8
Naïve Bayes	0.80	0.43	80.9
SVM	0.92	0.31	92.3

In Table 3 shows the accuracy results of the system with the filtered classifiers. In this analysis, the improvement of accuracy results can be seen with the effectiveness of filtering.

Table 3. Analysis of Accuracy Results after using Filtered Classifiers

Classifiers	F-Measure	RMS	Accuracy
C 4.5	0.89	0.38	88.5
Naïve Bayes	0.83	0.40	82.7
SVM	0.96	0.28	95.8

In table 4, the comparison result on the accuracy of before state and after state of using filtered classifiers. The accuracy of using filtered classifier is slightly increased according to the experimental analysis. The represented bar chart for analysis result on the different classifiers based on without filter and using filter approach are shown in Figure 2.

Table 4. Analysis of Accuracy Results of classifiers

Classifiers	With Filter	Without Filter
C 4.5	88.5	87.8
Naïve Bayes	82.7	80.9
SVM	95.8	92.3

According to the experimental result the accuracy result of SVM classifier provide the best result comparing with other two classifier. In Figure 2, the blue bar represents the accuracy result with filter classifier and the green bar shows the accuracy results without filter classifier. All accuracy results of approach with filtered classifier is better than the approach without filtered classifier.

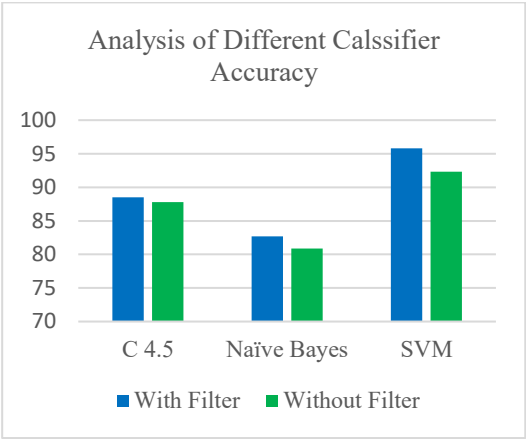

Figure 2. Analysis of Accuracy Results

5. Conclusion

In this paper, we presented Filtered Classifier on text documents, which classify the documents using the C4.5, Naïve Bayes and Support Vector Machine classifiers that are based on trained text document and testing data sets which are processed by the filter without changing the structural behavior. In addition, this system used discretization method as a preprocessing that transform the numerical attributes into nominal attributes of the text document. This system used C4.5 decision tree, Naïve Bayes and SVM classifiers for the classification of text documents. All the computations performed without and with filter on large amount of Newsgroups datasets. As the experimental result of this system, the classification text document approach with filer before classification provided the better performance result than the without using filtering step. As a conclusion, this proposed system approach based Filtered Classifier achieved the better performance than the normal other classification approach.

References

[1] Y. H. Li and et al., "*Text Document Clasification*," CJ, vol.41, Issue.8, pp.567-546, 1998.

[2] T. Taskar and et al., "*Learning on the test data with Leveraging unseen features*," TICML, pp. 744-751, 2003.

[3] G. Aydin and et al., "*Document Classification Using Distributed ML*,": 1802.03597, 2018.

[4] J.Kleinberg, "*Bursty and hierarchical structure in streams*," 8th ACM SIGKDD, pp. 91-101, 2002.

[5] B. Liu, *"A Survey of Opinion Mining and Sentiment Analysis,"* Mining Text Data Book, Springer, 2011.

[6] S. Chakrabarti, et al., *"EnhancedHypertext Categorization using Hyperlinks,"* ACM Conference, 1998.

[7] C.Huang, et al., "*An assessment of support vector machines for land cover classification,"* IJRS, Issue. 23, no. 4, pp. 725-749, 2002.

[8] R. Adwait, *"A maximum entropy model for part-of-speech tagging,"* Proceedings of the conference on NLP, vol. 1, pp. 133-142. 1996.

[9] M. Sherif, et al., "*Recognition of emotions in interactive voice response systems,"* INTERSPEECH, 2003.

[10] K. M. A. Chai, and et al., "*Bayesian online classifiers for text classification and filtering,"* ACM SIGIR Conference, pp. 97-104, ACM, 2002.

[11] M. Chau, H. Chen, *"A machine learning approach to webpage filtering using content and structure analysis,"* DSS, vol. 44, no. 2, pp. 482-494, 2008.

[12] D. Isa, et al., *"Using the self-organizing map for clustering of text documents,"* ESA, vol. 36, no. 5, pp. 9584-9591, 2009.

[13] A.K.Uysal and S.Gunal, *"A novel probabilistic feature selection method for text classification,"* KBS, vol.36, pp. 226- 235, 2012.

[14] K. Javed, and et al., *"A two-stage Markov blanket based feature selection algorithm for text classification,"* Nero Computing, vol. 157, pp. 91-104, 2015.

[15] G. Pang, S. Jiang, *"A generalized cluster centroid based classifier for text categorization,"* IPM, vol. 49, no. 2, pp. 576-586, 2013.

[16] T. Bruno, and et al., *"KNN with TF-IDF based Framework for Text Categorization,"* PE, vol. 69, pp. 1356-1364, 2014.

[17] S. Schmidt, and et al., *"Text classification based filters for a domain-specific search engine,"* CI, vol. 78, pp. 70-79, 2016.

[18] N. Majumder, et al., *"Deep Learning-Based Document Modeling for Personality Detection from Text,"* IEEE, vol. 32, no. 2, pp.74-79, 2017.