ISSN (online): 2250-141X
Vol. 6 Issue 3, July 2016

HYBRID ONE-PASS CLUSTERING WITH ENHANCED KNN CLASSIFIER COMBINED WITH EM FOR TEXT CATEGORIZATION (H3TC) ALGORITHM

K.Gayathri
Research Scholar of Computer Science
Karpagam University
Coimbatore,India

Dr.A.Marimuthu MCA,M.Phil,Ph.D.,
PG & Research Dept. of Computer Science
Government Arts College,
Coimbatore, India

ABSTRACT

K-Nearest Neighbor (KNN) is one of the most popular algorithms for pattern recognition. Many research have found that the KNN algorithm accomplishes very good performance in their experiments on different data sets. The traditional KNN Text Classification Algorithm has 3 limitations: (a) Calculation complexity due to the usage of all the training samples for classification, (b) the performance is solely dependent on the training sets and (c) there is no weight difference between the samples. To overcome these limitations, the KNN is combined with one pass clustering and EM Algorithms to improve its classification performance. To classify the test samples, calculate the distance. Before classification, initially, the reduced feature set is received from TF/IDF method which was discussed in the earlier work. The classifications are according to various parameters, measurements and analysis of results.

Keywords: Text Categorization, KNN, One Pass Clustering, EM –Algorithm.

I .INTRODUCTION

Document classification is a problem in today's library science information and computer science. In the past 20 years, the number of text documents in digital form has been grown exponentially. As a consequence of exponential growth great importance has been put on the classification of documents into groups that describe the content of the documents. The function of classifier is to merge text document into one or more predefined categories based on their content. Each document belong to several categories or may present its own category. A very rapid growth in the amount of text data leads to expansion of different automatic methods aim to improve the speed and efficiency of automated document classification with textual content. Many classification methods have been applied to text categorization, for example, Naïve Bayes

ISSN (online): 2250-141X
Vol. 6 Issue 3, July 2016

Probabilistic classifiers [1], Decision tree classifiers [2], Regression methods [3], Neural Network [4], KNN classifiers [3,5] & Support Vector Machine (SVM) [6]. In many applications, dynamically mining large web repositories the computational efficiency of these schemes is often the key element to be considered. Sebastiani pointed out in his survey on text categorization [7].K-Nearest Neighbors is one of the most popular algorithms for text categorization. Many researchers found that KNN algorithm achieves very good performance in their experiments on different data sets. [3]. The Nearest Neighbor rule identifies the category of unknown data point on the basis of its nearest neighbor whose class is already known this rule is widely used in pattern recognitions and text categorization. [8], T.M. Cover & P.E. Hart purpose K-Nearest Neighbor (KNN) in which nearest neighbor is calculated on the basis of value of K, that specifies how many nearest neighbors are to be considered to define a class of sample data point. But the computational complexity and memory limitation size of data set are reduced [9]. Clustering a common descriptive task in which one seeks to identify a finite set of categorization or clusters to describe the data. The clustering or the cluster analysis is a set of methodologies for classification of samples in one group or grouped and samples and the process of clustering is to measure the similarity or dissimilarity between given samples. The output of the clustering is a number of groups or clusters in the form of graphs, histograms and normal computer results showing the group numbers [10]. Expectation Maximization (EM) improves the classifier by the current classifier to guess the hidden variables and then using the current guess to advance the classifier training. EM consequently finds the classifier parameters that locally maximize the probability of both the labeled and unlabeled data [11].

