A MIXTURE MODEL OF HUBNESS AND PCA FOR DETECTION OF PROJECTED OUTLIERS

[1]Kamal Malik and [2]Harsh Sadawarti

[1]Research Scholar PTU and [2]Director RIMT,Punjab Technical University,India

ABSTRACT

With the Advancement of time and technology, Outlier Mining methodologies help to sift through the large amount of interesting data patterns and winnows the malicious data entering in any field of concern. It has become indispensible to build not only a robust and a generalised model for anomaly detection but also to dress the same model with extra features like utmost accuracy and precision. Although the K-means algorithm is one of the most popular, unsupervised, unique and the easiest clustering algorithm, yet it can be used to dovetail PCA with hubness and the robust model formed from Guassian Mixture to build a very generalised and a robust anomaly detection system. A major loophole of the K-means algorithm is its constant attempt to find the local minima and result in a cluster that leads to ambiguity. In this paper, an attempt has done to combine K-means algorithm with PCA technique that results in the formation of more closely centred clusters that work more accurately with K-means algorithm .This combination not only provides the great boost to the detection of outliers but also enhances its accuracy and precision.

KEYWORDS

PCA, Hubness, Clustering, HDD, Projected outliers

1.INTRODUCTION

Outlier or anomaly detection refers to the task of identifying the deviated patterns that do not give confirmation to the established regular behaviour [1].Although there are no rigid mathematical definitions of the outliers, yet they have a very strong practical applicability [2]. They are widely used in many applications like intrusion, fraud detection, medical diagnosis where the crucial and actionable information is required. The detection of outliers can be categorised into supervised, semi-supervised and unsupervised depending upon existence of labels for the outliers and existing regular instances. To obtain the accurate and representative labels are comparatively very expensive and difficult to obtain, hence among these three, unsupervised methods are usually applied. Distance-based methods are one of the very important methods of unsupervised approach that usually rely on the measure of distance or similarity in order to detect the deviated observations. A common and a well known problem that usually arises in higher dimensional data due to "curse of dimensionality" is that the distance becomes meaningless [3] since the distance measures concentrates i.e., pairwise data becomes equidistant from each other as the dimensionality enhances. Hence the every point in higher dimensional space becomes neutrally an outlier, it is quiet challenging. According to the nature of the data processing, outlier detection methods could be categorised as either Supervised or Unsupervised. The former is based on supervised classification method that needs the availability of ground truth in order to derive a suitable training set for the learning process of the classifiers. The later approach does not have the training data available with it and hence do not confirm any generalised pattern.

DOI : 10.5121/ijitmc.2015.3401

Hubness phenomenon consists of an observation that for increasing dimensionality of the data set, the distribution of the no of the times the data points occur around the k-nearest neighbours of the other data points becomes increasingly skewed to the right. To effectively and efficiently detect the projected outliers from the higher dimensional data is itself a great challenge for the traditional data mining techniques. In this paper, instead of attempting to avoid the curse of dimensionality by highlighting the lower dimensional feature subspace, the dimensionality is embraced by taking the advantage of some inherent higher dimensional phenomenon of hubness. Hubness is the tendency of the higher dimensional data to contain hubs that frequently occur in the k-nearest list of the other points and can be successfully exploited in the clustering.

Our Motivation is based on the following factors:

- The notion that the distance becomes indiscernible as the dimensionality of the data enhances and each point becomes equally a good outlier in higher dimensional data space.
- Moreover, the hubness can be termed as a good metric of a centralised point in higher dimensional data clusters and the major hubs can be effectively used as cluster prototypes while searching the centroid based cluster configuration.
- Our Contribution in this paper is:
- First of all, the various challenges of higher dimensionality in the case of unsupervised learning are discussed which includes curse of dimensionality, Attribute relevant analysis.
- The phenomenon of hubness is discussed which has a great impact on the reverse nearest neighbourhood counts.
- Projected outliers are highlightened by linking the K-means algorithm with PCA and the hybrid model formed Gaussian mixture.

