Distance Based Algorithm for Effective Outliers Classification and Prediction of WDBC Dataset

Dr. D. Rajakumari
Assistant Professor, Department of Computer Science
Nandha Arts and Science College, Erode -628052, Tamilnadu, India

***Abstract*— Knowledge Discovery on Database (KDD) is an essential process on data processing. Many features selection and classification algorithms are used to select the relevant features and classified in data mining applications. The outlier detection is presently growing as an extensive task in the data mining applications. Many outlier detection techniques were developed previously to overcome the challenges in the detection of outliers. Many feature selection and classification algorithms are used to select the relevant features and classify them according to criteria in data mining applications. These techniques suffer some limitations due to increasing complexity, size and variety of data sets. This paper proposes a novel boundary based classification approach for the effective prediction of outliers.**

***IndexTerms*—: Boundary Based Classification, Data Mining Outlier Detection, Wisconsin Breast Cancer Dataset**

I. INTRODUCTION

Nowadays there is collection and storage of huge amount of data across the world. It is really hard to find the database having Terabytes of data in the enterprise and research facilities. Data Mining is used for extracting useful information from a collection of databases. It is practically impossible for mining data from the huge amount of database, without the automatic methods. Throughout the years, many algorithms were created to extract nuggets of knowledge from large sets of data. Classification is a data mining technique used to predict group membership for data instances. Classification is the process of assigning class labels to the instance based on their attribute values. Generally classification is categorized into two major types such a supervised classification and unsupervised classification. The presence of outlier in the classification system causes major issue in the classified labels. The accuracy of the classification approach is affected due to the presence of outliers in the data set and the inability to correctly classify data records near the boundary. These outliers are mostly related to the boundary but they are functionally different. It may create very harmful effect in the medical oriented classification system.

Outlier is defined as the set of objects that are considerably dissimilar from the rest of the data. Generally outliers are categorized as erroneous or real. Real outliers are defined as the observations whose actual values are different from the observed value for the remaining data. Erroneous outliers are the observations that are distorted due to the misreporting errors occur during the classification process. Both the outliers exert more influence on the classification results. Outlier in the classification problem is roughly distinguished as attribute noise or class noise. Attribute noise affects the observed value of the input pattern during the classification process. The class noise changes the labels assigned to the instances.

Outlier detection techniques are used to reduce the influence of outliers in the final model to develop, or as a preliminary pre-processing stage before the elaboration of the information conveyed by a signal. It is really useful in various applications, such as network intrusion, medical diagnosis or fraud detection. Outlier (anomaly) detection refers to the task of identifying patterns that do not conform to established regular behavior. Despite the lack of a rigid mathematical definition of outliers, their detection is a widely applied practice. The interest in outliers is strong since they may constitute critical and actionable information in various domains. Hence, outlier detection has emerged as an important task in the field of data mining.

Outlier detection approaches mainly focus on discovering patterns that occur infrequently in the data, as opposed to many traditional data mining techniques, such as association rule mining or frequent item set mining, that attempt to end patterns that occur frequently in the data. The task of detecting outliers can be categorized as supervised, semi-supervised, and unsupervised, depending on the existence of labels for outliers and/or regular instances. Supervised outlier detection techniques have an explicit notion of the normal and outlier behavior and hence accurate models can be built. The drawback here is that accurately labeled training data might be prohibitively expensive to obtain. Labeling is often done manually by a human expert and hence requires a lot of abort to obtain the labeled training data set. Certain techniques inject artificial outliers in a normal data set to obtain a fully labeled training data set and then apply supervised outlier detection techniques to detect outliers in test data. Semi-supervised techniques assume the availability of labeled instances for only one class; it is often difficult to collect labels for other class. The unsupervised techniques do not make any assumption about the availability of labeled training data.

Among these categories, unsupervised methods are more widely applied, because the other categories require accurate and representative labels that are often prohibitively expensive to obtain. Unsupervised methods include distance-based methods that mainly rely on a measure of distance or similarity in order to detect outliers. Formulation of outlier detection depends upon the various factors such as input data type and distribution, availability of data and resource constraints introduced by application domain.

II. RELATED WORK

The naïve approach of distance-based outlier detection takes quadratic time with respect to the number of points in the dataset by comparing each point with the rest. This paper include the various feature selection algorithms and their implementation in outlier prediction.