II RESEARCH METHODOLOGY

- Jiang et al., (2012)- Improved KNN to solve these issues by combining it with a Single Clustering Algorithm.
- The Algorithm, treated as Improved KNN for Text Categorization INNTC, uses one pass clustering and to the results of clustering applies KNN classification. Steps in INNTC:

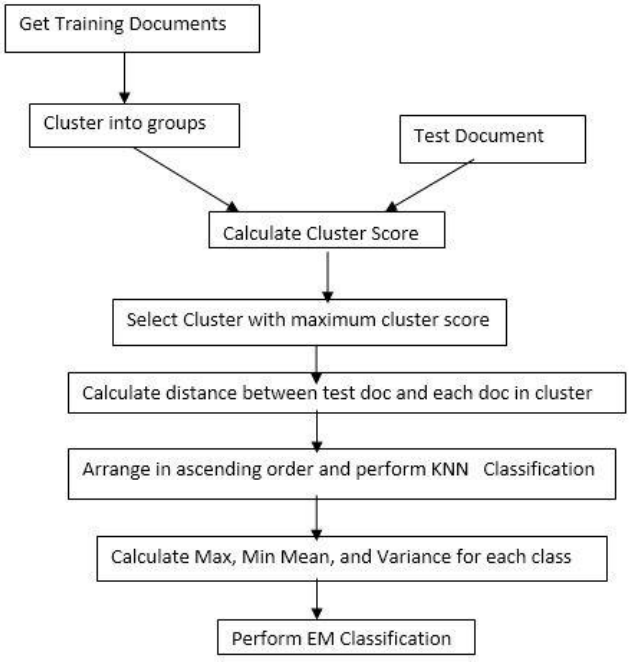


Fig-1 KNN Classifier combined with Expectation Maximization for Text Categorization (H3TC) Algorithm.

III. K-NEAREST NEIGHBOR (K-NN) ALGORITHM

KNN is a simple but effective method for text categorization, but it has three fatal defects: first, the complexity of its sample similarity computing is huge; second, its performance is

ISSN (online): 2250-141X
Vol. 6 Issue 3, July 2016

easily affected by single training sample, such as noisy sample; third KNN does not build the classification model since it is a lazy learning method. As a result, it is not suited in many applications. In this paper, proposes an improved KNN Algorithm for text categorization (INNTC). The classification model is obtained by employing constrained One Pass Clustering Algorithm which uses the least distance principle to constrainedly divide training text samples into hyper spheres with almost the same radius. Then employ with KNN approach to classify the test collections based on the obtained model. INNTC builds the classification model, which reduces the test similarity computing complexity.

3.1 Improved K-Nearest Neighbor Algorithm for Text Categorization

KNN is a sample based learning method, which uses all the training documents to predict labels of test document and has very huge text similarity computation. As a result, it cannot be widely used in the real-world applications. To undertake this problem, proposes an improved KNN algorithm for text categorization based on One-Pass Clustering Algorithm and KNN Algorithm.

3.2 Traditional K-Nearest Neighbor Algorithm for Text Categorization

The process of KNN algorithm is as follow: Given a test document x, find the K nearest neighbor of x among all the training documents, and score the category candidates based on the grouping of K neighbors. The similarity of x and each neighbor document is the score of the category of the neighbor document. If several of the K nearest neighbor document belongs to the same category, then the sum of the score of that category is the similarity, keep count of the category in regard to the test document x. By sorting the scores of the contestant categories, system assigns the candidate category with the highest score to the test document x. This approach is effective, non-parametric and easy to implement. However its classification time is long and the accuracy is severely corrupted by the presence of strident training document [12].

3.3 Conventional KNN Classification:

- Given ‘K’ value, the goal of KNN Algorithm is to find K nearest neighbors for a document among all documents

TRAINING:

Step 1: Select a ‘K’ value.

Step 2: Consider a document X.

Step 3: Calculate Similarity of a document ‘X’ with its neighbors.

Step 4: Arrange in ascending order of similarity score and group all documents.

Within a minimum distance.

Step 5: Repeat Steps 2 to 4, until process converges.

TESTING:

Step 1: Calculate similarity of an input documents.

Step 2: Calculate Sum score of each group.

Step 3: Find group with nearest similarity score.

Step 4: Assign input document to that group.

