ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A REVIEW PAPER ON IMPROVED K-MEANS TECHNIQUE FOR OUTLIER DETECTION IN HIGH DIMENSIONAL DATASET

Mr. Pushpendra Bhatt.[*], Prof. Tidake Bharat

[*1] Student Pursuing M.E. in Computer Networks from FIT,Khopi,Pune, India

[2]P.G. Co-ordinator at Department of Computer Engineering, FIT,Khopi,Pune, India

ABSTRACT

In many data mining application domain outlier detection is an important task, it can be regard as a binary asymmetric or unbalanced classification of pattern where one class has higher cardinality than the other, finding outlier is very challenging in high dimensional dataset where data contain large amount of noise which causes effectiveness problem, they are more useful based on their diagnosis of data characteristics which deviate significantly from average, this paper present Improved K-Means Technique for Outlier Detection in High Dimensional Dataset. Various subspace based method has been proposed for searching abnormal saprse density unit in subspace, this paper proposes a Clique density based clustering algorithm that attempt to deal with subspace that create dense reason when projected onto lower subspace in high dimensional data set and then apply the improved K-Means algorithm on generated subspaces for effectively and efficiently identifying outliers for getting the more meaningful and interpretable result.

KEYWORDS: Outlier Detection, Improved K-Means, Subspace, High dimension.

INTRODUCTION

An observation can be outlier which appears to be inconsistent with the rest of the data set based on some measure [15]; in many data mining application domain outlier detection is an important task, one of the application is detecting measurement errors by removing the outlier as form noise from the dataset as opposed to Clustering, Clustering is concerned with grouping together sets that are similar to each other and dissimilar to the sets belonging to other clusters. Cluster is used to group items that seem to fall naturally together, outlier are those deviated sets which shows different characteristics and does not fit to any category of data set, in very huge datasets detecting outliers has attract much attention over the past several years in data mining.

Most of the Application in the field of fraud detection for credit card, military surveillance for enemy activity, in cyber security as intrusion detection most of the application are high dimensional in which data contains hundreds of dimension. The finding of meaningful outlier in high dimension has substantially more complex.

This paper makes the contribution as follows: The proposed approach aims to find subspace by applying the existing Clique method high dimensional dataset and then by applying the improved K-Means algorithm for effectively and efficiently identifying outlier in generated subspaces. The rest of the paper organized as follows Section 2. Motivation Section 3 Background study Section 4. Literature survey section 5 Design of the System and finally section 6. Conclusion.

MOTIVATION

To Outlier is very challenging in high dimensional dataset where data contain large amount of noise which causes effectiveness problem, the sparsity of high dimension data implies that every point is an almost equally good outlier more complex to find, it is very difficult to introduce the general behavior or a normal region and imprecise boundary between normal and outlier behavior since at time outlier observation lying close to the boundary could actually be normal, and vice versa. Noise in the data tends similar to outlier difficult to differentiate and remove.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

RELATED WORK AND LITERATURE SURVEY

Barnett and Lewis define an outlier as [12]:"An outlier is an observation or subset of observation which appear to inconsistence with the remainder of data set". Another definition given by Hawkins is as [13]:" An observation which deviates so much with other observations as to arouse suspicious that generated by different mechanism.

Knorr and Ng [8], introduced distance based method that computes the distances of every point of data to its neighborhood to determine whether it is outlier or not. It can define as follows: An object K in a dataset S is a distance based outlier with parameter P and Q i.e. DB (P, Q). If at least a fraction P of the object in S lie distance greater than Q from K.

