INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND CREATIVE ENGINEERING (ISSN:2045-8711)
VOL.6 NO.1 JANUARY 2016

A Comparative analysis on Boundary-based Classification Techniques for Outlier Detection in WDBC Datasets

D.Rajakumari
Ph.D Research Scholar, Department of Computer Science,
Erode Arts & Science College (Autonomous), Erode, Tamil Nadu, India.
Email: rsrajakumarid@gmail.com
Dr.S.Pannirselvam
Associate Professor & Head, Department of Computer Science,
Erode Arts & Science College (Autonomous), Erode, Tamil Nadu, India.
Email: pannirselvam08@gmail.com

Abstract— **Record and sample of large size in data mining requires the outlaying observation detection (Outlier Detection (OD)), since they carry the necessary information. The large size and diversity introduces the limitations in outlier detection techniques. The classifiers involved in machine learning algorithms deteriorate the OD performance, since they are sensitive to noise, irrelevant features. This paper discusses the influence of Triangular Boundary-based Classification (TBC) and the Weighing Based Feature Selection and Monotonic Classification (WFSMC) on the Wisconsin Diagnosis Breast Cancer (WDBC) dataset for an effective outlier prediction. The imputation, weight computation, and ordinal feature selection prior to TBC predicts the relevant features. The normal distribution function-based triangular area support boundary region analysis to provide treatment or precautions to the patients. The points nearer to the boundary region lead to misclassification. Hence, the inclusion of monotonic constraints in the classification phase improves their accuracy. The comparative analysis between the TBC and WFSMC regarding accuracy, precision, and recall proves the effectiveness of WFSMC in real-time data mining applications.**

Keywords- Data mining, Monotonic classification, Ordinal Feature selection, Outlier Detection, Weight Computation, Wisconsin Diagnosis Breast Cancer (WDBC)

1. INTRODUCTION

Outlier prediction plays a significant role in data mining applications namely, fault identification and diagnosis. A multi-disciplinary effort for extracting the knowledge of data refers to data mining. The accuracy of knowledge extraction is affected due to the presence of outliers in the data set and inability to classify data records near the boundary. 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. Hence reliable detection methods are required for the identification of outliers. Depending on the existence of labels, the outlier detection process is classified as supervised [1], Semi-supervised [2], and unsupervised [3] methods. Unsupervised methods are widely used for outlier detection, since other categories require accurate and representative labels that are really expensive. Selection of relevant attributes improves the performance of data mining and avoids the over fitting. Feature selection is an essential part in successful data mining. The formation of subset of original features on the basis of specific criterion is feature selection process [4], [5]. Multi- view learning algorithms are applicable to analyse the presence of correlated and complemented information on data. But, the selection of discriminating features via multi-view learning on single view data is the challenging task. They utilizes the three information namely, data cluster centre, similarity between data and correlation between different views. Approximation of cluster labels introduced the noise due to the discrete nature leads to misguidance in feature selection. Robust spectral learning employs in unsupervised feature selection for the improvement of robustness. In general, the data in high dimensional and a large space are labelled.

The simultaneous explore of labelling and un-labelling of data difficult by using the un-supervised feature selection approaches. The computation of feature score is complex for high dimensionality. Semi-supervised algorithm based on noise insensitive trace ratio criterion analyses the dimensionality reduction. The semi-supervised selection process utilizes the special label propagation method in order to explore the distribution of labelled and unlabelled data. But, the semi-supervised approaches, introduces the small-labelled sample problem in which labelled data are small and unlabelled data are large. Semi-supervised approaches are sensitive to noise from the constraints set is used. Since it urges to implement techniques that can effectively handle the outliers, by previously eliminating the outlier from the training data or by minimizing their consequences if training data is not reliable. This paper proposes a novel triangular boundary based classification approach for the effective prediction of outliers.

