An Efficient Dimension Reduction in Text Categorization Using Clustering Technique

[1]Ms. V. INDUMATHI, [2]Dr. P. VIJAYAKUMAR, M.Sc., M.Phil., Ph.D

[1]Research Scholar, Department of Computer Applications, Sri Jayendra Saraswathy Maha Vidyala College of Arts & Science, Coimbatore, India.

[2]Associate Professor & Head, Department of Computer Applications, Sri Jayendra Saraswathy Maha Vidyala College of Arts & Science, Coimbatore, India.

Abstract: - Dimension reduction is a process for reducing random variables, which can be divided into feature selection and feature extraction [4]. When the dimension increases, the performance efficiency decreases in index configurations. Dimensional reduction algorithms, which is the only solution to support retention material and the satisfaction of query decisions.

Clustering is the process for identifying commodity groups that are different from the commodities of the commands of the groups and are different from the other groups. The dimensional reduction is to transform the higher dimensional quality into a meaningful representation of the lower dimension similar to the inherent dimension of the data. k-Nearest Neighbors (kNN) clustering algorithms often do not work well for high dimensions, so as to increase efficiency, the IG in the original data set can be achieved and the lower the dataset with less variable variables. The key component analysis and linear transformation in this paper is used for dimensional reduction and the initial center is calculated, then it is the k-Nearest Neighbor (kNN) clustering algorithm.

Keywords: - Mobile, Privacy, Preserving, Sensing, System, Security.

I. INTRODUCTION

Dimension reduction is a method of obtaining information from the high-dimensional point of view using less intricate dimensions. The mechanical learning method is crucial to cutting high-level data packages for better classification, setbacks, presentation and visualization of data. It is also useful to better understand the relationships within the data. It helps us to find the internal dimension of the database and the best generality. There are many approaches to dimensional reduction, which can be linear or linear. The choice of text classification is a well-researched problem; Its targets improve classification efficiency, computational performance or both. Highlighting a high loss is the use of [12] for performance increases in many cases. A comparative study of the multi-feature examination criteria used in the filter model [4] and the efficiency of the $\chi 2$ stats and database classification results are efficient and performance is a minor degradation. [5], [11] and [2] used hybrid approaches to relate to dependencies and volatility. These algorithms require intense computations of interactions between astrologers, which are difficult to measure the greatest highlight.

A clustering technique that gives better clusters to fit inside the cluster is low and performance is better when compared to external clatters. Clustering algorithm is usually superficially measured by comparative action used in both intelligence and its processing. It is also measured by detecting the arbitrary forms of clusters. However, objective assessment is complex and is usually done by man or expert analysis.

In section 2 we review some current works that make dimension reduction to do affective computing that are concerned about data but they do not provide a current solution to this problem. Section 3 presents the technique and proposals to provide a dimension reduction solution that can be applied in affective computing. The system proposal and analytical verification is presented in Section 4. Finally, some conclusions are derived in Section 5.

II. LITERATURE REVIEW

DR is the process of transforming a big amount of data into a far smaller, much less noisy illustration that preserves important relationships from the authentic records. DR strategies have been proven to effectively simplify large geometric datasets, but haven't begun to be adequately evaluated for textual records. This task evaluated five DR strategies (Principal Components

 DOI : 10.18231/2454-9150.2018.1130

Analysis, Multidimensional Scaling, Isomap, Locally Linear Embedding, and Laplace-Beltrami Diffusion Maps) from two awesome views.

G.N. Ramadevi and K. Usharani (2013), proposed, Data is not collected for data mining. Data is accumulating at an unprecedented speed. Data initiative for efficient machine learning and data mining is an important part. Data Mining is finding an interesting interest from large-scale data, which is an integral part of KDD (Knowledge Discovery Databases), an overall process for converting raw data into useful information. Transfer reduction not only for computational efficiency, but also for improving analysis accuracy. The techniques used for transfer reduction can be divided into two main ways, which are feature selection and feature extraction [1]. They may apply for a study that is not supervised or supervised. In this paper, we analyze dimensional reduction techniques and provide its applications in real applications. PCA (Principle Component Analysis) is a dimensional reduction technique of the feature reduction mechanism to reduce the dimension of the database without losing data. The clustering ranking can be used in PCA rankings before reducing substantial timing. PCA is used for data visualization and noise reduction.

