INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

DIMENSIONALITY REDUCTION OF HIGH DIMENSIONAL DATASET USING CLUSTERING TECHNIQUES

[1]**Mrs. Juliet Rozario,** [2]**Mrs. S. Sasikala.,** MCA., M.Phil.,

[1]Research Scholar, Department of Computer Science, Sree Saraswathi Thyagaraja College,Pollachi.
[2]HOD, UG Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi.

Abstract: - Dimensionality reduction studies methods that effectively reduce data dimensionality for efficient data processing tasks such as pattern recognition, machine learning, text retrieval, and data mining. We introduce the field of dimensionality reduction by dividing it into two parts: feature extraction and feature selection. Feature extraction creates new features resulting from the combination of the original features; and feature selection produces a subset of the original features. Both attempt to reduce the dimensionality of a dataset in order to facilitate efficient data processing tasks.

We introduce key concepts of feature extraction and feature selection, describe some basic methods, and illustrate their applications with some practical cases. Extensive research into dimensionality reduction is being carried out for the past many decades. Even today its demand is further increasing due to important high-dimensional applications such as gene expression data, text categorization, and document indexing.

On web search providing exact result to the user is the most important task. The previous existing models and search engines are lagging with providing personalization in an exact manner. They provide results according to the ranking algorithm which it's using. Identifying the user interest and providing search result according to that is still a challenging task. We identified the problem of identifying the way of user interest prediction, where the data set or the visiting history is huge. When the dimensionality of the data increases in the web search, how we going to identify the user interest in a efficient manner.

Keyword: - Dimension, Clustering, Dataset, Feature extraction, Feature Selection.

1. INTRODUCTION

In statistics, Dimension Reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. In physics, dimension reduction is a widely discussed phenomenon, whereby a physical system exists in three dimensions, but its properties behave like those of a lower-dimensional system.

Data exploration refers to the search of structures or features that may indicate deeper relationships between variables. Normally data exploration relies heavily on visual methods because of the power of human eye to detect structures. However when the number of dimensions of the data gets very large , it becomes necessary to reduce the number of dimensions of the datasets by applying dimensionality reduction techniques. Dimensionality reduction is the search of a small set of features to describe a large set of observed dimensions. Besides the obvious fact that reducing the number of dimensions makes it easier to visualize the data. Dimensionality reduction is also useful in discovering a compact representation, thus decreasing the computational processing time. In addition, the exercise of dimensionality reduction may serve to separate the important features or variables from the less important ones, therefore providing additional insight into the nature of the dataset that might otherwise be left undiscovered.

We will review some existing statistical methods for linear and non-linear dimensionality reduction, and provide some evaluation of the advantages and disadvantages of each. First, a linear dimensionality reduction technique, Multidimensional scaling (MDS) heavily applied to solve many problems in the social and behavioral sciences will be reviewed and other techniques also. Two recently introduced non-linear methods isometric feature mapping (ISOMAP) and locally linear Embedding, which have been implemented for many image processing and pattern recognition problems is also described. Finally summary and evaluation of the algorithm is presented.

A number of methods currently exist for accomplishing this reduction. These methods are broadly grouped into linear and non-linear approaches. These approaches include principal component analysis (PCA), multi-dimensional scaling (MDS), Isomap, locally linear embedding (LLE), and most recently LDM’s method. Each of these methods seeks to find a mapping which can represent the important features of the original data in a smaller space with substantially fewer dimensions. This can be expressed mathematically as mapping the original space RD to a new space Rd where d << D.

This work evaluates the effectiveness of several methods for dimensionality reduction as they relate to two distinct text mining applications. First, how dimensionality reduction impacts the ability of standard algorithms to effectively classify documents among known categories has been studied. It was theorized that some newer dimensionality reduction methods which stress local relationships would perform best. Results from classification, however, contradict this hypothesis. Nonetheless, results did show that many DR techniques are able to reduce the data such that classification accuracy is improved when comparing against a classifier that performs no DR but uses the same number of dimensions. In addition, results showed that many of the DR techniques could produce strong accuracies when using only a small number of dimensions.

2. EXISTING METHODOLOGY

Clustering is considered an unsupervised learning process, where the main aim is to group a collection of unlabeled documents into meaningful clusters that are similar within themselves and dissimilar to documents in other clusters. Clustering documents is attractive because it frees organizations from the need of manually organize document bases, which could be too expensive, or even infeasible given the time constraints of the application and/or the number of documents involved. Machine learning algorithms used for text clustering can be categorized into two main groups (i) hierarchical clustering algorithms, and (ii) partition-based clustering algorithms.

