International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol 03 Issue 01 | 2015

Performance Analysis of Data Clusterization & Classification Using Hybrid Algorithm: A Review

Pranjali N. Kubde[1], Dhananjay M. Sable[2]
1Department of CSE, Agnihotri College of Engineering, Wardha
[2]Head, Department of CSE, Agnihotri College of Engineering

ABSTRACT

This paper presents analysis of data clusterization for high dimensional data. Clustering becomes very difficult due to the increasing sparsity of data. To avoid this sparsity of data and curse of dimensionality, the hubness phenomenon is used. Hubnes is a good centrality point within a high dimensional data cluster.

Index Terms— Clustering, Curse of dimensionality, Hubs.

I. INTRODUCTION

Text Categorization system consists of various essential parts which includes feature extraction and feature selection. After preprocessing of text document, the feature extraction is used to transform the input text document into a feature set. The feature selection is applied to the feature set to reduce the dimensionality of it. Feature selection method will apply to text clustering task to improve the clustering performance and explore the possibility of hubness and k means are used for feature selection by exploiting some characteristics of the data. The objective is that it would be actively select instances with higher probabilities to be informative in determining feature relevance so as to improve the performance of feature selection without increasing the number of sampled instances. The purpose of text classification is to use the contents of a text or document to assign it to one or more categories. It has application in document organizing and management, information retrieval. Text classification is an area within pattern recognition and classification that has been studied with increasing frequency as internet usage becomes more commonplace. There are many varied practical applications of text classification. The most well known of these applications is likely improved spam filtering techniques. Search engines also take advantage of text classification techniques to return more accurate results to the user.

Feature selections form an important subset within the much large area of text classification. Correctly identifying the relevant features in a text is of vital importance to the task of classification. Other methods are used for reducing dimensionality, such as pruning clustering, can improve performance of text classification. The purpose of feature selection is to determine which features are the most relevant to the current classification task. In text classification, features are typically from a document. Choosing an appropriate feature selection method for text classification can be vital because of the large number of features usually present in text document.

In this paper, applying a combination of support vector machine, feature selection and unsupervised method along with clustering to get a better output for text classification with respect to the methods available and will perform abcomparative study on a variety of feature selection methods for text clustering, with other algorithm and finally will evaluate the performance of hybrid feature selection method based on clustering.

II. LITERATURE REVIEW

High dimensional data arises naturally in various domains and have regularly presented a great challenge for the traditional data mining techniques both in terms of effectiveness and efficiency. Clustering becomes very difficult due to the increasing sparsity of data as well as increasing difficulty in distinguishing distance between the data points. The hubness is used to avoid the curse of dimensionality in clustering by observing lower dimensional feature subspaces and therefore we use the hubness phenomenon. The tendency of high dimensional data is to contain the points or hubs that frequently occur in k nearest neighbor of other points and can be successfully exploited in clustering.The hubness is a good centrality within a high dimensional data cluster. The methods are mostly detecting the hyper spherical cluster and need to properly handle the clusters in arbitrary shapes. The algorithms are using hubness for improving high dimensional data clustering. The data repository has a high

dimensional data which makes complete search problems, to solve this problem clustering plays a very important role in handling high dimensional data and low dimensional data.

Data mining is the non-trivial process of extracting information from the very large amount of database. In recent years, data repository has a high dimensional data, which makes a complete search in the most of data mining problems leads computationally infeasible. To eradicate this problem clustering plays a vital role in handling low dimensional data and high dimensional data. Low dimensional data makes a task very simple and easy to cluster. High dimensional data is a crucial fact to cluster and it has to resolve using hubness phenomenon. Here hubness refers a data point which may frequently occur among the groups. The hub based clustering techniques are used to improve the quality of cluster in terms of effectiveness and accuracy and only avoid hyper spherical clusters. The biological dataset for clustering is more difficult because it has a very high dimensional data. To solve this problem the hub based clustering algorithms are used to automatically calculate the number of clusters from the biological data set. By using hub based clustering technique to improve the quality of cluster in terms of effectiveness and accuracy and avoid only detecting hyper spherical clusters.