V. Arul Kumar, and N. Elavarasan (2014), proposed, Data mining is the automatic extraction of useful and often unknown information from databases or databases. Data collected from real world applications has a lot of misinformation. Data initiative is a crucial technology in the data mining correction of erratic data in the dataset. Many data mining applications contain more dimensional data. The higher the dimension will minimize the performance of mining algorithms and increase the time and space required for data mining. High dimensional problematic dimension reduction (DR) technique is solved. DR is divided into two: feature selection and feature extraction. In this sheet, she studied extensively about how the dimensional problem is resolved by using two different techniques.

Tajunisha and Saravanan (2011), proposed, Cluster analysis is one of the key analysis methods in data processing. K is a very popular and shared-based clustering protocol. But this is a predictable price and the result of clusters is largely dependent on the size of the initial center and the selection. K-means methods have been proposed in literature to improve clustering algorithm efficiency. Principle Component Analysis (PCA) is an important approach to unmanaged dimensional reduction technique. This paper is proposed once the method is very useful and efficient using the PCA and the modified k-algorithm. In this paper, the primary keypad is the primary hub for k-objects, and the dimension reduction and k-algorithm were first used in the primary molecular analysis, assigning data-value.

Prof. Rasendu Mishra, and Dr. Priti Sajja (2018), proposed, This review article uses the calculation of various dimension reduction techniques and efficiently uses group computing and timely calculation. A dimensional reduction comment can be used to collect and document recommendations for the question. The study lists various comparison reduction technologies with descriptions, advantages, and disadvantages of test comparisons.

III. EXISTING METHODOLOGY

Clustering is considered an unsupervised learning process, the main purpose of which is to compile unnamed documents with meaningful shelters that are similar to those of the other clusters. Clustering documents are attractive because it creates settings based on documents that may be very expensive or may be given by the use of the time limits of the application and / or the relevant documents. Machine caching algorithms used for text clustering are classified into two main groups (i) phase clustering algorithms, and (ii) as partition clustering algorithms.

Stages clustering algorithms create a local partition of the data, linking or splitting clusters based on their unity. On the other hand, partition-based clustering methods compile data into non-overview partitions, which can usually improve a clustering parameter. If the data is naturally in hierarchy, the stage cluster provides better visualization capabilities. However, it does not have the weakness because it is very important for seizures. In addition, computational clustered calculation time is huge, which controls its use in large data.

IV. PROPOSED METHODOLOGY

Dimension reduction can take different forms. Feature selection, TRR contains less significant features from the package, eg, information gain [2]. Instead of extracting feature on DR, we focus on where new features are created based on some (potentially complex) transformation of input elements. DR modes can be divided into supervised and non-supervised techniques, which involve the supervision techniques of labels associated with input events (or documents). We'd like to simplify text processing applications such as labels that are not often known (eg, with aforementioned emotional analysis), because we take note of unpleasant DR Techniques.

In our experiments, we begin with a collection of documents, encode the documents into a term-document matrix, and then perform dimensionality reduction on this large matrix. Using the reduced matrix, we then classify each document into one of several known categories. Finally, we evaluate the results. Below we elaborate on each of these steps.

 DOI : 10.18231/2454-9150.2018.1130

4.1. Document Encoding

To make more processing easier, we have to tag our documents first. For general text mining, each document is labeled as a team described in a row in a row. We have the possible matrix and ways to calculate them, but based on the previous work [12, 16] we follow the following linear plan. This time, each column is akin to a word found in the corpus.

The resulting team is known as a (gained) term-document matrix (TDM). In our process, we reject words before we calculate TDM and reject any word at least three times in the corps.

4.2. Dimensionality Reduction

The encoded TDM contains 600-3000 rows and 5000-11000 columns. We reduce the next dimension to significantly reduce the columns of this number.

We implement every node in Java and then check everything against the modal or R-code that we have from others. How close to neighboring countries is to refer to Isomac and LLE when creating a nearby neighborhood; We tested 10 neighbors and gave good results. Each technique first reveals the most important dimensions and helps us analyze the impact only using the first M dimensions of classification. In addition, we have been comparing against two independent types of T. Firstly, no-rand uses randomly selected MM features from TMM Matrix. Secondly, no variables use M constant features, but N selects the maximum number of TF-IDF scores (more keywords). These two variants replace the potential classification effectiveness without having to operate any DR.