Zhang et al. [9] proposed LDOF determines the degree deviation from its neighborhood, and takes over all complexity $O(N_2)$ for calculating LDOF of all point in the data set. Where N is the number of point in dataset, more deviation of point from neighbors has high LDOF value and probably it may be an outlier, the LDOF factor can be calculated as [9]:

$$LDOF(p) = dp/DP \quad (1)$$

Where

LDOF(p): The local distanced based outlier factor of p, dp: kNN distance of p and Dp: kNN inner distance of p. Let N_p be the set of k-nearest neighbors of object p excluding p. the k-nearest neighbors of distance of p equals to the average distance from p to all object in N_p. let dist(p, q) $\geq$ 0 be a distance measure between object p and q, then the k-nearest neighbors of distance of object p is defined as:

$$Dp=1/k \sum dist(p,q) \quad (2)$$

Dp: Given N_p of object p, k-nearest neighbors inner distance of p is defined as:

$$Dp=1/ k^{2-k}\sum dist(p, q) \quad (3)$$

Breuning et al.[5], proposed LOF Technique based on the local density of given sample's neighborhood to identify local outlier, the LOF mines outlier that deviate from their cluster belongs to, the LOF may not be effective in density with sparse neighbors and fails to calculate outlier when neighbors have similar density.

K-means Clustering Algorithm [7]: - K-means algorithm discovers K non overlapping cluster by finding K centroids and then assigning each point to the cluster associated with its nearest centroid.

Algorithm: The k-means clustering algorithm

Input:

$D = \{d_1, d_2,.....,d_n\}$

D is *n* data items set.

k = No. of desired clusters

Output:

A set of *k* clusters.

Steps:

1. Select *k* data-items from set D as initial centroid;

2. Repeat

Assign each item d_i to the cluster which has the closest centroid;

Calculate new mean for each cluster;

Until convergence criteria is met.

Aggarwal and Yu[14],proposed Subspace method to identify outliers by searching sparse density units in subspace, this is the grid based method in which search process start from one dimensional projection to higher dimension but it is hard to find subset of dimension in most negative saprsity coefficient, to overcome this problem Jinsong Leng[2], proposed Novel method that identify outlier in interesting subspace with tight cluster by using subspace based clustering algorithm.

LITERATURE SURVEY

Multi- Yaling Pei et al.[3] proposed an efficient referenced based method that uses the relative degree of density with respect to a fixed set of reference point to calculate the neighborhood density of a data point this method performs better in identifying local outlier that deviate from the given dataset also this method evaluate data clustering and outlier analysis by developing a synthetic data generation system that can produced datasets with various cluster and outlier pattern, referenced based method reduced the number of distance evaluations, the running time of referenced based algorithm is O(Rn log n) where n is the size of dataset and R is the number of reference point this method is effective, efficient and highly scalable to very large dataset.

Jinsong Leng [2], proposed a novel subspace outlier detection approach in high dimensional datasets for finding outlier in interesting subspace, this approach aim to group the meaningful subspace and then identifies outlier in projected subspace with tight cluster, to analyze correlation among the dimension of the dataset this method introduced entropy and

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

joint entropy measures, in this method interesting subspace are ranked by goodness of clustering to calculate the z-score in the limited subspace using distance based and density based algorithm a novel subspace outlier detection approach describes the criteria for measuring the degree of correlation among dimensions, in this aspect the problem turn to investigate the group in subspaces.

Arthur Zimek et al. [1] proposed Unsupervised Ensembles Subsampling method for inducing diversity among individual outlier detector, subsample can renders performance gain not only more robust but can improve the performance even further, Subsampling is flexible, fundamental and does not rely on specific types of data.

Fabrizio Angiulli et al. [6] proposed Detection and Prediction Method that enhances the state of distance-based outlier detection research; a subset of the data set is used to predict new unseen objects. The false positive rate is negligible, and ROC analysis uses the solving set instead of the data set for getting accuracy, but at a lower computational cost.

DESIGN

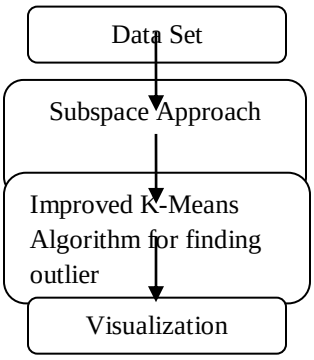

Fig 1: System Architecture for detecting outlier in high dimension dataset.

