Document Categorization using Improved KNN Classification

Neha[1] Dr. RK Chauhan[2]
[1]M. Tech. Research Scholar [2]Senior Professor
[1,2]Department of Computer Science & Applications
[1,2]Kurukshetra University, Kurukshetra, India

Abstract— Document categorization is the method of classifying the documents from mixed documents into particular specific documents such that they belong to the same classes. Classification is a data mining technique used to predict group membership for data instances. The relevance of keywords in documents and text mining has become very essential. An easy way of storing creates the need for a convenient way of retrieval which simplifies that what is the use of storing documents if they cannot be found. Resultantly, categorization of documents has been applied to make it easier to find relevant information. Classifying the documents is more convenient and virtuous. Thus, the main aim of this research is to design an improved KNN classifying technique so as to classify large sets of documents with improved accuracy in lesser time in terms of F-measure and G-measure.

Key words: Data Mining, Improved KNN, Centre Prediction, TF-IDF, Confusion Matrix, Euclidian Distance

I. INTRODUCTION

Data mining is the subfield of computer science which deals with the computational process of discovering patterns in large data sets. The main aim of data mining process is to extract vital information from a data set and convert it into an understandable format for future use. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from various different categorize it and summarize the relationships which are identified. Technically, data mining is the process of finding patterns among a large number of fields in large relational databases. Knowledge Discovery Diagram (KDD) is used to make sure that useful information is resulting from the data [2] It is basically used to retrieve useful information from the database. KDD have the following steps:

1) Data cleaning: It is a first phase in which irrelevant data are removed from the collection of data.
2) Data integration: It is a stage in which heterogeneous multiple data sources or similar data sets are combined to a common source.
3) Data selection: In this step the data related to the analysis is decided on and retrieved from the data collection.
4) Data transformation: It is also called as data consolidation in which the selected data is transformed into various forms which are appropriate for the mining process.
5) Data mining: It is the most important step in which knowledgeable techniques are applied to retrieve potentially useful patterns.
6) Pattern evaluation: In this phase the strictly patterns representing the knowledge are recognized based on the given measures.
7) Knowledge representation: It is the final step in which the discovered knowledge is visually represented to the user.

The basic block diagram of the data mining process is shown in Fig. 1 below:

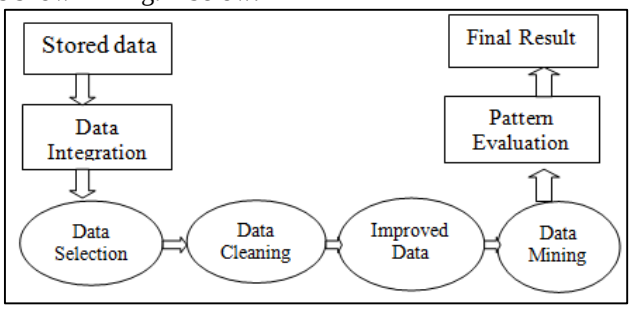

Fig. 1: Block Diagram of Data Mining Process

Document categorization refers to the process of dividing the large amount of text to one or multiple categories according to the contents or attributes of the text. Document categorization mainly includes two parameters-

- Clustering: It is the task of discovering structures and groups in the data that are similar in some way or another.
- Classification: It is the task of generalizing known structure to apply to the new data.

The document classification process is mainly divided into various steps.

Decision Tree classification: Decision tree classification is a flowchart like a tree structure where each internal node or the non-leaf node represents a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. When a decision tree is built, many branches will reflect the anomalies in the training data because of the noise or outliers. Basically the training data will not fit in the memory so Decision tree construction becomes inefficient due to swapping of the training.

Bayesian classification is a statistical classification. They can predict class membership probabilities such as the probability that a given data belongs to a particular class. A Bayesian network or directed acyclic graphical model represents a set of random variables and their conditional dependencies through a Directed Acyclic Graph.

Frequent patterns and their correlation rules characterize interesting relationships between attribute conditions and class labels and thus widely used for classification. Association rule shows a strong association between attribute-value pairs that occur frequently in the data set. Association rule is mainly used to analyze the purchasing of customers in a store.

