ISSN:2379-3686 International Journal Of Science Research And Technology Volume 3 Issue 1,PP 32-36, 25th April 2017

OUTLIER DETECTION USING CLUSTER-BASED APPROACH: A REVIEW

Ms. Mayuri A. Bhangare
ME student
Department of Computer Engineering
K. K. W. I. E. E. R., Nashik

Prof. J. R. Mankar
Assistant Professor
Department of Computer Engineering
K. K. W. I. E. E. R., Nashik

Abstract : Outlier detection is a crucial task in data mining which aims to detect an outlier from given data set. The data is said to be an outlier which appears to have inconsistent observation with the remaining data. Outliers are generated because of improper measurements, data entry errors or data arriving from various sources than remaining data. Outlier detection is the technique which discovers such type of data from the given data set. Several techniques of outlier detection have been introduced which requires input parameter from the user such as distance threshold, density threshold, etc. The goal of this proposed work is to partition the input data set into the number of clusters and then outlier is detected for each cluster. The computational time is affected as the data set size is reduced in first phase due to clustering. This work aims at studying two-three different methods of outlier detection with two different data sets. Also analyzing the performance of each method based on system accuracy.

Keywords - Clustering, Clustering center identification, Density Metrics

I. INTRODUCTION

Outlier is a data which appears to have inconsistent observation with the remaining data and outlier detection is the technique which discovers such type of data from the given data set. Outlier is generated because of data entry errors, improper measurements or data arriving from various sources than rest of the data [14].

Outlier detection is an important task in data mining which aims to detect an outlier from given data set. Outlier detection is the first step towards obtaining a coherent analysis in many data-mining applications. The technique of outlier detection is used in many fields such as data cleansing, environment monitoring, criminal activities in e-commerce, clinical trials, network intrusion detection etc.

The cluster-based outlier detection the method works in two phases. In first phase the data set needs to be clustered using Unsupervised Extreme Learning Machine [13]. Unsupervised learning machine (US-ELM) deals with unlabeled data and performs clustering efficiently. US-ELM can be used for multicluster clustering for unlabeled data. In second phase the outliers are detected from each cluster.

Proposed system extends ELM to Unsupervised Extreme Learning Machine. It deals only with unlabeled data and also handles clustering task efficiently. Proposed system works in two phases where in first phase k-number of clusters is generated

using US-ELM from input data set and in second phase an outlier is detected from each cluster using different methods viz. Outlier Detection Algorithm (ODA) and Pruning Technique for different data sets. Then the systems final output is the set of outliers. The main focus of this paper is on the analysis of the methods used for outlier detection. In analysis the outlier detection accuracy, cluster accuracy and computational time are considered.

In pruning all the clusters generated by US-ELM given as an input and the process is repeated for each cluster. The distance between centroid and each point is calculated and the points are sorted in descending order. For point p if the, find its k-nearest neighbor's (kNNs), the distance of point which is farthest from centroid in set of kNNs is considered as temporary distance. The point q is said to be an outlier if it follows conditions explained in following section in detail.

In ODA all the clusters generated by US-ELM are given as an input and then the radius of each cluster is calculated [17] .The pruning operation is performed in each cluster for the point whose distance from centroid is less than the radius of the cluster and other point remains unpruned. Once the pruning is over for all clusters the Local outlier Factor (LOF) for unpruned points is calculated which gives information about how much the point differs from its neighbor's. If the LOF of unpruned point is higher that means it is highly deviated from its neighbor's and the point is declare as an Outlier.

II. RELATED WORK

Huang et al. [1] introduced Extreme Learning machine (ELM) used for training Single Layer Feed Forward Network (SLFNs).The bias and parameters of SLFNs are randomly generated and ELM updates the output weights between hidden layer and output layer. ELM solves regularized least squared problem quicker than the Support Vector Machine's (SVM) quadratic programming problem. But ELM only works with labeled data.

D. Liu [2] extended ELM to the Semi-Supervised Extreme Learning Machine (SS-ELM) where the manifold regularization architecture was imported into the ELMs model to manage both unlabeled and labeled information. When the number of instances is higher than the number of neurons the SS-ELM and SS-ELM are work effectively. But SS-ELM is not able to achieve this because the data is not sufficient as compared to the number of hidden neurons.

J. Zhang [3] proposed co-training technique to train ELMs in SS-ELM. The labeled training sets grows progressively by transferring a subset of most positively judged unlabeled data at each iteration to the labeled set, and pseudo-labeled set is generated. This newly generated pseudo-labeled set is used to train ELMs regularly. The ELM's needs to train regularly in this algorithm, it makes effects on computational cost.