ISSN (online): 2250-141X
Vol. 6 Issue 3, July 2016

3.4 Improved K- Nearest Neighbor Algorithm for Text Categorization based on Clustering (INNTC)

3.4.1. Build Classification Model Based Clustering

Clustering is a process of partitioning data into clusters of similar objects. It is an unsupervised learning process of hidden data. In text, clustering assumes the similarity degree of the substance of the documents in the same cluster is the most, while in different clusters to the least. Therefore to preprocess the documents using clustering are useful for discovering the distribution and structure of corpus. To build the classification model with the training text documents, use One-Pass Clustering Algorithm to constrainedly cluster the text collection.

The details about the clustering are illustrated as follow:

1. Initialize the set of clusters m_0, as the empty set and read a new text 'p'.
2. Create a new cluster with the p; its label is regarded as the label of the new cluster.
3. If no texts are left in the text collections, go to (6), otherwise read a new text p, compute the similarities between p and all the clusters and find the cluster that is closest to the text p.
4. If similarity is less than a threshold 'r', the group text to that cluster else go to (2).
5. Merge text p into cluster and update the weight of the words of cluster; go to (3).
6. Stop clustering, get the clustering results.

3.5. Up-dating the classification

INNTC puts together the classification model using constrained One-Pass Clustering Algorithm, it changes the learning way of KNN algorithm. The number of clusters gain by constrained clustering is much less than the number of trained model. As a result, when uses KNN methods to classify the test documents, the text similarity computation is significantly reduced and the contact of its performance affected by single training samples are also shaped.

3.6 Analysis of INNTC

Time complexity reduced, when compared to conventional KNN, but still was high with large sized data sets.

- The choice of k has a great impact on the performance of classification.
- The accuracy of the INNTC improved classification accuracy by 4.22% when comparing to conventional KNN Algorithm, but careful analysis shows that there are steps which can improve this algorithm.

IV.EM Algorithm

The Expectation-Maximization (EM) Algorithm (Dempsteret al., 1977) is a popular class of iterative Algorithms for maximum likelihood estimation in the problems with incomplete data. It is often used to fill the missing values in the data using existing values by computing the expected value for each missing value. The EM Algorithm consists of two steps, the *Expectation* step, and the *Maximization* step. The *Expectation* step basically fills in the missing data. The parameters are estimated in the *Maximization* step after the missing data are filled or reconstructed. This leads to the next iteration of the Algorithm [13].

ISSN (online): 2250-141X
Vol. 6 Issue 3, July 2016

V. Experimental Data Set

5.1 Retures- 21578.

The Retures-21578 data set and used the standard "modApte" train/test split. These documents appeared on the Reuters newswire in 1987 and were manually classified by personnel from Reters Ltd. ModApt'e split-9603 training and 3299 testing documents. Total-12902 documents. Out of 135 categories of documents only the top five were selected. Distinct words - 31715; Average numbers of words per document-126 words of which 70 were distinct.

5.2.20-NewsGroup Corpus:

The 20-News Group Corpus contains approximately 20,000 News Group documents being partitioned evenly across 20 different News Groups. Already used 20 news 18828 version for evaluation.

Table-2

Dataset	TDRF	LSI	MLSI
Reuters			
2000	0.86	0.82	0.89
4000	0.86	0.82	0.89
6000	0.87	0.83	0.90
8000	0.88	0.84	0.91
10000	0.88	0.84	0.90
12000	0.87	0.83	0.90
14000	0.88	0.85	0.90
20 News Group			
2000	0.87	0.83	0.89
4000	0.87	0.84	0.90
6000	0.88	0.83	0.91
8000	0.89	0.84	0.91
10000	0.89	0.84	0.92
12000	0.89	0.85	0.92
14000	0.89	0.86	0.93

Table-1

VI. Performance Metric

The evaluation of a classifier is done using the precision and recall measures. To derive a robust measure of the effectiveness of the classifier. It is able to calculate the break-even point, the 11-point precision and average precision to evaluate the classification for a threshold ranging from 0 (recall=1) up to a value where the precision value equals 1 and the recall value equals 0, incrementing the threshold with a given threshold step size. The break-even point is the point where recall meets precision and the eleven values 0.0, 0.2…0.9. "Average precision" refines the eleven point precision as it approximates the area "below" the precision /recall curve.