4.3. Classifiers

We evaluated with three commonly-used classifiers:

4.3.1. k-Nearest Neighbor

It presents a type of k class (k) of the closer events (near neighbor) class (H) in training data to determine the synchronization function. In the course of our work, using the distance of the cos used to calculate the similarity and selecting the most segments of the use of the votes with the neighborly family.

We calculate classification accuracy, which is the percentage of documents sent to the correct category. A test for testing all tests is used once in a procedure.

The k-NN Algorithm

Step 1 - Load the data
Step 2 - Initialize K to your chosen number of neighbors
Step 3 - For each example in the data
Step 4 - Calculate the distance between the query example and the current example from the data.
Step 5 - Add the distance and the index of the example to an ordered collection
Step 6 - Sort the ordered collection of distances and indices from smallest to largest by the distances
Step 7 - Pick the first K entries from the sorted collection
Step 8 - Get the labels of the selected K entries
Step 9 - If regression, return the mean of the K labels
Step 10 - If classification, return the mode of the K labels.

V. PERFORMANCE EVALUATION

Common estimates are used to evaluate our results with literature.

5.1. Data Sets

High-dimensional data contains several thousands of columns. Any database operates under a corresponding model, which is selected as a high-dimensional database. The following five different datasets like, 20NG, sports, health, society and local news for experiment.

Category	No. of Documents
20NG	412
Sports	300
Health	669
Society	442
Local News	254

Table 1: - The Category of Dataset

Used databases differ in vocabulary size and type distribution. Each document is related to the 1000 attributes or dimensions and is classified using dimensions reduction techniques. The details of each database are listed below,

5.1.1. 20NG Dataset

20 Newsgroups (20NG) is a collection of nearly 412 articles for Usenet newsgroups. Some news packages are very relevant (e.g. autos, motorcycles), others are much unrelated. The standard split "byte" was used, with copies and titles removed. This decision is for 18,941 papers, 60% of which are allocated for training packages, while the test package contains 40%.

5.1.2. Sports

It is a collection of 300 game documents. These documents were collected from the five types of sports fields. The categories used in this work are: players, places, posts, materials and science. However, the number of documents in this database is small and the difference in the nature of the type is larger. This makes the

 DOI : 10.18231/2454-9150.2018.1130

classification process easier. In order to compare the results of these job results, the same split was used.

5.1.3. Health

It is a collection of 669 notes collected from 270 medical journals published in various years. In this work, the use of two sub-applications was called patient and doctor. In this subset, only 669 documents containing the summaries published each year resulted in only 23 versions. The first 669 was used as a training suite and the remaining test suite.

5.1.4. Society

The database contains 442 articles. The selection of these articles selected at least one article each week. The documents of this database are distributed almost in eight categories. The diagram is evaluated as there is no fixed division for the training and test documentation of this database. Five randomized selective divisions were constructed for each of the four extracurricular training documents reflecting the fifth of the total number of documents.

5.1.5. Local News

It has been a standard benchmark in Text Categorization for the last 10 years [48]. It consists of over 254 news stories appeared.

5.2. Performance Evaluation Parameters

The following performance parameters are commonly used in Dimensionality Reduction technique evaluation. The existing approach is compared with proposed scheme using these evaluation parameters. The performance of the TC process can be measured by one or more of the following methods:

5.2.1. Recall and Precision

They are two well known measures of effectiveness in text mining. While Recall is a measure of correctly predicted documents by the system among the positive documents, Precision is a measure of correctly predicted documents by the system among all the predicted documents.

Recall

Dataset Categories	No. of Documents	Clusters	Existing (K-Means)	Proposed (k-NN)
20NG	412	13	50 Sec	45 Sec
Sports	300	9	70 Sec	60 Sec
Health	669	22	81 Sec	70 Sec
Society	442	15	83 Sec	76 Sec
Local News	254	7	86 Sec	82 Sec