David Sathiaraj and EvangelosTriantaphyllou [1] On Identifying Critical Nuggets of Information during Classification Tasks. They mentioned the proposed an innovative domain-independent method to measure criticality and suggested a heuristic to reduce the search space for finding critical nuggets. In classification tasks, data set that conforms to a certain representation or a classification model was considered. If one were to perturb a few data instances by making small changes to some of their attribute values, the original classification model representing the data set was changed. Also, if one were to remove those data instances, the original model could change significantly. The magnitude of changes to the original model provided clues to the criticality of such data instances, as more critical data instances tend to impact the model more significantly than data instances that are comparatively noncritical. This idea is exploited in this paper to introduce the notion of critical nuggets, to define a metric for criticality and for the eventual mining of critical nuggets. The critical nuggets were isolated and validated on some real-world data sets. This work also identified certain properties of critical nuggets and provided experimental validation of the properties

M. Radovanovic, A. Nanopoulos, and M. Ivanovic [2] Reverse nearest neighbors in unsupervised distance-based outlier detection. The demonstrated that the distance-based outlier methods have produced more contrasting outlier scores in the high dimensional data. This was done by reexamining the reverse nearest neighbors in the unsupervised outlier detection process. This paper focused on the effects of high dimensionality on unsupervised outlier-detection methods and the hotness phenomenon, extending the previous examinations of anti-hotness to large values of k, and exploring the relationship between hotness and data sparsely. Based on the analysis, the Anti Hub method was formulated for unsupervised outlier detection and a derived method was proposed to improve discrimination between scores. The unsupervised methods can detect outliers which are more pronounced in high dimensions, under the assumption that all (or most) data attributes are meaningful, i.e. not noisy

F. Angiulli, S. Basta, S. Lodi, and C. Sartor Angiulli [3] Distributed strategies for mining outliers in large data sets. The introduced a distributed method to detect the distance-based outliers in very large data sets. The method exploits parallel computation in order to obtain vast time savings. Indeed, beyond preserving the correctness of the result, the proposed schema exhibits excellent performances. The key point of our approach is to exploit the locality properties of the problem at hand to partition the computation among the processors of a multi-processor system or the host nodes of a communication network to obtain vast time savings. From the theoretical point of view, the temporal cost of our algorithm is expected to be more than the classical nested-loop like approach to detect outliers. Experimental results show that the algorithm is efficient and that its running time scales quite well for increasing number of nodes. Importantly, the solving set computed by our approach in distributed environment has the same quality as that produced by the corresponding centralized method.

K. Bhaduri, B. L. Matthews, and C. R. Giannella [4] Algorithms for speeding up distance-based outlier detection.The problem of distance-based outlier detection is difficult to solve efficiently in very large datasets because of poten-tial quadratic time complexity. Sequential and distributed algorithms were developed to address this problem, by combining the indexing and disk block accessing techniques. By combining simple but effective indexing and disk block accessing techniques, a sequential algorithm iOrca was developed. The indexing scheme was based on sorting the data points in order of increasing distance from a fixed reference point and then accessing those points based on this sorted order. To speed up the basic outlier detection technique, two distributed algorithms were developed for modern distributed multi-core clusters of machines, connected on a ring topology. The first algorithm passes data blocks from each machine around the ring, incrementally updating the nearest neighbors of the points passed. By maintaining a cutoff threshold, it is able to prune a large number of points in a distributed fashion. The second distributed algorithm extends this basic idea with the indexing scheme discussed earlier.

N. Pham and R. Pagh [5] A near-linear time approximation algorithm for angle-based outlier detection in high-Dimensional data. Pham and Pugh suggested a novel random projection-based technique to estimate the angle-based outlier factor for all data points. The random projection-based algorithm approximates the variance of angles between pairs of points of the data set. A robust outlier score detects high-dimensional outlier patterns. The theoretical analysis of the quality of approximation was introduced to guarantee the reliability of estimation algorithm. Also, our approach is suitable to be performed in parallel environment to achieve a parallel speedup. The variance of angles (VOA) is used as the outlier factor to evaluate the degree of out liernes of each point of the data set. The naive Angle-Based Outlier Detection (ABOD) algorithm computes the VOA for each point of the data set and return the top 'm' points having the smallest VOA as outliers. The outlier rankings are computed based on the variance of cosine spectrum with or without weighting factors and the variance of angle spectrum are likely similar in high-dimensional data.