ISSN (online): 2250-141X
Vol. 6 Issue 3, July 2016

VII. Results

Experimental Results Precision

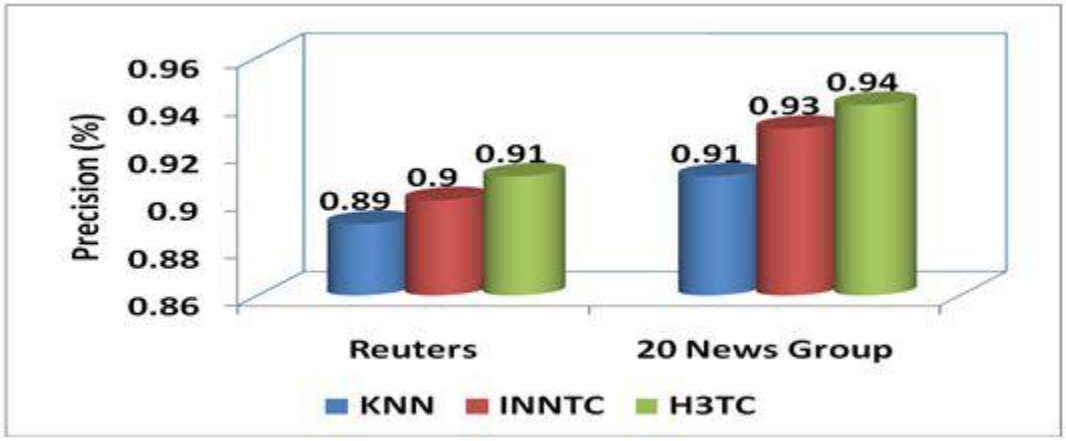


Fig-2.1

Experimental Results Recall

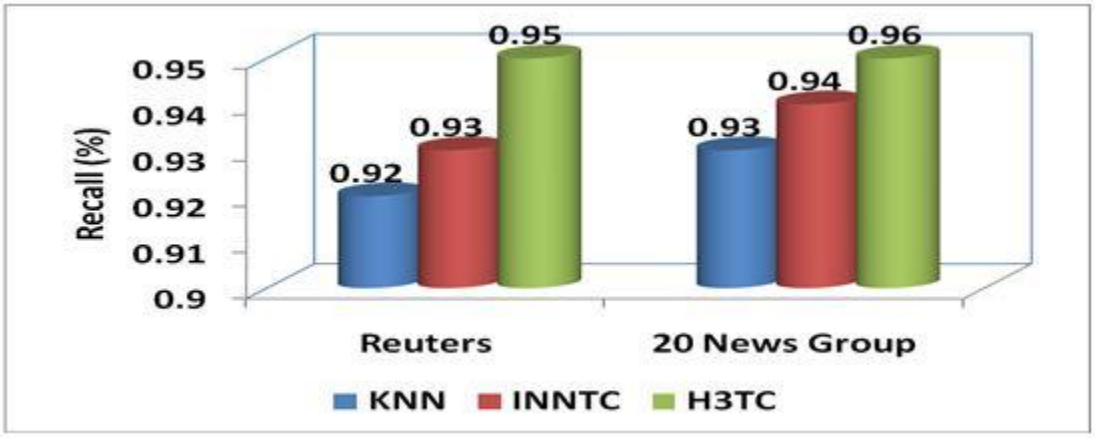
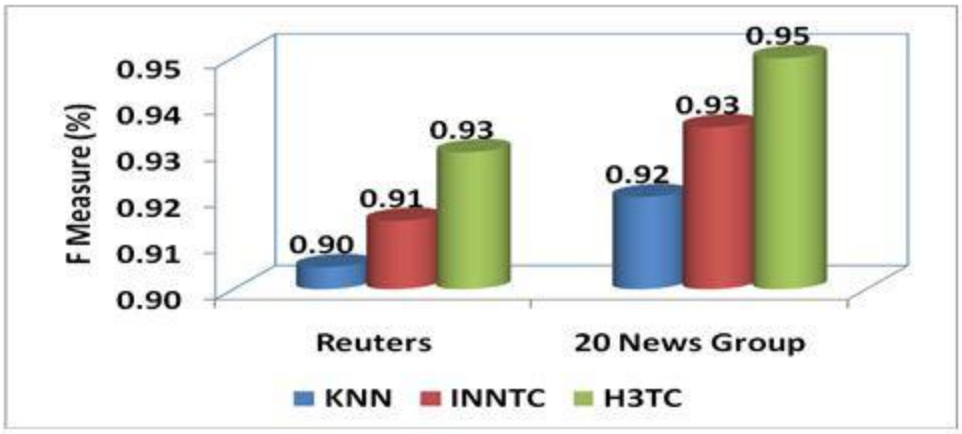