2.RELATED WORKS

In the past, many outlier detection methodologies have been proposed [4] [5] [6] [7] [8][9][10][11][12]. The outlier methodologies are broadly categorised into distribution based, distance based and density based methods. Statistical based methods usually follow the underlying assumptions of the data and this type of approach aims to find the outliers which deviate from such distributions. Most of the statistical distribution models are univariate, as the multivariate distributions lack the robustness. The solutions of the statistical models suffer from noise present in the data as the assumptions or the prior knowledge of the data distribution is not easily determined for the practical problems. In case of the distance based methods the distance between each point of interest and its neighbours are calculated. This distance is compared with the predetermined threshold and if the data points lie beyond the threshold, then those points are termed as outliers, otherwise they are considered as the normal points[13]. In case of the multiclustered structured data, the data distribution is quiet complex as no prior knowledge of the data distribution is needed. In such cases, improper neighbours are determined which enhances the false positive rate.

To alleviate this problem, the density based methods are proposed. LOF is one of the important density based methods to measure the outlierness of each data instance as LOF not only determines the outliers rather highlight the degree of outlierness, which provide the suspicious ranking scores for all the samples. Although it identifies the outliers in the local data structure via density estimation and also awares the user of the outliers that are sheltered under the global data structures, yet the estimation of the local density for each instance is computationally expensive, usually when the data size is very large. Apart from the above work, there are other outlier detection approaches that are recently proposed [6][9][10] .Among them, ABOD i.e., angle based outlier detection is the one which is very crucial and unique as it calculates the variation of the angles between each target instance and the remaining data points and it’s usually observed that the outliers always produce a smaller angle variance than the normal one. Then further it’s extension named as fast ABOD was proposed to generate the approximation of the original ABOD solution. K-means algorithm is an important clustering algorithm to cluster n-objects based on the attributes into k-partitions where k<n. It is quiet similar to the expectation-maximisation algorithm for the different Gaussian mixture in a manner that they both attempt to find the centres of the natural clusters in the data . Moreover, it assumes that the object attribute forms a vector space.A major drawback of the K-means algorithm is its constant attempt to find out the local minima and results in a cluster that leads to ambiguity. In this paper, we have tried to mortise the K-means algorithm and Robust Gaussian model with Principal Component analysis and hubness to enhance their effectiveness, efficiency and precision.

This section describes related works on clustering higher dimensional data. Due to the implicit sparsity of the points, HDD has become an emerging challenge for the clustering algorithms [14]. In [14], they provided a complete survey of a usual concept of the projected subspaces for searching the projected clusters. To eliminate the sparse datasets for every cluster, and further adjusting the points into the subspaces in which the most likelihood occurs is a much generalised idea of clustering in full dimensional subspaces. This technique is used in [15] for the effective HDD visualisation. SUBCLU [16] (density-connected Subspaces Clustering) is an indispensible approach to the subspace clustering problem. In [16], the various paradigms in a usual framework are used systematically. Projected clustering is a usual data mining task for unmanned grouping object [17]. In paper [18], they presented a detailed probabilistic approach to k-nearest neighbour classification. The feasibility of incorporating the hubness data for Bayesian class prediction is examined in [15].In [16], it is shown that the nearest neighbour methods can be further enhanced by taking a point in the hubness. In HDD, it’s very difficult to handle the ordinary machine learning algorithms, which particularly characterize the problem of curse of dimensionality. In [19],they provided better assured proposed labels by exposing the fuzzy nearest neighbour classification and they also enlarged the already existing crisp hubness based approach into a fuzzy counterpart. Moreover, hybrid fuzzy functions are available and tested in detail in [19].

3. CHALLENGES IN HIGHER DIMENSIONAL DATA

3.1 CURSE OF DIMENSIONALITY

In higher dimensional space, unsupervised methods detect every point equally a good outlier because the distance becomes indiscernible as the dimensionality increases. All the data points become equidistant from each other showing that all of them carry useful and important information. Due to the sparsity of the data, the outliers are hidden in the lower dimensional subspaces and are usually do not prominently highlighted while dealing with independent components.

3.2 ATTRIBUTE RELEVANT ANALYSIS

In case of the higher dimensional clustering, the relevant attributes contain the projected clusters and the irrelevant ones contain outliers or noise [20]. Cluster structure is the region with the higher density of the points than its surroundings. Such dense regions represent the 1 dimensional projection of the cluster. So, by detecting the dense regions in each dimension, it becomes easier to differentiate between the dimensions of the relevant and irrelevant clusters. The dense regions can be distinguished from the sparse one using the predefined density threshold, but in such a case, this value of density threshold may affect the accuracy and the goodness of the cluster.

4. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is one of the very important techniques of dimensionality reduction where “respecting structure” means “variance preservation”. PCA has already been rediscovered by so many researchers in so many fields and is basically known as a method of empirical orthogonal functions, and singular value decomposition and determines the principal directions of the data distributions. To obtain these Principal directions, one needs to construct the data covariance matrix and calculate its dominant eigenvectors [23]. Among the vectors in the original data space, these eigenvectors are the most informative so are considered as the principal directions. If we want to summarize p-dimensional feature vectors into a q dimensional feature subspace, then the principal components are the projections of the original vectors onto q directions which spans the subspace. Mathematically, although there can be various different ways of deriving the principal components yet the most prominent one is to find out the projections whose variance is maximum. The first principal component will be taken in the feature space where the projections have the largest variance and the second one is taken in the direction which maximises variance among all the directions orthogonal to the first. In a generalized way, the kth component is the variance maximizing direction orthogonal to the k-1th previous component. The variance can be maximised using equation

$$\sigma^2 = 1/n \sum_i (x_{i.} w) 2 \qquad (1)$$

Where n data vectors can be stacked into an n*p matrix ,X is the projection and is given by X_w which is an n*1 matrix. Then,

$$\sigma^2 = 1/n\ (Xw)^T\ (Xw) \qquad (2)$$

$$\sigma^2 = 1/n\ w^T X^T Xw \qquad (3)$$

$$\sigma^2 = wT\ ((X^T\ X)/n)\ w \qquad (4)$$

$$\sigma^2 = w^T Vw \qquad (5)$$

The unit vectors are need to taken into consideration and we need to constrain the maximisation.

5. HUBNESS

Hubness is usually treated as a metric to detect the local centrality which further enhances the clustering in an efficient and an effective way keeping the notion in a mind that the hubs are located near the centres of compact sub clusters in higher dimensional data which is an obvious way to check the feasibility of using them to calculate their centres and compare the hub based approach with some centroid based technique. That’s why we are using K-means algorithm with iterative approach to define the clusters around separated high hubness data elements The mediods and centroids in K-means tend to converge the hubness data elements leading to the promising subspaces in the in the large datasets. Mediods and centroids usually depend upon the current cluster elements whereas hubs carry locally centralized information as they depend on their neighbouring elements. Acc to [22], two types of hubness are there i.e., local hubness and global hubness where local hubness restricts the global hubness on any given cluster.

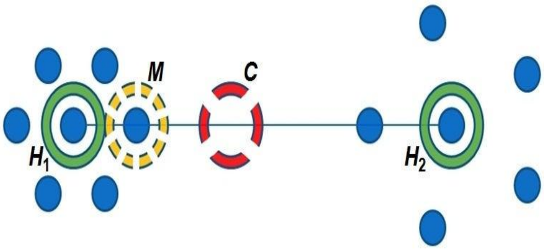

Fig 1. Hubness based Clustering [22]

6. PROPOSED METHODOLOGY

To validate the centroids formed from K-means clustering algorithm, the hubness based clustering technique is employed. Hubness can be considered as local centrality measure. It is usually considered that the hubs are located near the centres of the compact sub clusters in the higher dimensional data to test the feasibility of the centroids. In this way, the hubness based approach helps us to compare the proximity and validity of the centroids formed from K-means clusters. The stepwise procedure employed for the detection of the true projected outliers:

- The large dataset or higher dimensional data is uploaded or taken .In our experiment, we have taken the iris dataset from UCI repository where n=7500 d-dimensional points whose components are independently drawn.
- Then, the dimensionality of this large dataset is reduced using Principal Component Analysis technique where the k^{th} Principal component is the variance maximising direction orthogonal to the $(k-1)^{th}$ previous Principal components.
- Then on these reduced dimensions, the iterative k-means algorithm is applied to have the concrete clusters using the central centroids and Mediods. Hence, the higher dimensional data space in this way is divided into the proper subspaces and the outliers can be detected and defined properly.
- To obtain more refined and true outliers, the phenomenon of hubness is also used which is based on the local centrality. Hubness not only converges the Mediods and centroids more to their accurate and approximate locations but also improves the quality of clustering as well as subspace in order to provide the true projected outliers.