The use of semi-supervised feature selection of high dimensional data introduces the challenge that the presence of irrelevant features affects the speed and accuracy. The occurrence of unreliable inferences during the uninformative feature selection introduces the label noise problem. The evaluation mechanism effectively performed data analysis. The feature selection process categorized into three processes based on evaluation mechanism namely, filter, wrapper and embedded. Filter based approaches consider the redundant

variable without the knowledge about the interaction between the variables. Wrapper based models consider the interaction between the variables. But, they suffer two problems such as over fitting risks and high computational time. Embedded methods combined the advantages of both filter and wrapper based models. The main contributions of the proposed approach are

- A Weighing based Feature Selection and Monotonic Classification including the imputation methods and ordinal classification methods, for the effective feature selection.
- The feature weighing scheme based on imputation improves the performance of selection and classification.
- A novel triangular boundary based classification approach including the training and testing algorithms, for the effective prediction of outliers.
- The comparative analysis between TBC and WFSMC regarding accuracy, precision, and recall on Wisconsin Diagnosis Breast Cancer (WDBC) dataset demonstrates the efficiency.

2. Related Work

This section describes the conventional works related to Outlier Detection (OD), influence of feature selection algorithms on OD. Due to the high-dimensionality, OD was a challenging task. Radovanovic et al. [6] demonstrated the distance-based outlier methods that produced more contrasting outlier scores in the high dimensional data by re-examining the reverse nearest neighbours. Angiulli et al. [7] introduced a distributed method for distance-based OD in very large data sets. They realized the efficiency and scalability improvement through experimental results with increasing node size. Pham and Pagh [8] suggested a novel random projection-based technique to estimate the angle-based outlier factor for all data points. Kriegel et al. [9] proposed a novel outlier detection model was introduced to detect the outliers that differ from the normal instances by considering the combinations of different subsets of attributes. Albanese et al. [10] proposed a Rough Outlier Set Extraction method for the theoretic representation of the outlier set by using the rough set approximations. The time consumption for an effective OD was more.

Kim et al. [11] utilized the kd-tree indexing and an approximated k-nearest neighbour (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. The selection of high-contrast projection for maximum accuracy was difficult. Keller et al. [12] proposed a novel subspace search method was proposed for selecting the high contrast subspaces for the density-based outlier ranking for quality improvement. The determination of most likelihood outliers was difficult in the categorical data. Suri et al. [13] performed clustering and ranking 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. An accurate detection of spatial outliers was the difficult process due to the high-dimensionality. Cai et al. [14] addressed the high-dimensional problems of spatial attributes and outliers with irregular features and detected accurately by iterative self-organizing map approach. The simultaneous evaluation of mean and variance was the necessary task in OD validation. Buzzi-Ferraris et al. [15] developed the new technique for correctly OD and simultaneous evaluation. They combined density based and partitioning clustering method was presented for data streaming process. The high-dimensionality is the major problem observed in conventional OD techniques.

The feature selection was an effective technique and an important step for dimension reduction in data mining applications. The irrelevant attributes were filtered before data mining process. Padungweang et al [16] proposed an unsupervised feature selection based on Fourier transform of the probability density function (PDF). The discrimination score extended the data orientation. The time complexity also employed in this category. Traditional feature selection methods were developed to detect the single view data. Feng et al. [17] proposed an Adaptive Unsupervised Multiuser Feature Selection (AUMFS) utilized cluster centre, similarity and correlation. Sparse regression models employed in AUMFS to predict the cluster labels. The noise and irrelevant features affected the estimation of graph laplacian and approximation of cluster labels introduced the noise. Shi et al [18] proposed the Robust Spectral learning based Feature Selection (RSFS) for the improvement of robustness in sparse spectral regression. The multimedia and web based applications handled high dimensional data. The limited number of labelled data leads to the development of semi-supervised feature selection algorithm. The semi-supervised algorithm effectively handled both labelled and unlabelled data. Liu et al [19] analysed the noise insensitive trace criterion for dimensionality reduction. They utilized the special label propagation method explored the distribution of labelled and unlabelled data. The feature selection in data mining extends the applications to video recognition. The supervised approaches were informal in the identification of relevant features due to limited number of labelled videos. Han et al [20] presented the Semi-Supervised Feature Selection via Spline Regression (S2FS2R) that combined the discriminative information and geometric structure. The speed and accuracy of semi-supervised approaches were affected by the irrelevant features.