Table 2: - Characteristics of Datasets used in the Recall Comparative Study

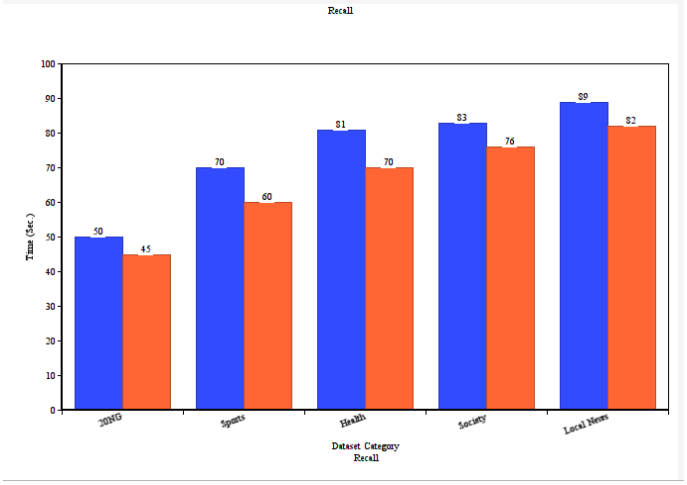

Fig 1: - Evaluation of Recall using kNN

In the above plotting, the existing hierarchical approach takes more time for extract the result from the dataset. The blue line represents the existing approach and the red line represents the k-NN for executing the dimensionality reduction of the documents with various categories..

Precision

Dataset Categories	No. of Documents	Clusters	Existing (K-Means)	Proposed (k-NN)
20NG	412	13	75%	90%
Sports	300	9	61%	81%
Health	669	22	58%	70%
Society	442	15	65%	80%
Local News	254	7	68%	73%

Table 3: - Characteristics of Datasets used in the Precision Comparative Study

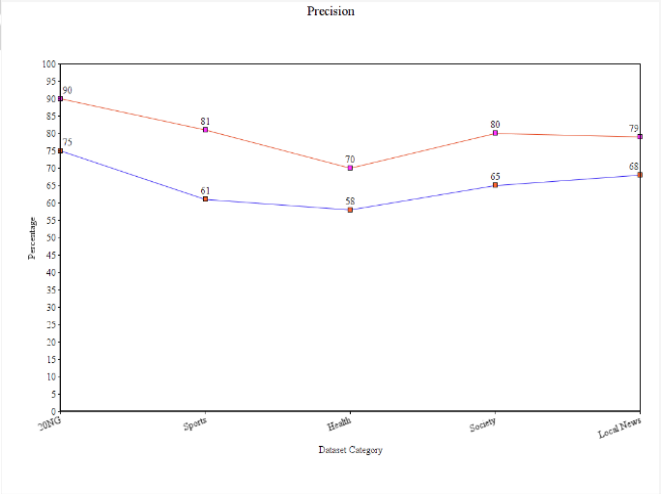

Fig 2: - Evaluation of Precision using kNN

The precision represents the accuracy of retrieval or categorizing the data. Existing approach accuracy level is poor compare with the k-NN Approach of the proposed one.

 DOI : 10.18231/2454-9150.2018.1130

In the above result, the red line represents the existing approach and the blue line represents the k-NN for executing the dimensionality reduction. Existing approach accuracy level is poor compare with the k-NN Approach of the proposed one.

5.2.2. F-Measure

F-measure combines precision and recall and is the harmonic mean of precision and recall.

Dataset Categories	No. of Documents	Clusters	Existing (K-Means)	Proposed (k-NN)
20NG	412	13	75	80
Sports	300	9	82.35	87
Health	669	22	86.35	90
Society	442	15	87.89	92
Local News	254	7	81.40	92

Table 4: - Characteristics of Datasets used in the F-Measure Comparative Study

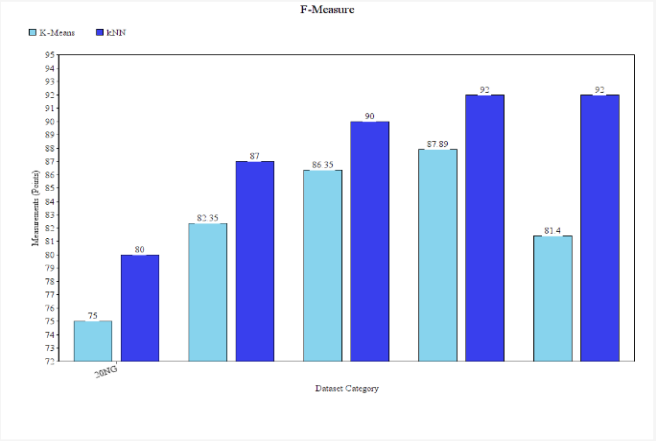

Fig 3: - Evaluation of F-Measure using kNN

The F-Measure represents the measure of recall and precision of retrieval or categorizing the data. In the above result, the light blue line represents the existing approach and the dark blue line represents the k-NN for executing the dimensionality reduction. These measures are very helpful in evaluating the performance of both frequent and rare categories.