Fig-2.2

ISSN (online): 2250-141X
Vol. 6 Issue 3, July 2016

Experimental Results F-Measures



Speed

Fig-2.3

ISSN (online): 2250-141X
Vol. 6 Issue 3, July 2016

VIII. Conclusion

The advantage of the proposed approach is the classification Algorithm learns the importance of attributes and utilizes them in the similarity measures. In future, the classification model can be build, which analyses terms on the concept sentence in document.

IX. REFERENCES

I. Lewis D (1998) Naïve Bayes at forty: "The independent assumption in information retrieval", In: Proceedings of ECML-98, 10th European conference on machine learning pp 4-15.

II. CohenW,SingerY (1999) "Context-sensitive learning methods for text categorization",ACM Trans Inform Syst 17(2):141-173.

III. Yang Y, Liu X (1999) "A re-examination of text categorization methods", In: Proceedings of SIGIR-99, 22nd ACM international conference on research and development in information retrieval, pp: 42-49.

IV. Ruiz M, Srinivasan P (1999) "Hierarchical neural networks for text categorization", In Proceedings of SIGIR-99, 22nd ACM international information retrieval, pp281-282.

V. Mitchell T (1996)" Machine learning",McGraw Hill, New York .

VI. Joachims T (1998) "Text categorization with support vector machines: learning with many relevant features", in: Proceedings of 10th European conference on machine learning, Chemnitz, germany, pp 137-142.

VII. Sebastiani F (2002) "Machine learning in automated text categorization",ACM Comput Surv;1-40.

VIII. T.Bailey & A.K. Jain "A note a Distance weighted K-Nearest Neighbour rules". Trans.systems, Man Cybernatics, Volume-8, pp: 311-313.

IX. T. M.Cover & P.E.Hart "Nearest Neighbor pattern classification",IEEE .Inform Theory Vol:it-1, jan-1967.

X. Taeho jo "The Implementation of Dynamic Document Organization using the Integration of Text Clustering and Text Categorization " Ottawa-Carleton Institute for Computer Science, School of information Technology and Engineering (SITE).canada.

XI. Andrew Kachites Mc Callum, Kamal Nigam "Employing EM & Pool Based Active Learning for Text Classification".

XII. Shengyi jiang,"An improved K-Nearest –Neighbor algorithm for text categorization" Elsevier, School of Informatics, china.

XIII. Dempster, A., Laird, N. M., & Rubin, D. (1977)."Maximum likelihoodfrom incomplete data via the EM algorithm", *Journal of the Royal Statistical Society B*, 391–38.