Hierarchical clustering algorithms produce nested partitions of data by merging or splitting clusters based on the similarity among them. On the other hand, partition-based clustering algorithms group the data into non–overlapping partitions that usually locally optimize a clustering criterion. Hierarchical clustering provides good visualization capabilities especially if data is naturally exist in hierarchy. However, it lacks robustness as it is very sensitive to outliers. Additionally, the computational time of hierarchical clustering is very large which limits its usage in large data.

3. PROPOSED METHODOLOGY

The main contribution of this thesis is to enhance TC using the filter approach of DR. In order to compare the results of this work with the benchmark results in TC, the most used technique is chosen in each stage of the TC process. The following section summarizes the how the experiments of this work have been setup.

i) Document Pre-Processing

All punctuation and special characters are first removed from documents. It is worth noting that the removal of rare words was not considered, since it might be harmful especially in highly-skewed datasets. These datasets may contain categories contain a limited number of documents, and without rare words the discrimination of these categories may be difficult.

ii) Document Representation

Every document was represented as a BOW which is the simplest representation available.

iii) Dimensionality Reduction

Four feature scoring methods have been chosen to examine the performance of the thresholding techniques as well as the combining operators. These methods are the DF, IG, MI, and CC. These feature selection methods have been widely used, and have shown promising results.

iv) Feature Weighting

Classical normalized tfidf (ltc) method has been adopted since it is the most commonly used feature weighting method.

v) Classification

SVM has been the method of choice of this work. Studies have shown that it is among the best performing classifiers in TC applications.

vi) Performance Evaluation

The common MicroF1 and MacroF1 measures have been used for performance evaluation to benchmark our results with the literature.

3.1. K-Means Clustering

K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represented by their centroids, by minimizing the square error function [13]. Although K-means is simple and can be used for a wide variety of data types. The K-means algorithm is one of the partitioning based, nonhierarchical clustering methods. Given a set of numeric objects X and an integer number k, the K-means algorithm searches for a partition of X into k clusters that minimizes the within groups sum of squared errors. The K-means algorithm starts by initializing the k cluster centers [1]. The input data points are then allocated to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest. The mean (centroid) of each cluster is then computed so as to update the cluster center [1]. This update occurs as a result of the change in the membership of each cluster.

- **Step 1**: Initialization: choose randomly *K* input vectors (data points) to initialize the clusters.
- **Step 2:** Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster.
- **Step 3:** Mean update: update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster.
- **Step 4:** Stopping rule: repeat steps 2 and 3 until no more change in the value of the means.

Reduced datasets done by principal component analysis reduction method is applied to K-Means clustering. As a similarity metric, Euclidean distance has been used in k-means algorithm. The steps of the Amalgamation k-means clustering algorithm are as follows.

Phase-I: Apply PCA to Reduce the Dimension of the High Dimensional Data Set

Step 1: Organize the dataset in a matrix *X*.
Step 2: Normalize the data set using Z-score.
Step 3: Calculate the singular value decomposition of the data matrix. $X = UDV T$

Step 4: Calculate the variance using the diagonal elements of *D*.
Step 5: Sort variances in decreasing order.
Step 6: Choose the *p* principal components from *V* with largest variances.
Step 7: Form the transformation matrix *W* consisting of those *p* PCs.
Step 8: Find the reduced projected dataset *Y* in a new coordinate axis by applying *W* to *X*.

Phase-II: Find the Initial Centroids
Step 1: For a data set with dimensionality, *d*, compute the variance of data in each dimension.
Step 2: Find the column with maximum variance and call it as max and sort it in any order.
Step 3: Divide the data points of *cvmax* into *K* subsets, where *K* is the desired number of clusters.
Step 4: Find the median of each subset.
Step 5: Use the corresponding data points (vectors) for
Step 6: each median to initialize the cluster centers.

Phase-III: Apply K-Means Clustering With Reduced Datasets.
Step 1: Initialization: choose randomly *K* input vectors (data points) to initialize the clusters.
Step 2: Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster.
Step 3: Mean update: update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster.
Step 4: Stopping rule: repeat steps 2 and 3 until no more change in the value of the means.

High Dimensional Dataset is reduced using principal component analysis reduction method. Dataset consists of 569 instances and 30 attributes. Here the Sum of Squared Error (SSE), representing distances between data points and their cluster centers have used to measure the clustering quality.

The experiment conducted by [48] also illustrates that selecting 10% of features using the filter approach exhibits the same classification performance when using all the features by SVM. This is in contrast to other classifiers such as k-means in which using all the features 19 degrades the performance significantly.