Data Set: - Let D = {d_1, d_2,.....,d_n}, where D is set of *n* data items

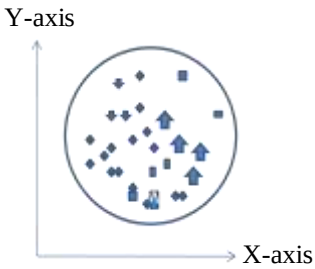

Fig 2: A simple view of two dimensional Data sets

Subspace Approach: - Various subspace based method has been proposed for searching abnormal saprse density unit in subspace, this paper proposes a Clique density based clustering algorithm that attempt to deal with subspace that create dense reason when projected onto lower subspace.

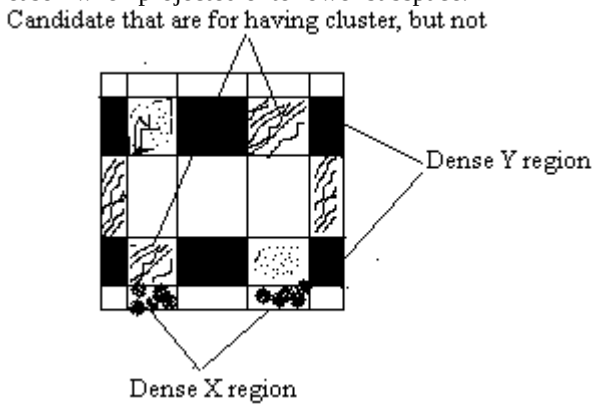

Fig 3: Illustration that density in high dimension implies density in low dimension not vice-versa.

Improved K-Means Algorithm [16]: -

Algorithm: The Improved K-Means clustering algorithm

Input:
D = {d_1, d_2,.....,d_n}
D is *n* data items set.
k = No. of desired clusters
j= Id of the closest cluster
Output:
A set of *k* clusters.
Steps:
1. Set m = 1;
2. in the set D Compute the distance between each data point and all other data points;
3. Find the closest pair of data points from the set D; Delete those point in data set which form a data-point set Am (1<= m <= k);
4. Find the data point in D that is closest to the data point set Am, Add it to Am and delete it from D;
5. Repeat step 4 until the number of data points in Am reaches 3/4*(n/k);
6. If m<k, then m = m+1, find another pair of data points from D between which the distance is the shortest, form another data point set Am and delete them from D, Go to step 4;
7. In Am find the arithmetic mean of the vectors of data points for each data point set Am (1<=m<=k), these means will be the initial centroids.
8. Compute the distance of each data-point *di* (1<=i<=n) to all the centroids *cj* (1<=j<=k) as *d(di, cj)*;

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

9. Find the closest centroid *cj* and assign *di* to cluster *j*, for each data-point *di*
10. Set ClusterId[i]=j;
11. Set Nearest_Dist[i] = *d(di, cj)*;
12. Recalculate the centroids, for each cluster *j* (1<=j<=k);
13**. Repeat**
14. for each data-point *di*,
14.1 Compute its distance from the centroid of the present nearest cluster;
14.2 If this distance is less than or equal to the present nearest distance, the data-point stays in the cluster;
Else
14.2.1 For every centroid *cj* (1<=j<=k)
Compute the distance *d(di, cj)*;
End for;
14.2.2 Assign the data-point *di* to the cluster with the nearest centroid *cj*
14.2.3 Set ClusterId[i]=j;
14.2.4 Set Nearest_Dist[i]= *d(di, cj)*;
End for;
15. For each cluster *j* (1<=j<=k), recalculate the centroids;
Until the convergence criteria is met.
Visualization: - In Dataset by applying the Subspace Approach the subspaces generated from the dataset and hence by applying the Improved K-Means Algorithm [16] it generates the desired outlier as output.