Albanese et al [6] proposed a Rough Outlier Set Extraction method for the theoretic representation of the outlier set by using the rough set approximations. A novel Kernel set was also introduced for describing the original data set. The same outlier set was detected efficiently and quickly by using the kernel set. .*Kim et al* [7] utilized the kd-tree indexing and an approximated k-nearest neighbor (ANN) search algorithm to reduce the computation time of the density-based outlier detection. Hence, the local outlier was detected effectively within a short time. A novel outlier detection approach was presented to address the data with imperfect labels and incorporate limited abnormal examples into learning. [8] The outlier detection performance was improved by effectively handling the data with imperfect labels using the combination of local and global outlier detection approaches. A better compromise between the false alarm rate and outlier detection rate was achieved when compared to the advanced outlier detection approaches.

A novel subspace search method was proposed for selecting the high contrast subspaces for the density-based outlier ranking. The quality of traditional outlier rankings was improved by computing the outlier scores in high contrast projections [9]. A two-phase algorithm was proposed for detecting outliers in the categorical data. Clustering and ranking were performed to determine the set of most likely outliers. The computational complexity of the proposed algorithm was not affected by the number of outliers to be detected [10].An iterative self-organizing map approach with robust distance estimation was proposed for spatial outlier detection. The high dimensional problems of spatial attributes was addressed and spatial outliers with irregular features was detected accurately by the proposed process [11]A novel method for correctly detecting the outliers was introduced based on the new technique developed for simultaneous evaluation of mean, variance and outliers[12]. An unsupervised outlier detection scheme including the combination of density based and partitioning clustering method was presented for data streaming process. The outlier detection rate of the proposed approach was better than the existing approaches [13].

The distance-based outlier detection method was actually expensive. The limitations of the clustering based outlier detection move toward were difficult to find the clusters with irregular shapes and specify the number of clusters. The major drawbacks of the density-based clustering approach were the sensitivity of the algorithm with respect to the density of the starting object and inefficient identification of adjacent clusters of various densities. The computational complexity was also too high.

III. NOVAL BOUNDARY BASED CLASSIFICATION APPROACH (NBBC)

The proposed Novel Boundary based Classification including the imputation methods and ordinal classification methods are explained in this section. The detailed description of WDBC dataset as follows.

WDBC dataset:

The Wisconsin Diagnostic Breast Cancer (WDBC) contains various attributes namely, diagnosis, ID number and real valued features. There are ten real valued features namely, radius, area, perimeter, smoothness, texture, compactness, concave points, concavity, symmetry and fractal dimension computed from digitized image of breast mass.

The training dataset for the class label c is defined as

$$X^c = \begin{bmatrix} x^1 \\ x^2 \\ . \\ . \\ x^n \end{bmatrix} \tag{1}$$

Where the single ith instance x^i represents the 'm' number of features is expressed as

$$x^i = [\, f_1^i\,, f_2^i\,, f_3^i \dots f_m^i] \tag{2}$$

The triangular area is computed to find the relationship between the each features present in the instancex^i. The triangular matrix for the ith instance with class label 'c' is written as

$$TA_i^c = \begin{pmatrix} 0 & T_{1,2}^i & T_{1,3}^i & \dots & T_{1,m}^i \\ T_{2,1}^i & 0 & T_{2,3}^i & \dots & T_{2,m}^i \\ T_{3,1}^i & T_{3,2}^i & 0 & \dots & T_{3,m}^i \\ T_{4,1}^i & T_{4,2}^i & T_{4,3}^i & 0 \dots & T_{4,m}^i \\ T_{5,1}^i & T_{5,2}^i & T_{5,3}^i & \dots & 0 \end{pmatrix} \tag{3}$$

