ISSN:0975-9646

Spam Outlier Detection in High Dimensional Data: Ensemble Subspace Clustering Approach

Suresh S. Kapare, Bharat A. Tidke

Computer Department, Savitribai Phule Pune University
Flora Institute of Technology, Pune, Maharashtra, India

***Abstract*— High Dimensional data is need of world as social networking sites, biomedical data, sports, etc. Many data sets are represented with hundreds or thousands of dimensions. Dimensions are increasing, so due to "Curse of Dimensionality", traditional outlier detection methods not working efficiently. Increasing dimensions of data objects, makes difficult to find out points, which are not fitting in group (cluster), called Outlier. The outlier detection method has important applications in the field of fraud detection, network robustness analysis, error elimination in scientific data, sports data analysis and intrusion detection. Most such applications are high dimensional domains in which the data can contain hundreds of dimensions. Spam can be linked based or content based. Ensemble subspace clustering is paradigm in which spam outlier detection is done for high dimensional data sets is proposed in this paper. The proposed method divides original high dimensional data set in subspace clusters using subspace clustering algorithm. By using improved k-means algorithms outlier cluster is found, which is further merged with other clusters depending upon consensus function. Outlier cluster, which is not going to merge with any other subspace cluster, is called as final outlier.**

***Keywords*— Outlier, high dimensional data, subspace, ensemble, clustering.**

I. INTRODUCTION

As we all proceeding with the automation in life, use of computer is mandatory. We are trying to handle all type of processing's with the help of computers. Many fields of work like Engineering, Science, Biomedical, Agriculture, Finance, Sports, Education, Telecommunication, etc. are intended for automation. Automation includes computers. While accessing such a processes database is generated, but as all advances in fields makes these databases with multiple dimensions and these dimensions are counted in hundreds or thousands. Such a database is called as High Dimensional data. Now whenever we want to make analysis of such a high dimensional data. We need to make groups of data objects. The process of making groups is called as Clustering. As dimensions are increasing data is becoming sparse which is called as "curse of dimensionality", means by applying set of dimensions over data point if we try to make clusters, then in that region, data points densely satisfying dimensions will be crowded and some points which are satisfying less dimensions in cluster will spread in region. Such a data points are called as outliers with respect to that cluster. But it may happen that outlier of on cluster may belong to other cluster or other set of dimensions densely. So finding outliers with respect to individual cluster is not efficient. It needs to check with other clusters too. Therefore various methods of outlier detection are not working efficiently in high dimensional data sets.

II. LITURATURE SURVEY

Now a days as technology is increasing in the field of medical, animation, sports, neural, analysis, reviews, etc. Data collected from such field is not simple data. Collected data might be defined with hundreds or thousands of its dimensions. Information retrieval from such high dimensional data is actually crucial work. And finding outliers from it is challenge.

A. Outlier Detection in High Dimensional Data

Jonathan Von Brunken [1] proposed method works that how measure of intrinsic dimensionality can be used to improve outlier detection technique in data with strongly varying intrinsic dimensionality. Intrinsic Dimensionality of every data object is calculated which gives number of variables required to describe a given data set. This local Intrinsic Dimensionality gives Intrinsic Dimensionality Outlier Score (IDOS) which is used to distinguish between inliers and outliers. ID of inliers increases the agreement with remaining inliers member of subspace and weakened agreement with non-members of cluster. This helps to separate out outliers from subspace.

C. C. Aggarwal [2] proposed method work by finding lower dimensional projections which are locally sparse, and cannot be discovered easily by brute force techniques because of the number of combinations of possibilities. They proposed algorithm for outlier detection which find the outliers by studying the behaviour of projections from the data set.

B. Ensemble Subspace Clustering

Reza Ghaemi, Md. Nasir Sulaiman [4] proposed a survey on clustering ensembles techniques. They elaborated that many real world application domains such as data compression, data mining and pattern recognition need data clustering analysis. All these data has multidimensional. Therefore single clustering algorithm is not able to achieve clustering analysis. Therefore clustering ensemble technique will integrate all consensus of multiple clustering solutions into single consensus one.

The principle of ensemble is to generate a set of different models, and then aggregated them into only one. We identify the main consensus functions commonly used in the clustering ensemble. Random subspaces are an excellent source of clustering diversity that provides different views of the data. Projective clustering is an active topic in data mining. Here, however, we are only concerned with the use

of random projections for the purpose of clustering combination. Here, however, we are only concerned with the use of random projections for the purpose of clustering combination. There are some types of consensus function such as: Hypergraph Partitioning, Voting Approach, Mutual Information Algorithm, Co-association based functions and Finite Mixture model.