Nearest Neighbor Classification is the classification using instance-based classifier which simply used for locating the nearest neighbor in instance space and labeling the unknown instance with the same class label as that of known neighbor forming a VSM vector space model [1]. The nearest neighbor classifier can be considered as a special case of more general K-nearest neighbor classifier therefore it is referred to as a KNN classifier. The best

choice for K generally depends on the value of data, larger values of K decreases the effect of noise on the classification but make boundaries between less dissimilar classes. A good K value can be determined by various heuristic techniques. The special case where the class is predicted as a class to be closest training sample (K=1) is considered to be the nearest neighbor algorithm. Nearest neighbor rules in effect absolutely compute the decision boundary. It is possible to compute the decision boundary explicitly so as to conclude that the computational complexity is a function of the boundary complexity. [3].

II. RELATED WORK

A. Using Association Rule with Decision Tree

According to the analyzed paper on Text classification using decision tree, out of the total data set using 76% training data an acceptable accuracy was obtained while it is suitable to obtain good accuracy using only 40 to 50% of total data sets as training data [3]

B. Using Association Rule with Naïve Bayes Classifier

According to the research done on Text Classification Using the Concept of Association Rule of Data Mining where Naive Bayes Classifier was used to classify the text, it showed the dependency of the Naïve Bayes Classifier with Associated Rules [4]. But as this method ignores the negative calculation for some specific class determination in some cases accuracy may be decreased. As a conclusion if the test set matches with a rule set, which has weak probability to the actual class then it may cause wrong classification.

C. Using document categorization using K-means

Document categorization is very vital method for selecting the proper required documents among the thousands of documents. The research has shown that the improved document categorization method gives improvement in the clustering of the same documents. The research paper shows that the problem of automatic clustering is obtained by considering features as cluster labels.[1]

D. Using KNN Classification Algorithm

This research paper proposed that the KNN classification can be used which saves calculation, improves the computing speed and thus is much better for the large amount of the data.

When the sample text classification increases then the calculation of the amount of the reduced amplitude goes up. This paper holds some disadvantages also such that the choice of the threshold has a crucial impact on the classification which reduces the accuracy of the classification [4]. Therefore it is better used for accuracy by the systems less demanded.

III. PROPOSED WORK

This proposed method uses improved KNN classification method for the better accuracy, efficiency and better results. The K-Nearest Neighbor (KNN) Classifier [5] is a very simple classifier that works well on basic recognition problems. The comprehensive process of categorization of the data is as follows. Firstly three data sets are taken from a newsgroup each having 100 words consists in it. Then these data sets are pre-processed so that the noise values are removed which involves 4 steps ie. Removing of the special symbols and stop words. Through porter algorithm all the words are changed into base word and lower case conversion method will convert all the words to the lowercase. The data set thus obtained will be consisting of the required data sets which need to be categorizing into different classes. After that Euclidian distance method is used instead of cosine method for predicting the center because cosine method gives false values with respect to zero entity. After the center is predicted there will be three sets of classes having the homogeneous data sets belonging to particular classes. This forms a vector space model (VSM) [6] which splits into training and test data percentage [7]. Then training percentage applied on the data due to which the data sets will be matched with those classes and the distance is labeled. Finally this gives a good prediction matrix and better accuracy.

A. Basic Flow Chart of Proposed Approach

Center prediction: Euclidian distance method is used for predicting the center points.

1) Algorithm 1: for predicting center points

Input: VSM, Ki terms, weight

Output: Set of Ki initial centroids

1) Create initial center matrix for K terms and set the default values as zero.

2) T=1

3) For Ki=1 to K

If (K==1)

Center (Ki,:)=w

Set first terms from terms t to t_{29}

t=t+29

Set from t_{30} to t_{60}

end

2) Algorithm 2: for improved KNN

Input: VSM, Ki terms, weight, training percentage

1) Split VSM into test and train data using %
2) Training = n*train%
3) Select random train size data from VSM and put it into train data and remaining into test data.
4) Prepare prediction matrix
5) For each t_i
6) Vi=feature of t_i
7) For k_i-= 1 to k
8) V2= feature of k^{th} row of prediction matrix
9) D= calculate each using formula of Euclidian
10) dist(ki) = d
11) Min dist = d
12) Apply Precision formula

IV. EXPERIMENTS AND RESULTS

This approach has been proposed that the process time of proposed work is less than from the previous approach because previous approach was less trained and uses cosine

similarity approach for finding the center points but this work uses improved KNN classification with greater training percentage and Euclidian distance method is used for predicting the center points because cosine similarity method could give the false values with respect to zero values. This approach gives better results for precision and recall and calculating these values is easier because of the confusion matrix formed. Finally, the accuracy of this proposed work is high and takes less time for calculating the results. The algorithm is tested on mini Newsgroup datasets. The work is implemented in MATLAB. Following three categories of the data set is taken

Each mini-group consists of the data sets. These data sets are applied with the i-KNN classification.