In the triangular matrix main diagonal values are must be zero. Upper diagonal and lower diagonal triangular matrix values are same. In case of any updating in the upper triangular value, it will affect the lower diagonal value. Hence the correlation between the features in the instance x^i is represented by upper or lower diagonal triangular value. Triangular value is calculated using the vector values $\|\mathrm{V}_\mathrm{a}^\mathrm{i}\|$ and $\|\mathrm{V}_b^\mathrm{i}\|$ of the ath and both features

$$T_{a,b}^i = \frac{\|V_a^i\| * \|V_b^i\|}{2} \tag{4}$$

Where $1 \leq i \leq n$, $1 \leq a \leq m$, $1 \leq b \leq m$. Here 'n' and 'm' denote the number of instances and number of features in the dataset. The values of ath and bth features lie within the number of features instances. The vector values lie between the 0 and 1 except the same feature combination values are 1 in $D_{a\,,1}$.To obtain the triangular value, the transformation technique is applied for the vectors. $Y_{a,b}^i$is defined as the form of Cartesian coordinate system ($V_a^i V_b^i$). The temporary value of the ath and bth features is

$$Y_{a,b}^i = \left[V_a^i V_b^i\right]^T = \begin{bmatrix} D_{a\,,1} & D_{a\,,2} & \dots & D_{a\,,m} \\ D_{b\,,1} & D_{b\,,1} & \dots & D_{b\,,m} \end{bmatrix} \times x^i \tag{5}$$

The distance between the pth and qth features is calculated by using the equation

$$D_{p,q} = \sqrt{\left(x_p - x_q\right)^2} \tag{6}$$

Triangular average for the class is calculated as

$$\overline{T_i^c} \leftarrow \frac{1}{q*r}\sum_{j=0}^{q}\sum_{k=0}^{r} TA_i{}^{j,k} \qquad (7)$$

The covariance between the features in the triangular matrix is used to find the linear changes of all features. Dimension of the covariance matrix is the number of features * number of features. Covariance Matrix for the class c is given as

.

$$Cov^c = \begin{bmatrix} \sigma(T_{2,1}^c, T_{2,1}^c) & \sigma(T_{2,1}^c, T_{3,1}^c) & \dots & \sigma(T_{2,1}^c, T_{m,m-1}^c) \\ \sigma(T_{3,1}^c, T_{2,1}^c) & \sigma(T_{3,1}^c, T_{2,1}^c) & \dots & \sigma(T_{3,1}^c, T_{m,m-1}^c) \\ & \dots & & \\ & \dots & & \\ & \dots & & \\ \sigma(T_{m,m-1}^c, T_{2,1}^c) & \sigma(T_{m,m-1}^c, T_{2,1}^c) & \dots & \sigma(T_{m,m-1}^c, T_{m,m-1}^c) \end{bmatrix} \qquad (8)$$

he standard deviation between the triangular values of two instances is defined as

$$\sigma(T_{u,v}^c, T_{x,y}^c) = \frac{1}{z-1}\sum_{j=1}^{z}(T_{u,v}^{c,j} - \mu_{T_{u,v}^c})(T_{x,y}^{c,j} - \mu_{T_{x,y}^c}) \qquad (9)$$

Where

$$\mu_{T_{u,v}^c} = \frac{1}{z}\sum_{i=1}^{z} T_{u,v}^{c,i} \qquad (10)$$

The distance for ith Instance for class label is computed by using the equation

$$MD^{c,i}(TA_i{}^c, \overline{T^c}) = \left|\sqrt{\frac{(TA_i{}^c - \overline{T^c})^T(TA_i{}^c - \overline{T^c})}{Cov^c}}\right| \qquad (11)$$

Training Algorithm:

```
Step1: Compute Triangular Area // using eqn(3)
Step:2 For I = 1,2, … z do
Step:3 $TA \leftarrow TA_i{}^c$
Step:4 End For
Step:5 Compute $\overline{T^c} \leftarrow \frac{1}{z}\sum_{i=1}^{z}\overline{T_i^c}$
Step:6 Compute Covariance Matrix $Cov^c$// using eqn(8)
Step:7    For I = 1,2, … z do
Step:8    Compute $MD^{c,i} \leftarrow MD^{c,i}(TA_i{}^c, \overline{T^c})$
Step:9    End For
Step:10 Compute $\mu$
Step:11  Compute $\sigma$
Step:12Return
```

If the distance value is greater than or lesser than the boundary layer, then it is considered as an outlier and the other class label for the appropriate instance is predicted.