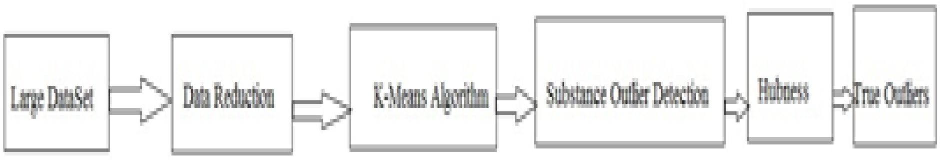

Fig 2. Block Diagram for the detection of outliers

7. RESULTS AND DISCUSSIONS

The results of this experiment are implemented on matlab and the iris data is taken from UCI repository . This is an n=7500 d- dimensional data with independent components.Then Principal Component analysis is done which is simply based upon maximising the variance and according to which the first principal component is in the direction which maximizes the variance among all the directions orthogonal to the first.

Then K-means algorithm is applied iteratively and the various cluster centroids and clusters are formed and shown in fig 3(a). below . Now we want to find out the exact lower dimensional subspaces because outliers mostly hidden in H.D.D equidistant from each other ,so it becomes important to find out the subspaces in lower dimensional area. In order to find the appropriate lower dimensional subspaces the hubness phenomenon is used which is a tactic to overcome the problem of curse of dimensionality.Hubness is a very effective tool in eliminating the sparse subspaces for every cluster and projecting the points into those subspaces in which the greater similarity occurs. Moreover, this K-means algorithm is used as an itrerative approach for defining clusters around separated high hubness data elements and a very important point is that the centroids obtained from K-means depend on all the cluster elements while hubs depend mostly on their neighboring elements and therefore they contain the localized centrality information . Moreover,Table1 shows the comparason of the various outlier detection methodologies.

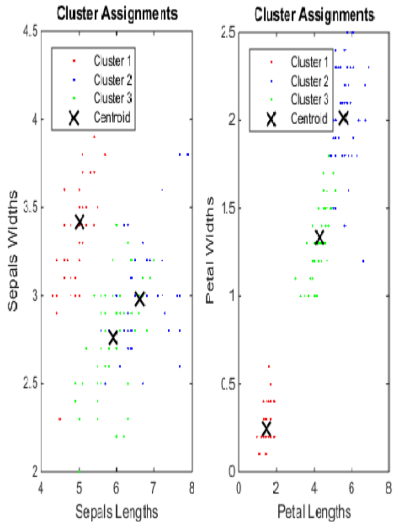

Fig 3 (a) Cluster Centroids along with clusters

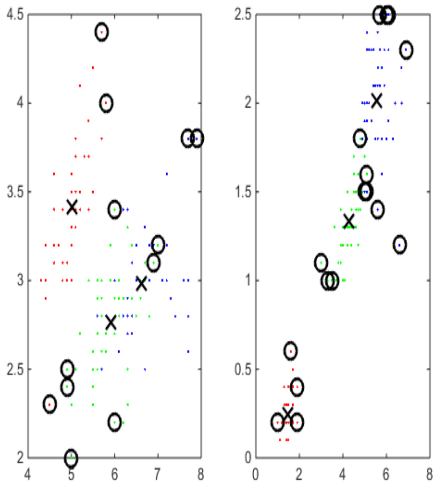

Fig 3 (b) Refined subspaces along
With projected outliers

Table 1 Comparison of various existing outlier detection Methodologies

Outlier Detection Methodology	**Anomaly detection Ratio for H.D.D**	**Compatibility with H.D.D**
Statistical Based Methods	Low	No
Distance Based Methods	Very Low	No
Density Based Methods	Low	No
Clustering based Methods	High	Not All
Subspace Methods	Very High	Yes
Hubness Method	Very High	Yes

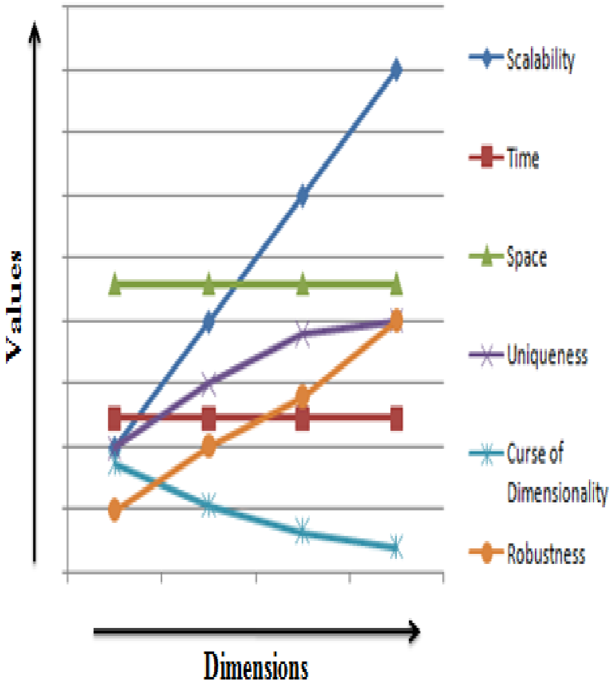

Fig 4. Graphical Representation of various parameters against values and dimensions