Hubness is discovered the general problem of machine learning in high dimensional spaces. The hub objects have a small distance to an exceptionally large number of data points and the anti hubs are far from all the other data points. It is related to the concentration of distances which impairs the contrast of distances inspired by shared nearest neighbor approaches has been shown to reduce hubness and concentration and there already exists some work on direct application of SNN in the context of hubness in image recognition. The shared nearest neighbor is applied to a large number of high dimensional data sets from the diverse domains and compares it to the other secondary approaches. The shared nearest neighbor i.e. SNN is used to reduce the hubness but less than the other approaches. It is only able to improve the accuracy for the half of the data sets. The SNN is algorithm is used to reduce in high dimensional spaces.

Several application domains such as molecular biology and geography produces a tremendous amount of data which can be no longer to be managed without help of efficient and effective data mining methods. The primary data mining task is clustering. The traditional algorithms are failed to detect the meaningful clusters because most of the real world datasets are characterized by high dimensional inherently sparse data space. The data sets contain the clusters which are hidden in various subspaces of the original feature space. Therefore the concept of subspace clustering is used which aims at automatically identifying subspaces of the feature space in which the clusters are exist. The SUBCLUE algorithm introduced which is based on formal clustering notion and SUBCLU is able to detect arbitrarily shaped and positioned clusters in subspaces.

The density connected subspace clustering which is an efficient and effective approach to the subspace clustering problem. The density connectivity is used to efficiently prune subspaces in process of generating all clusters in a bottom up way. The SUBCLU density based clustering algorithm for detecting clusters in high dimensional data. An efficient greedy algorithm is use to compute all density connected sets hidden in subspaces of high dimensional data.

III. PROBLEM STATEMENT

- To selecting the features using feature selection.
- To designing feature selection algorithm by identifying the best features of the existing methodology using hubness phenomenon and unsupervised method with clustering.
- To designing the support vector machine algorithm for classification.
- To evaluating the result and comparing the performance of above methods with the other algorithms.

IV. RECENT RESEARCH

To overcome the above review problems, the two algorithms are used in this paper. The SVM means support vector machine and feature selection algorithm. Support vector machine has been successfully applied to solve a large number of classification problems. The support vector machines are use for classification. This algorithm calculates retrieval time, efficiency and avoids redundancy of data.

The second algorithm is feature selection algorithm. It is used for reducing the input to a manageable size for processing and analysis. This is used to select the appropriate features from the database to retrieve the data. It is use to find out data relevant features and use to improve the performance.

V. OBJECTIVES OF THE PROPOSED SYSTEM

From the above discussion, this paper analyzed that, in high dimensional data, the problems are arises due to the increasing sparsity of data and the curse of dimensionality of data, and hyper spherical clusters.
To overcome this problems, this paper proposes to select the features using feature selection algorithm and to designing the feature selection algorithm by identifying the best features of the exiting methodology using hubness phenomenon and unsupervised method with clustering and also designing the support vector machine algorithm for classification. And finally evaluating the result and comparing the performance of above methods with the other algorithm.

VI. CONCLUSION

This paper shows that efficient algorithms are used to find out data relevant features to improve the performance and used to select the appropriate features database to retrieve the data and reduce noisy ratio and use parameters i.e. accuracy and precision to calculate the performance. It is also used for reducing the input to a manageable size for processing and analysis. The algorithms are also applied to solve a large number of classification problems and calculate retrieval time, efficiency and avoid redundancy of data. This paper also gives the analytical study of different hubness algorithms.

REFERENCES

[1]. Karin Kailing, Hans-Peter Kriegel, Peer Kroger," Density-Connected Subspace Clustering for High-Dimensional Data", In Proc. 4th SIAM Int. Conf. on Data Mining, pp. 246-257, Lake Buena Vista, FL, 2004.
[2]. Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic," The Role Of Hubness In Clustering High Dimensional Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 3 MARCH 2014.
[3]. .Suganthi, S.Tamilarasi," A study on clustering high dimensional data using hubness phenomenon", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2,
[4]. Ver. VIII (Mar-Apr. 2014), PP 22-30 www.iosrjournals.org
[5]. Arthur Flexed, Dominica Schnitzer," Can Shared Nearest Neighbor Reduce Hubness in High-Dimensional Spaces?",Austrian Research Institute for Artificial Intelligence Freyung 6/6, Vienna, Austria
[6]. J. Han and M. Kamber, Data Mining: Concepts and Techniques, second ed. Morgan Kaufmann, 2006.
[7]. C.C. Aggarwal and P.S. Yu, "Finding Generalized Projected Clusters in High Dimensional Spaces," Proc. 26th ACM SIGMOD Int'l Conf. Management of Data, pp. 70-81, 2000.