Testing Algorithm:

```
Step 1: Compute Triangular Area
Step2:  For I = 1,2, … z do
Step:3$TA_i \leftarrow TXA_i$
Step:4For c= 1,2, … C do
Step:5 Compute $MD^{c,i} \leftarrow MD^{c,i}(TA_i, \overline{T^c})$
Step:6if( $\mu - \sigma * 0.5$ ) ≤ $MD^{c,i}$ ≤ ( $\mu + \sigma * 0.5$ )then
Step:7$TX_i \leftarrow c$
Step:8End if
Step:9End For
Step:10End For
```

In the testing phase, the testing instance to be calculated by using equation (3). Then, the distance between the triangular area for testing instance and the mean triangular value is computed by using equation (11). The computed distance is compared with each class boundary layer, and the class label is stored in the respective satisfied boundary layer class.

The distance measurement approach is used to find the dissimilarity for each instance between the mean values of the triangular matrix. Finally, the outlier analysis is performed for predicting the nuggets in the classified results. Our proposed approach is experimented with the Wisconsin Diagnosis Breast Cancer (WDBC) dataset. The proposed TBC approach achieves better performance in terms of Precision, Recall, Accuracy, Sensitivity, and Specificity, Error rate, F1-score, Correct and Incorrect classification rate than the existing classification algorithms.

IV. PERFORMANCE ANALYSIS

The performance analysis results of the proposed Novel Boundary based Classification (approach. Our proposed approach is tested by using the WDBC dataset. It is the publically available label standard dataset which is used in the classification research. Testing our approach on the WDBC dataset, contributes a credible estimate and makes the evaluation with other state of art technique by using WEKA. WEKA is one of data mining tool which contains state of art mechanisms for classification, clustering preprocessing, etc. The performance metrics used for the evaluation of the proposed approach are

- Incorrectly classified instance
- Correctly classified instance

4.1 *Incorrectly classified instance*

The incorrectly classified instance indicates the percentage of incorrectly classified instances. It shows the ability of the classifier for the correct classification of the outliers.

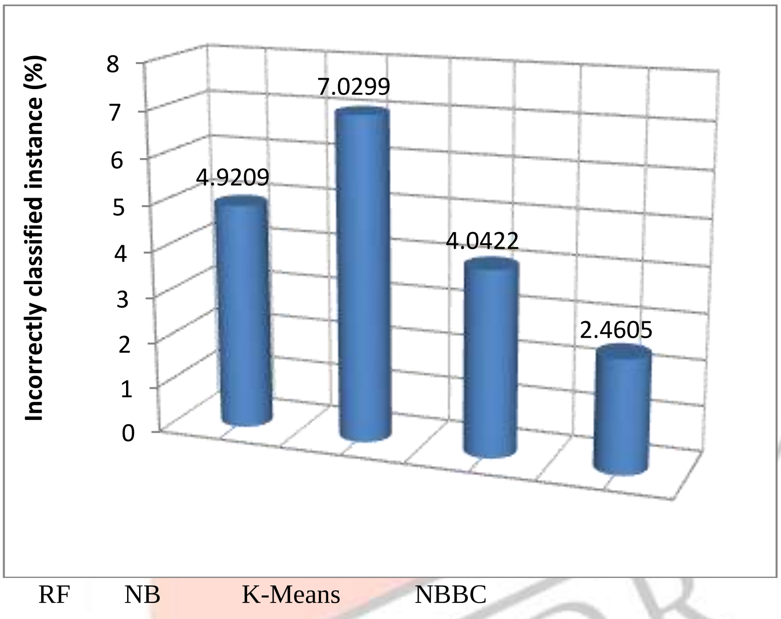

Figure.1 Comparative Analysis of incorrectly classified instance of the proposed approach and existing algorithms

Figure.10 shows the comparison graph for the incorrectly classified instance of the proposed approach and existing algorithms. The incorrectly classified instance of the proposed approach is relatively lower than that of the existing algorithms. This shows the effectiveness of the proposed approach.

4.2 *Correctly classified instance*

The correctly classified instance shows the percentage of correctly classified instances. It shows the measure of incorrectness of the classifier for the dataset. Shows the comparison graph for the correctly classified instance of the proposed approach and existing algorithms. The correctly classified instance of the proposed approach is higher than the existing algorithms. The performance of the proposed method is higher than the existing algorithms.