8.CONCLUSION

In this paper, we have used the technique of Principal Component Analysis for the reduction of data and Hubness based clustering to obtain the appropriate and concrete clusters. Firstly the various clusters are formed using iterative K-means clustering and then their centroids are converged in an effective way using hubness that enhances the goodness and the quality of the clusters and finally linked with the principal component analysis technique to find out the global and the refined outliers. Fig 4 provides the compatibility of various outlier detection methodologies with the higher dimensional data.

9. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Comput. Survey, vol. 41, no. 3, pp 15, 2009.

[2] P. J. Rousseuw and A. M. Leroy, Robust Regression and Outlier Detection. Hoboken, NJ, USA: Wiley, 1987.

[3] PhD Thesis by Dr. Ji Zhang entitled as “Towards Outlier Detection for higher Dimensional data streams using projected outlier analysis strategy, 2009.

[4] D.M. Hawkins, Identification of Outliers. Chapman and Hall, 1980.

[5] M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, “LOF: Identifying Density-Based Local Outliers,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2000.

[6] H.-P. Kriegel, M. Schubert, and A. Zimek, “Angle-Based Outlier Detection in High-Dimensional Data,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and data Mining, 2008.

[7] F. Angiulli, S. Basta, and C. Pizzuti, “Distance-Based Detection and Prediction of Outliers,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 145-160, 2006.

[8] V. Barnett and T. Lewis, Outliers in Statistical Data. John Wiley&Sons, 1994.

[9] W. Jin, A.K.H. Tung, J. Han, and W. Wang, “Ranking Outliers Using Symmetric Neighborhood Relationship,” Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2006.

[10] N.L.D. Khoa and S. Chawla, “Robust Outlier Detection Using Commute Time and Eigenspace Embedding,” Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2010.

[11] E.M. Knox and R.T. Ng, “Algorithms for Mining Distance-Based Outliers in Large Data Sets,” Proc. Int’l Conf. Very Large Data Bases, 1998.

[12] H.-P. Kriegel, P. Kro¨ger, E. Schubert, and A. Zimek, “Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data,” Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2009.

[13] C.C. Aggarwal and P.S. Yu, “Outlier Detection for High Dimensional Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2001.

[14] C. C. Aggarwal and P. S. Yu, ―Finding generalized projected clusters in high dimensional spaces, in Proc. 26th ACM SIGMOD Int. Conf. on Management of Data, 2000, pp. 70–8

[15] K. Kailing, H.-P. Kriegel, P. Kr ¨oger, and S. Wanka, ―Ranking subspaces for clustering high dimensional data, in Proc. 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), 2003, pp. 241–252.

[16] D. Francois, V. Wertz, and M. Verleysen, ―The concentration of fractional distances, IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 7, pp. 873–886, 2007.

[17] A. Kab´an, ―Non-parametric detection of meaningless distances in high dimensional data, Statistics and Computing, vol. 22, no. 2, pp. 375–385, 2011.

[18] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, ―On the surprising behavior of distance metrics in high dimensional spaces,‖ in Proc. 8th Int. Conf. on Database Theory (ICDT), 2001,pp. 420–434.

[19] I. S. Dhillon, Y. Guan, and B. Kulis, ―Kernel k-means: spectral clustering and normalized cuts, in Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2004, pp. 551

[20] M.Radovanovi´c, A. Nanopoulos, and M.Ivanovi´c, ―On the existence of obstinate results in vector space models,‖ in Proc.33rd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2010, pp. 186–193.

[21] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang, Member, IEEE, Anomaly Detection via Online Oversampling Principal Component Analysis. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013.

[22] Nenad Tomasev, Milos Radovaovic, Dunja Mladenic Ivanovic. "The Role of hubness in clustering higher dimensional data"

[22] Sindhupriya.R. Ignatius Selvarani.X "K-means based Clustering in Higher Dimensional Data" International Journal of Advanced Research in Computer Engineering and Technology, vol 3,Issue 2 ,Feb,2014

[23] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang, Member, IEEE "Anomaly Detection via Online Oversampling Principal Component Analysis"