Statistical community [4, 5] is the first to study the problem of outlier and proposed model based outliers. They assumed that the data set follows some distribution or at least statistical estimates of unknown distribution parameters. An outlier is the data from data set that deviates from assumed distribution of data set. These model based approaches degrades their performance with high dimensional data set and arbitrary data set since there is no chance to have prior knowledge about distribution followed by these type of data set.

K. Li [6] proposed some model free outliers methods to overcome the drawback of model based outliers.

There are two model free outliers detection approaches viz. Density based and Distance based. But these two model free outlier approaches required some input parameter to declare an object as an outlier e.g. distance threshold, number of objects nearest neighbor, density threshold etc.

Knorr and Ng [7-9] proposed another algorithm Nested-Loop (NL) to compute distance-based outlier. In this algorithm the buffer is partitioned into two halves viz. first array and second array. It copies data set into both arrays and computes the distance between each pair of objects. The count of neighbor is maintained for objects in first array. It stops counting neighbors of an object as soon as count of neighbors reaches to the D. Drawback of this algorithm is it takes high computation time. Typically nested loop algorithm requires $O(N^2)$ distance computations where N is number of objects in data set.

Angiulli et al. [12] proposed a method Detecting Outliers Pushing objects into an Index (DOLPHIN) which works with data sets resident to disk. It is easy to implement and also can work with any data type. It has I/O cost of successive reading two times the input data set file is inputted. Its performance is linear in time with respect to data set size since it performs similarity search without pre-indexing the whole data set. This method is improved further in efficient computations adopting spatial indexing by other researchers e.g. R-Trees, M-Trees etc. But these methods are sensitive to the dimensions.

III.PROBLEM FORMULATION

To design and develop a system for Outlier Detection using Cluster-based approach

IV. SYSTEM ARCHITECTURE

Fig.1. gives the detail idea about working of the system. The system works in two phases in first phase k number of clusters of input data sets are formed using US-ELM whereas in second phase the outliers from each cluster is detected and finally system gives set of outliers as an output.

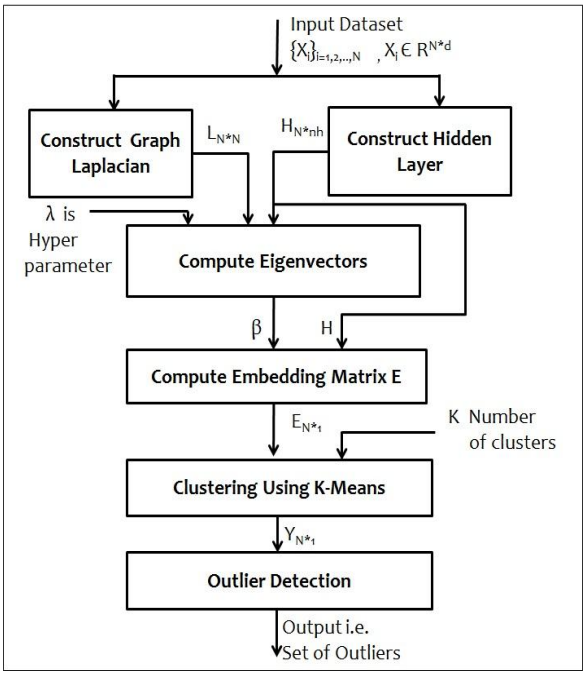

Figure 1: System architecture

Here for clustering US-ELM algorithm is used to form good quality clusters. An input given to the outlier detection block is the clusters created in first phase.The different techniques are used to find outlier from each cluster.

Figure 1 represents proposed system architecture. Processing of proposed work is takes places as following way :

A. Graph Laplacian : Graph Laplacian(L) is computed using similarity matrix(W) and diagonal matrix (D) using formula L=D-W.

W is Similarity or weight matrix of size N*N and the nonzero weights are computed by (1.1)

$$\exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right). \quad \text{------} \qquad (1.1)$$

D is Degree Matrix of size N*N. It is diagonal Matrix and it is computed by (1.2)

$$D_{ii} = \sum_{j=1}^{N} W_{i,j} \quad \text{-------} \quad (1.2)$$

B. Initiate ELM Network : Initiation of ELM Network consist of three blocks from block diagram:

1. Construct Hidden Layer
2. Compute Eigenvectors
3. Compute Embedding Matrix.

1. Construct Hidden Layer : ELM's goal to learn an approximation function or a decision rule rely on the training data. The training of ELMs has two stages. In the initial stage hidden layer is constructed using randomly generated fixed number of mapping neurons. The mapping function can be any linear piecewise continuous functions such as Sigmoid function (2.1) and Gaussian function (2.2).

$$g(x_i, \theta) = \frac{1}{1 + \exp(-(a_j{}^T x_i + b_j)} \quad \text{------} \quad (2.1)$$

$$g(x_i; \theta) = \exp(-b_j \parallel x_i - a_j{}^T \parallel) \quad \text{-------} \quad (2.2)$$

Where θ = {a, b} are the parameters of the mapping function and ||. || denotes the Euclidean norm.