VI. CONCLUSION

The main goal is to achieve the highest performance with simple techniques. This work demonstrates this approach to the DR process due to the simplicity and efficiency of the feature filtration approach. Using the key component analysis results in significant reduction in the amount of vocabulary that results in a significant loss of accuracy. On the other hand, the use of k-Nearest Neighbor approach decreases storage requirements. However, the current approach leads to some degradation in significant performance in large databases.

The k-Nearest Neighbor (kNN) has improved the proposed integration and development of clustering approach. This also leads to improved classification accuracy and the saving of the set size. In addition, this reduces reduction savings and reduces computational resources. The proposed systems have shown comparable results with benchmark data.

REFERENCES

[1] D. L. Donoho, High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Lecture on August 8, 2000, to the American Mathematical Society "Math Challenges of the 21st Century". Available from http://www-stat.stanford.edu/~donoho/.

[2] M. Turk, A. Pentland, Eigenfaces for Recognition, *J. Cognitive Neuroscience*, **3**-1 (1991) 71-96.

[3] R. Bellmann, Adaptive Control Processes: A Guided Tour. Princeton University Press, 1961.

[4] A. Barron, Universal Approximation Bounds for Superpositions of a Sigmoidal Function, *IEEE Tr. On Information Theory*, **8**-3 (1993) 930-945.

[5] D. W. Scott, J. R. Thompson, Probability density estimation in higher dimensions. In: J.E. Gentle (ed.), Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface, Amsterdam, New York, Oxford, North Holland-Elsevier Science Publishers, 1983, pp. 173-179.

[6] P. Comon, J.-L. Voz, M. Verleysen, Estimation of performance bounds in supervised classification, *European Symposium on Artificial Neural Networks*, Brussels (Belgium), April 1994, pp. 37-42.

[7] B.W. Silverman, Density estimation for statistics and data analysis. Chapman and Hall, 1986.

[8] P. Demartines, Analyse de données par réseaux de neurones auto-organisés. Ph.D. dissertation (in French), Institut National Polytechnique de Grenoble (France), 1994.

[9] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, *Physica D*, **56** (1983) 189- 208.

[10] T. Kohonen, Self-Organizing Maps. Springer Series in Information Sciences, vol. 30, Springer (Berlin), 1995.

[11] A. Choppin, Unsupervised classification of high dimensional data by means of self-organizing neural networks. M.Sc. thesis, Université catholique de Louvain (Belgium), Computer Science Dept., June 1998.

[12] R. N. Shepard, The analysis of proximities: Multidimensional scaling with an unknown distance function, parts I and II, *Psychometrika*, **27** (1962) 125-140 and 219-246.

 DOI : 10.18231/2454-9150.2018.1130

[13] R.N. Shepard, J.D. Carroll, Parametric representation of nonlinear data structures, *International Symposium on Multivariate Analysis*, P. R. Krishnaiah (ed.) pp. 561-592, Academic Press, 1965.

[14] J.W. Sammon, A nonlinear mapping algorithm for data structure analysis, *IEEE Trans. on Computers*, **C-18** (1969) 401-409.

[15] P. Demartines, J. Hérault, Curvilinear Component Analysis: a self-organizing neural network for nonlinear mapping of data sets, *IEEE Trans. on Neural Networks*,. **8**-1 (1997) 148-154.

[16] J. Lee, A. Lendasse, N. Donckers, M. Verleysen, A robust nonlinear projection method, ESANN'2000 (European Symposium on Artificial Neural Networks), Bruges (Belgium), April 2000, pp. 13-20, DFacto publications (Brussels).

[17] A. Lendasse, J. Lee, E. de Bodt, V. Wertz, M. Verleysen, Input data reduction for the prediction of financial time series, ESANN'2001 (European Symposium on Artificial Neural Networks), Bruges (Belgium), April 2001, pp. 237-244, D-Facto publications (Brussels).

[18] A.N. Refenes, A.N. Burgess, Y. Bentz, Neural networks in financial engineering: a study in methodology, *IEEE Transactions on Neural Networks*, **8**-6 (1997) 1222-1267.

[19] A.N. Burgess, Nonlinear model identification and statistical significance tests and their application in financial modeling. In *Artificial Neural Networks*, Proceedings of the Inst. Elect. Eng. Conf., 1995.

 DOI : 10.18231/2454-9150.2018.1130