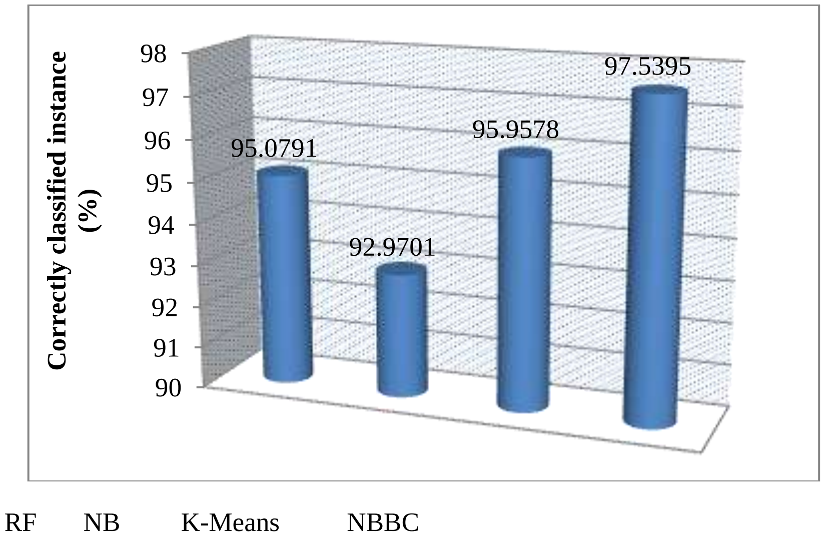

Figure.2 Comparative Analysis of correctly classified instance of the proposed approach existing algorithms

From the performance evaluation results, it is clearly evident that the proposed NBBC approach achieves high performance classified instance. Hence, the proposed NBBC approach is effective than the existing Random forest, Naivy Bayes and K-means IBK algorithms.

V. Conclusion

The novel boundary based classification approach is proposed to overcome the limitations in the existing classification approaches. The training algorithm and testing algorithm are used for training and testing the class. Results from the evaluation on the real-time WDBC data sets revealed that the proposed approach achieves better performance than the existing classification algorithms.

References

[1] David Sathiraj,Evangelos Triantaphyllou, "On Identifying Critical Nuggets of Information during Classification Tasks. " *IEEE Transactions on Knowledge and Data Engineering,* , vol. 25, pp. 1354-1367, 2013.

2] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Reverse nearest neighbors in unsupervised distance-based outlier detection," *IEEE Transactions on Knowledge and Data Engineering,* , vol. 27, pp. 1369-1382, 2015.

[3] F. Angiulli, S. Basta, S. Lodi, and C. Sartori, "Distributed strategies for mining outliers in large data sets," *IEEE Transactions on Knowledge and Data Engineering,*, vol. 25, pp. 1520-1532, 2013.

[4] K. Bhaduri, B. L. Matthews, and C. R. Giannella, "Algorithms for speeding up distance-based outlier detection," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, 2011, pp. 859-867.

[5] N. Pham and R. Pagh, "A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 877-885.

[6] A. Albanese, S. K. Pal, and A. Petrosino, "Rough sets, kernel set, and spatiotemporal outlier detection," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 26, pp. 194-207, 2014.

[7] S. Kim, N. W. Cho, B. Kang, and S.-H. Kang, "Fast outlier detection for very large log data," *Expert Systems with Applications,* vol. 38, pp. 9587-9596, 2011.

[8] B. Liu, Y. Xiao, P. S. Yu, Z. Hao, and L. Cao, "An efficient approach for outlier detection with imperfect data labels," *IEEE Transactions on Knowledge and Data Engineering,*, vol. 26, pp. 1602-1616, 2014.

[9] F. Keller, E. Müller, and K. Böhm, "HiCS: high contrast subspaces for density-based outlier ranking," in *28th International Conference on Data Engineering (ICDE), 2012 IEEE*, 2012, pp. 1037-1048.

[10] N. R. Suri, M. N. Murty, and G. Athithan, "An algorithm for mining outliers in categorical data through ranking," in *12th International Conference on Hybrid Intelligent Systems (HIS), 2012* 2012, pp. 247-252.

[11] Q. Cai, H. He, and H. Man, "Spatial outlier detection based on iterative self-organizing learning model," *Neurocomputing,* vol. 117, pp. 161-172, 2013.

[12] G. Buzzi-Ferraris and F. Manenti, "Outlier detection in large data sets," *Computers & chemical engineering,* vol. 35, pp. 388-390, 2011.

[13] Y. Thakran and D. Toshniwal, "Unsupervised outlier detection in streaming data using weighted clustering," in *12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012*, 2012, pp. 947-952.