As mentioned earlier, most of the research on subspace clustering is focused on defining the subspace clusters and how to efficiently mine them. The clusters are information extracted from the data, but not knowledge that is useful to the users. To convert information to knowledge, post-processing of the clusters is needed.

C. Spam Detection

Importance of reviews also gives good incentive for *spam*, which contains false positive or malicious negative opinions. Internet is too much essential for modern information systems. Every fields, such as e-commerce websites becoming popularly available for people to purchase different types of products online. Online shopping is crazing now a days. Everyone wants to deal in e-commerce to purchase anything online. And whenever people are purchasing something, they first check reviews online which are given by users. If product is having good rating then it will be purchased. But some vendors or customers are providing fake/spam reviews to misled customers. These spam reviews are less as compare to dataset. Therefore detecting spam plays important role here.

Al Najada. H [5] has proposed that spam is small portion of review reports so data is becoming imbalance. This naturally leads to a data imbalance issue for training classifiers for spam review detection, where learning methods without emphasizing on minority samples (i.e., spams) may result in poor performance in detecting spam reviews (although the overall accuracy of the algorithm might be relatively high). He proposed a bagging based approach to build a number of balanced datasets, through which we can train a set of spam classifiers and use their ensemble to detect review spams.

Nitin Jindal and Bing Liu [6] proposed work by performing spam detection using two methods,

1) Duplicates Detection: There are a large number of duplicate reviews and many of them are clearly spam. For example, different user_ids posted duplicate or near duplicate reviews on the same product or different products. Duplicate detection is done using the shingle method.

2) Model Building: Here logistic regression is used by author [5]. The reason for using logistic regression given is that it produces a probability estimate that each review is a spam review. It is almost certain that in the non-spam training or test data there are spam reviews which were not duplicated. This means that the labelled non-spam data has many errors.

D. Improved K-Means

There are many clustering algorithms such as text clustering, K -Means algorithm, as one of the classic algorithms of clustering algorithms, and a textual document clustering algorithms commonly used in the analysis process, is widely used because of its simple and low complexity. As there are limitations of K-means algorithm, hence advanced K-means algorithm is proposed.

Gang Liu, Shaobin Huang [8] has implemented advanced K-means algorithm based on association rules. In this he explained improved K-means algorithm based on minimum cover set [8].

His experiment on real data set about basic old-age insurance in social security area audit methods [8]. First experiment uses the classic K-means algorithm for clustering on given data set. Author has done analysis of association rules using data mining software weka. He has also given comparison of k-means and advanced k-means algorithms. Advanced K-means algorithm i.e. K-means algorithm based on association rule obviously has enhanced purity.

Here Improved K-means algorithm is used to partition subspaces into original clusters and outlier clusters.

III. RESEARCH MODEL

As discussed in survey, we get to know that how outlier detection is very important from high dimensional datasets for analysis purpose. By considering this challenge, a model for outlier detection is proposed. In this model high dimensional data set is provided as an input data. This input subspaces will be divided in data clusters by using subspace clustering algorithm. Then by using Improved K-means algorithm data points with dense dimensionality (original cluster) and sparse points called outlier cluster will be separated for every subspace. Now there are chances that data points present outlier of one subspace may present in original cluster of other subspace. Therefore real data can be removed from datasets. This is not expected in data analysis. So these outliers have to add in appropriate clusters of subspace. To achieve this again outlier cluster is of current subspace is compared with other subspaces and if it satisfies the dimensionality then added to that cluster. These steps are repeated until getting final Outlier Cluster. Complete research work is explained with the help of Figure [1].

A. Subspace Clustering

High dimensional data set is divided into subspaces and from each subspace original data cluster and outlier cluster is separated. For this task advanced K-means algorithm is implemented after subspace clustering. Data is sampled, then selected a set of k-medoids and repeatedly improved the clustering [2]. Medoids are the point at centre of region. This algorithm works in three steps: initialization, iteration, and cluster refinement. Initialization works on greedy approach to select a set of potential medoids, which must present at some distance. The iteration phase choose a random set of k-medoids from reduced dataset and replaces bad medoids with randomly chosen new medoids, and determines if cluster has improved. Cluster quality is based on the average distance between instances and the nearest medoids. The total number of dimensions associated to medoids must be k*l, where l is an input parameter that selects the average dimensionality of cluster subspaces [2]. Once the subspaces have chosen for each medoids, average

segmental distance is used to assign points to medoids, forming clusters. The refinement phase calculates new dimensions for every medoid based on the clusters produced and reassigns points to medoids, removing outliers. While clusters may be found in different subspaces, the subspaces must be of similar sizes since the user must input the average number of dimensions for the clusters. Clusters are representing as sets of instances with associated medoids and subspaces and form non-overlapping partitions of the dataset with possible outliers. Due to the use of sampling, PROCLUS is somewhat faster than CLIQUE on large dataset. Here Improved K-means algorithm is used to partition subspaces into original clusters and outlier clusters.