4. RESULTS & DISCUSSIONS

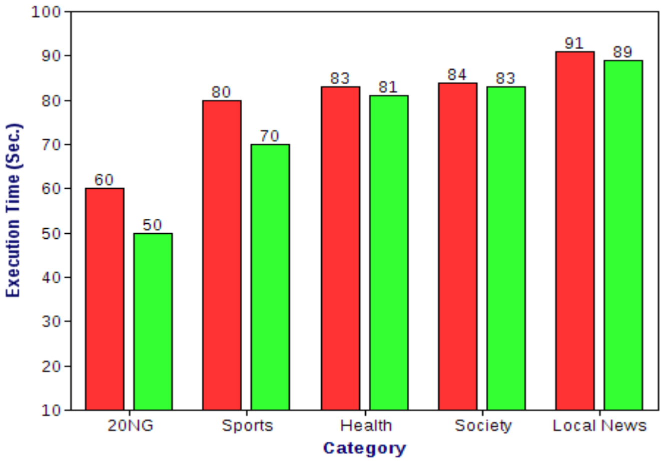
The main contribution of this paper is to enhance TC using the filter approach of DR. In order to compare the results of this work with the benchmark results in TC, the most used technique is chosen in each stage of the TC process.

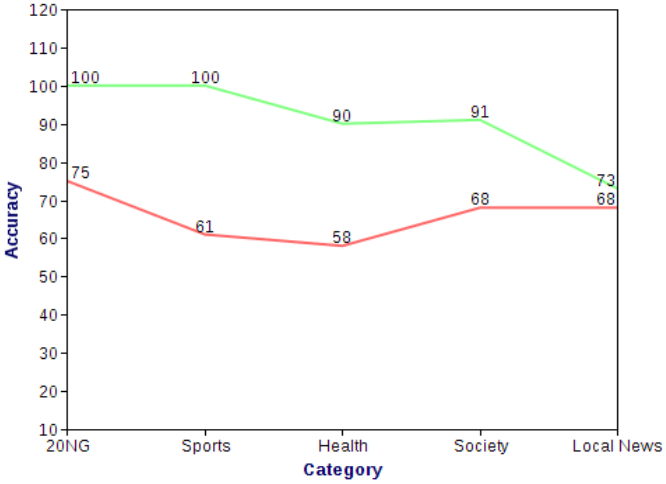
i. Recall & Precision

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. The system is evaluated in terms of precision, recall and Fmeasure. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

$$precision = \frac{number\ of\ correct\ results}{number\ of\ all\ returned\ results}$$

$$recall = \frac{number\ of\ correct\ results}{total\ number\ of\ actual\ results}$$





ii. F- Measure

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

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Several experiments were conducted with different query documents and the precision, recall and F-measure of the output was calculated. This higher improvement in precision value can compromise for the very small percentage of drop in the recall value. Moreover, the F-measure which combines precision and recall is much improved for similarity than existing system.

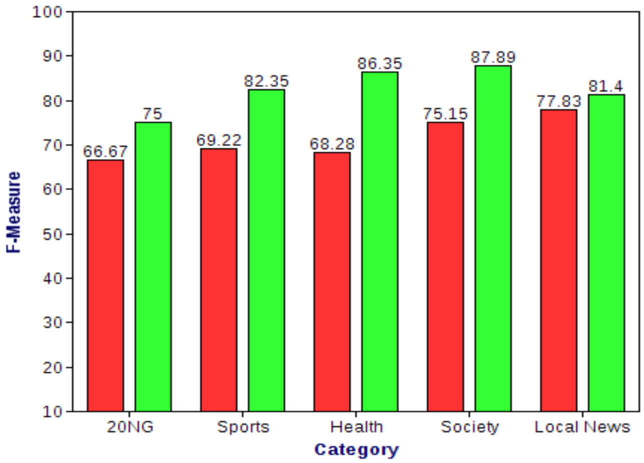


Fig 4.2: - F-Measure Result

The F-Measure represents the measure of recall and precision of retrieval or categorizing the data. In the above result, the red line represents the existing approach and the green line represents the k-means for executing the dimensionality reduction.

These measures are very helpful in evaluating the performance of both frequent and rare categories.

5. CONCLUSIONS

This main contribution of this work is to enhance existing DR techniques and second is to conduct a comparative study that allows users to make comprehensive choices among available techniques. The main objective is to achieve the highest performance with the simplest techniques. Due to the simplicity and efficiency of the feature filtering approach, this work demonstrates this approach in order to perform the DR process. The results indicate that using the principal component analysis does not lead to a significant loss in the accuracy while reducing the vocabulary size significantly. On the other hand, using the k-means approach dramatically reduces the storage requirements. However, the existing approach has shown to lead to some degradation in the performance notable in large dataset.

The proposed combining and improved the performance of the k-means clustering approach. This leads to improved categorization accuracy and a saving in the feature set size. In addition, this reduction would help reduce the storage and decrease the computational resources. The proposed systems have shown comparable results with benchmark datasets.

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.

[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.