Categories
Alt.atheism
Comp.graphics
Comp.os.ms-window.misc

Table 1: mini_newsgroup

The below figure shows the value of accuracy, sensitivity, specification, precision, f-measure and time complexity for both existing and proposed algorithm. The values of accuracy has greater value for proposed algorithm as compared to existing algorithm in lower time period thus shows that it is better than the existing algorithm

Trainper (60)	Accuracy	Senstivity	Specificity	Precision	F-Score	Time Complexity
Knn	0.83	0.74	0.87	0.75	0.74	0.018
i-knn	0.89	0.84	0.92	0.84	0.84	0.004

Table 2: Value of Accuracy for Existing and Proposed Algorithm

The precision has much improvement in the proposed values as compared to the existing algorithms. The result for the value of f-score has much better results than the existing work.

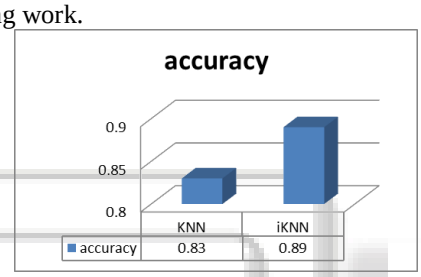

Fig. 2: Accuracy Measure

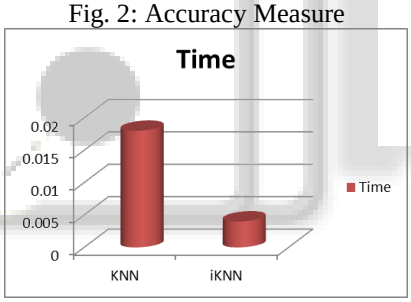

Fig. 3: Time Complexity Measure

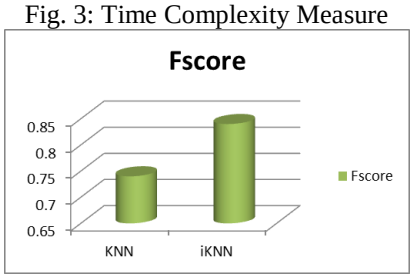

Fig. 4: F- Measure

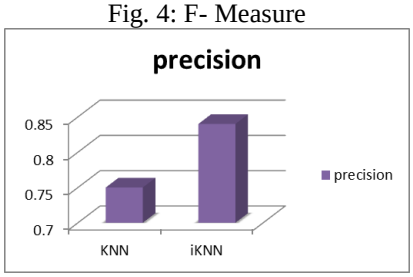

Fig. 5: Precision Measure

V. CONCLUSION & FUTURE SCOPE

Classification algorithm is a crucial mean in data mining process. There are several research papers those worked upon document clustering and document classification but clustering method does not give better results of accuracy because in classification data is already trained before testing. This research makes some improvements in the current KNN algorithm and generally focuses on the disadvantage of less accuracy and calculations. In the era of data explosion accuracy and time complexity plays a vital role. This improved KNN method can further be used on large number of the data sets as here few data sets are taken into consideration. Furthermost, the data will be secured only when it is well structured and some security or protection is applied on this.

REFERENCE

[1] Promod Bide, Rajashree Shedge, "Improved Document Clustering Using K-Means Algorithm", International Conference on Electrical, Computer and Communication Technologies (ICECCT- 2015), pp: 1-5, IEEE, 2015.

[2] A.Shameem Fathima, D.Manimegalai and Nisar Hundewale "A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue", IJCSI, vol 8 pp.322-328, 2011

[3] Masud M. Hassan, Chowdhury Mofizur Rahman, "TextCategorization Using Association Rule Based Decision Tree," Proceedings of 6th International Conference on Computer and Information Technology, JU, pp. 453-456, 2003.

[4] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining," In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation), New York, USA, 1998.

[5] Zhao, Y. and G. Karypis, "Evaluation of Hierarchical Clustering Algorithms for Document Datasets", Proceedings of CIKM, 2002.

[6] N. Suguna1, and Dr. K. Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010

[7] Guohua WU, L. W. (2015). Improved Expected Cross Entropy Method For Text Feature Selection. ieee, 49-54.