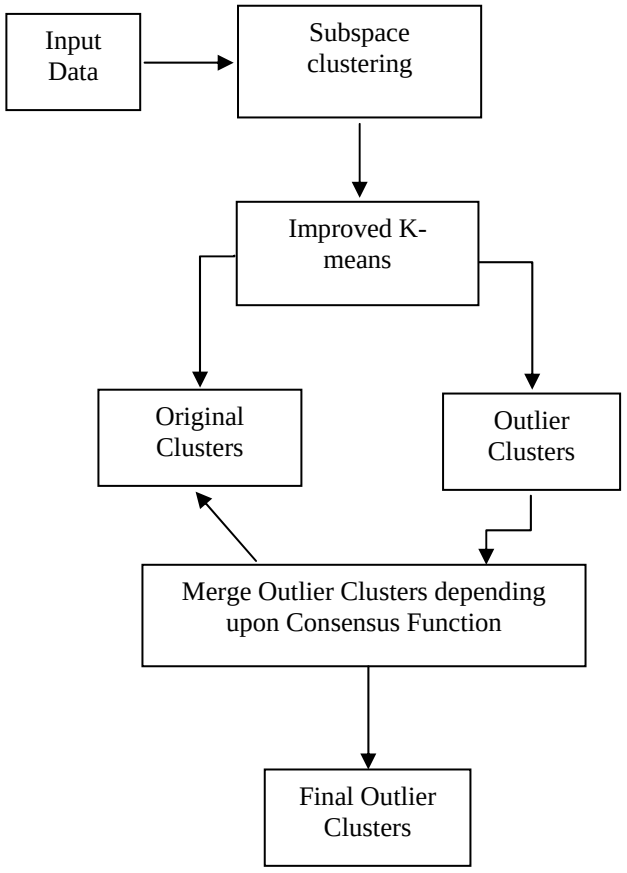

Fig. 1 Spam Outlier Detection Model

B. Merge Outlier Cluster

After partitioning subspace into original cluster and outlier cluster we are getting every subspace in to group of original clusters and one outlier cluster. Now suppose we have four subspaces partitioned in subspace clustering and by improved K-means again each subspace is divided in three original cluster and one outlier cluster. If we did not consider this scenario then might be original data consider as outlier. So finally it is important to merge outlier clusters to subspaces depending upon consensus function called ensemble [3]. The principle of ensemble is to generate a set of different models, and then aggregated them into only one. We identify the main consensus functions commonly used in the clustering ensemble. Random subspaces are an excellent source of clustering diversity that provides different views of the data. Projective clustering is an active topic in data mining. Here, however, we are only concerned with the use of random projections for the purpose of clustering combination. There are some types of consensus function such as: Hypergraph Partitioning, Voting Approach, Mutual Information Algorithm, Co-association based functions and Finite Mixture model [11].

In this way implementation can proceed. Most important thing in this is, Ensemble subspace clustering. This merges outlier detected in subspace to other. So no original data will consider as outlier.

IV. Mathematical Model

Let D is set of n high dimensional data sets of reviews.

$$\therefore D = \{d1, d2, d3, \ldots\ldots dn\} \ldots\ldots\ldots\ldots raw\ data$$

As high dimensional data, it has hundreds to thousands attributes.

$$Datt = \{a1, a2, a3, \ldots\ldots\ldots\ldots an\}$$

By applying subspace clustering algorithm, D is getting divided into subspaces, let 4 subspaces are created .i.e.

$D = SP1 \cup SP2 \cup SP3 \cup SP4$ and $SP1 \cap SP2 \cap SP3 \cap SP4 = \emptyset$

SP is defined as,

$$SP = \{x | x \in D \ \&\& \ D \xrightarrow{a+} x\}$$

Where a^+ = applying more than one attributes

Now consider single subspace SP1.

Applying improved K-means clustering algorithm,

If c1,c2, c3 are three centroids of subspace, as 3-means clustering is considered.

Let SP1={d1,d2,d3,…..,dx}

Here x data sets partitioned in sub spacing to SP1.