The output of the hidden layer is denoted as H and it is given by

$$H = \begin{pmatrix} g(x_{1,} a_1, b_1) & \ldots\ldots\ldots & g(x_1, a_{nh}, b_{nh}) \\ \vdots & & \vdots \\ g(x_{n,} a_1, b_1) & \ldots\ldots\ldots & g(x_{n,} a_{nh}, b_{nh}) \end{pmatrix}$$

Where n_h is the number of neurons considered n construction of ELM network.

2. Compute Eigenvectors : In the next stage, ELMs solves the output weights in such way that the sum of the squared losses of the prediction errors must be minimum. The computation of output weights is depending on the number of patterns and the number of neurons.It is denoted as β.

3. Compute Embedding Matrix : The final output matrix called embedding matrix(E) is computed by (3.1).

$$E = H * \beta \quad \text{----------} \quad (3.1)$$

C. K-Means Clustering : Now, to implement clustering task in embedded space k-means algorithm is adopted. Hence to perform clustering N patterns are cluster into k number of clusters using k-means by considering each row of matrix E as point.Clustering task gives Label vector y as an output with cluster index for all N patterns from given data set.

D. Outlier Detection : In this section first the Cluster Based Outliers are defined and then the algorithm to compute outlier from each cluster of given data set.

Here in this phase of system different algorithms namely pruning and ODA are used for outlier detection and the results are compared.

V. CONCLUSION

In this survey paper, several existing techniques have studied and analysed in section under related work. Traditional methods of outlier detection work effectively and efficiently to identify outliers from dataset but user needs to give some threshold value as input to the system. This paper aims to detect outliers is the task that finds objects that appears to have inconsistent observation with the remaining data. We analyzed an efficient outlier detection method. partitioned the input data set into the number of clusters and then outlier is detected for each cluster. Due to reduction in size of dataset, the computation time reduced considerably. Then the outliers are detected for each cluster. We get outliers within a cluster.

REFERENCES

[1] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in Proc. Int. Joint Conf. Neural Netw., vol. 2. 2004, pp. 985–990.

[2] L. Li, D. Liu, and J. Ouyang, "A new regularization classification method based on extreme learning machine in network data," J. Inf. Comput. Sci.,vol. 9, no. 12, pp. 3351–3363, 2012.

[3] K. Li, J. Zhang, H. Xu, S. Luo, and H. Li, "A semi-supervised extreme learning machine method based on co-training," J. Comput. Inf. Syst.,vol. 9, no. 1, pp. 207–214, 2013.

[4] Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, New York (1994).

[5] Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley, New York (2005)

[6] He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recog. Lett. 24(9), 1641–1650 (2003).

[7] Knorr, E.M., Ng, R.T.: Algorithms for mining distance based outliers in large datasets. In: Proceedings of the International Conference on Very Large Data Bases, pp. 392–403 (1998).

[8] Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Rec. 29(2), 427–438 (2000).

[9] Angiulli, F., Pizzuti, C.: Outlier mining in large high-dimensional data sets. IEEE Trans.

[10] Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. ACM Sigmod Rec. 29(2), 93–104 (2000).

[11] Bay, S.D, Schwabacher, M.: Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 29–38 (2003).

[12] Angiulli, F., Fassetti, F.: Very efficientmining of distance-based outliers. In: Proceedings of the Sixteenth ACMConference on Information and KnowledgeManagement, pp. 791–800 (2007).

[13] Huang, G., Song, S., Gupta, J.N.D.,Wu, C.: Semi-supervised and unsupervised extreme learning machines. IEEE Trans. Cybern. 44(12), 2405–2417 (2014).

[14] Hawkins, D.M.: Identification of Outliers. Springer, New York (1980).

[15] X. Wang, D. Shen ,M. Bai, T. Nie ,Y. Kou ,G. Yu : Cluster-Based Outlier Detection Using Unsupervised Extreme Learning Machines.Springer International Publishing Switzerland 2016J. Cao et al. (eds.), Proceedings of ELM-2015 Volume 1, Proceedings in Adaptation, Learning and Optimization 6, DOI 10.1007/978-3-319-28397-5_11.