CONCLUSION

On This paper explored the Improved K-Means Technique for Outlier Detection in High Dimensional Dataset there are two steps behind this method first is the subspace outlier detection method applied for high dimensional dataset that generate the subspaces of the dataset, second the Improved K-Means Algorithm is applied in the generated subspaces for getting the more meaningful and interpretable result in high dimensional data set.

REFERENCES

1. Arthur Zimek et al.” Subsampling for Efficient and Effective Unsupervised Outlier Detection Ensembles”, ACM 978-1-4503-2174-7/13/0, KDD’13, August 11–14, 2013
2. Jinsong Leng” A novel subspace outlier detection approach in high dimensional datasets” in International Conference on Computer and Electrical Engineering ,978-1-4244-7224-6, 2010.
3. Yaling Pei et al. “An efficient referenced based approach to outlier detection in Large Datasets” Sixth international Conference on Data Mining, 2006
4. Fabrizio Angiulli et al. “Distance-Based Detection and Prediction of Outliers” IEEE transactions on knowledge and data engineering, vol. 18, no. 2, February 2006.
5. Breuning et al.“LOF Identifying Local Density Based Outlier” In procedings of the 2000 ACM SIGMOD International Conference on management of data pages 93-104, 2000.
6. Fabrizio Angiulli et al.” Distance-Based Detection and Prediction of Outliers”, IEEE transactions on knowledge and data engineering, vol. 18, no. 2, February 2006.
7. Margaret H. Dunham, *Data Mining-Introductory and Advanced Concepts*, Pearson Education, 2006
8. M. Knorr R.T and V Tucakov “Distance based outlier: algorithm and application”, The VLDB Journal, 8(3-4): 237-253, 2000.
9. K. Zhang, M. Hutter, and H.jin “A new local distance based outlier detection approach for scattered real-world data” In PAKDD ‘09: Proceeding of the 13 th pacific-Asia Conference Advanced in Knowledge Discovery and Data Mining, pages 813-822, 2009.
10. Xuan Hong Dang, Raymond T. Ng, and Arthur Zimek, Erich Schubert” Discriminative Features for Identifying and Interpreting Outliers” in 2014
11. Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel, “Generalized Outlier Detection with Flexible Kernel Density Estimates” 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA, 2014
12. V. Barnett and T. Lewis, “Outlier in Statistical Data John Wiley and Sons, 1994.
13. D. Hawkins, “Identification of outlier Chapman and Hall London, 1980.
14. Aggarwal & Yu,” Outlier detection for high dimension” in proceeding of the 2001 ACM SIGMOD International Conference on management Data pages 37-46, 2001
15. V. Barnett and T. Lewis,”Outliers in Statistical Data” John Wiley&Sons, 3rd edition, 1994
16. K. A. Abdul Nazeer, M. P. Sebastian,” Improving the Accuracy and Efficiency of the k-means Clustering Algorithm”, Proceedings of the World Congress on

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

Table No.1 Comparison of various proposed Systems in Outlier Detection

Proposed System	Authors	Year	Limitation	Advantage
Subsampling for Efficient and Effective Unsupervised Ensembles	Arthur Zimek et al. [1]	2013	Building ensembles	Rely on no specific data types.
Novel Subspace outlier detection	Jinsong Leng[2]	2010	Criterion Formulation	Interesting subspace
LOF Distance based outlier detection	Breuning et al. [5]	2000	Sparse neighbors	Mines deviated outliers
LDOF- Local Distance based Outlier detection	K. Zhang et al. [9]	2009	Computationally expensive	Determinesdegree of Object.
LOGP-Local Outliers with Graph Projection	Xuan Hong Dang et al[10]	2014	performance improvement	Ranking list and Discriminative feature.
KDEOS	Erich Schubert et al. [11]	2014	local concentrations	Easily adjustment
Outlier detection for high dimension	Aggarwal & Yu [14]	2001	Negative sparsity coefficient.	Avoids intensive computation