Now, calculating length of all data sets from centroids, and selecting datasets having length till m.

m =thresh hold.

Now, we get three clusters CL1, CL2 & CL3.

Where,

$$Cl1 = \{dn | l(c1, dn) \leq l(c2, dn), l(c3, dn)\}$$
$$Cl2 = \{dn | l(c2, dn) \leq l(c1, dn), l(c3, dn)\}$$
$$Cl3 = \{dn | l(c3, dn) \leq l(c2, dn), l(c1, dn)\}$$

Where n= 0, 1, 2… x;

We considered bounding values till m.

If $l(c1, dn) > m$

Then that data point is called *outlier point*.

$$\therefore OT1 = \{dn | l(c1, dn), l(c2, dn), l(c3, dn) > m\}$$

Set OT1 is called as set of outlier points from subspace SP1.

In this way OT is separated from subspace.

V. Experimental Setup

For experiment purpose taken a database of reviews. These data sets are given below.

Amazon Product Review Data (more than 5.8 million reviews) used for opinion spam (fake review) detection. It can also use it for sentiment analysis. It has information about reviewers, review texts, ratings, product info, etc. Due to the large file size, it may need to use Download Accelerator Plus (DAP) to download. It is used in [7].

VI. Conclusion And Fututre Work

This paper explores literature survey over many concepts of high dimensional data mining, information retrieval, outlier detection in high dimensional data, ensemble subspace clustering, spam detection, improved k-means algorithm based on association rules. As High Dimensional data is need of information systems so all these concepts can be used for improvement in data mining.

All these approaches are helpful for designing many healthy applications for information retrieval. One application can be Spam Outlier Detection using Ensemble subspace clustering. In which spam outliers in review dataset of e-commerce can be detected. In this subspace clustering can be done followed by outlier detection and again ensemble with other subspaces for great accuracy. In advance if we add spam detection logic, then there will not be any issue for fraud reviews by someone. Whatever clusters are identified as a outlier cluster from high dimensional data sets, these can be highlighted or in some cases make some legal necessary actions against all these individuals. Second is we can implement rejection logic in datasets, so that while performing data analysis when outliers are detected initially, if coming data is belonging to same dimension set will be rejected form adding it to the database.

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

References

[1] Jonathan von Brünken, Michael E. Houle, and Arthur Zimek, “Intrinsic Dimensional Outlier Detection in High-Dimensional Data”, NII Technical Report, NII-2015-003E, Mar. 2015.

[2] Charu C. Aggarwal IBM T. J. Watson Research Center, “ HIGH-DIMENSIONAL OUTLIER DETECTION: THE SUBSPACE METHOD,” from: C. Aggarwal. Outlier Analysis, Chapter 5, Springer, 2013

[3] Imran Khan, Joshua Zhexue Huang, “Ensemble Clustering of High Dimensional Data with FastMap Projection”, Springer International Publishing Switzerland, W.-C. Peng et al. (Eds): PAKDD2014 workshops, LNAI 8643, pp. 483-493, 2014.

[4] Reza Ghaemi, Md. Nasir Sulaiman, “A survey: Clustering Ensembles Techniques”, World Academy of science, Engineering & Technology, Vol 3: 2009-02-25.

[5] Al Najada, H., Xingquan Zhu, “iSRD: Spam review detection with imbalanced data distributions”, Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on 13-15 Aug. 2014, DOI: 10.1109/IRI.2014.7051938.

[6] Nitin Jindal and Bing Liu, “Reviw Spam Detection”, WWW 2007, May 8-12, 2007, Banff, Alberta, Canada, ACM 978-1-59593-654-7/07/0005

[7] B.A.Tidke* et al [IJESAT] International Journal of Engineering Science & Advanced Technology, ISSN: 2250-3676, Volume-2, Issue-3, 645-651,june-2012.

[8] Gang Liu, Shaobin Huang, Caixia Lu, and Yudan Du,” An improved K-Means Algorithm Based on Association Rules”, International Journal of Computer Theory and Engineering, Vol.6, No. 2, April 2014.

[9] Chunfei Zhang*, Zhiyi Fang*, “An Improved K-means Clustering Algorithm”, JICS 10; 1 (2013) 193-199.

[10] Madhu Yedla, Srinivasa Rao Pathakota, “Enhancing K-means Clustering Algorithm with Improved Initial center”, [IJCSIT], Vol.1(2), 2010,121-125.

[11] Blaise Hanczar and Mohamed Nadif,”Study of consensus functions in context of ensemble methods for biclustering”, LIPADE, 2 april 2013