The diagnosis of disease and early prediction of stages of cancer required the quality improvement in feature selection process. Sharma et al [21] designed the data mining model by using Probabilistic Neural Network (PNN). The PNN based selection model improved the accuracy and effectiveness of the treatment. The discovery of knowledge patterns in multi-clinical categorization was an important process. Jacob et al [22] analysed the various feature selection algorithms namely, fisher filtering, runs filtering, stepwise discriminant analysis and RELIEF on error rate minimization. The comparative study improved the decision making process in classification. The simultaneous tracking of multi-level expressions of genes in the clinical data performed by using microarray. The introduction of reduction technique in feature selection and cluster analysis improved the performance. Zhu et al [23] proposed hybrid Case Based Reasoning (CBR) in order to reduce the redundant features. They selected the minimal set of features from the problem domain and the redundant ones were reduced by using the neighbourhood set algorithm. The feature weight factor was calculated by using RELIEF algorithm. But, the existence of uncertainty in feature weight calculation lead to poor evaluation. Sun et al [24] solved the uncertainty problem by novel mean-variance model based feature selection

algorithm. The mean and variance at discriminative instant considered in the feature weight estimation led the stable and accurate results. Several complex issues namely, class-specific nature, unbalanced data set handling, complexities in multi-labelled data and noise were addressed. Yilmaz et al [25] extended the RELIEF algorithm into RELIEF-MM for multimodal fusion. The representation and reliable capabilities and the discrimination capability were utilized in weight function estimation. The deterioration of feature selection and classification performance occurs due to the relevant or irrelevant feature and noise in the traditional algorithms. Hence, this paper investigates the Weighing based Feature Selection and Monotonic Classification (WFSMC) to improve the performance.

3. Triangular Boundary Based Classification

The novel triangular boundary based classification approach includes training algorithm and testing algorithms for boundary layer computation of class labels. The training dataset for the class label c is defined as

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

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

$$x^i = [\, f_1^i, f_2^i, f_3^i \ldots f_m^i \,] \quad (2)$$

The triangular area is computed to find the relationship between the each features present in the instance x^i. 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. Triangular average for the class is calculated as

$$\overline{T_i^c} \leftarrow \frac{1}{q*r} \sum_{j=0}^{q} \sum_{k=0}^{r} TA_i^{jk} \quad (3)$$

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

$$Cov^c = \begin{bmatrix} \sigma(T_{2,1}^c, T_{2,1}^c) & \sigma(T_{2,1}^c, T_{3,1}^c) & \ldots \sigma(T_{2,1}^c, T_{m,m-1}^c) \\ \sigma(T_{3,1}^c, T_{2,1}^c) & \sigma(T_{3,1}^c, T_{2,1}^c) & \ldots \sigma(T_{3,1}^c, T_{m,m-1}^c) \\ & \ldots & \\ & \ldots & \\ & \ldots & \\ \sigma(T_{m,m-1}^c, T_{2,1}^c) & \sigma(T_{m,m-1}^c, T_{2,1}^c) & \ldots \sigma(T_{m,m-1}^c, T_{m,m-1}^c) \end{bmatrix} \quad (4)$$

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

$$\sigma(T_{u,v}^c, T_{x,y}^c) = \frac{1}{z-1} \sum_{j=1}^{z} (T_{u,v}^{c,j} - \mu_{T_{u,v}^c})(T_{x,y}^{c,j} - \mu_{T_{x,y}^c}) \quad (5)$$

$$\mu_{T_{u,v}^c} = \frac{1}{z} \sum_{i=1}^{z} T_{u,v}^{c,i} \quad (6)$$

The distance for i[th] Instance for class label is computed by using the equation

$$MD^{c,i}(TA_i^c, \overline{T^c}) = \left| \sqrt{\frac{(TA_i^c - \overline{T^c})^T (TA_i^c - \overline{T^c})}{Cov^c}} \right| \quad (7)$$

The upper and lower boundary layer for each class present in the training dataset are formulated by using normal distribution. These values helps to predict the accurate class for each instance. Overall Mean and Deviation are computed by using the equations given below

$$\mu = \frac{1}{z} \sum_{i=1}^{z} MD^{c,i} \quad (8)$$

$$\sigma \leftarrow \sqrt{\frac{1}{z-1} \sum_{i=1}^{z} (MD^{c,i} - \mu)^2} \quad (9)$$

$$Layer \leftarrow \mu + \sigma * \alpha \quad (10)$$

In the Training algorithm, the dataset X containing 'C' number of class labels is considered. The input dataset containing 'c' number of class labels and 'n' number of instances is obtained. Each instance contains 'm' number of features. The boundary layer prediction is built through the normal distribution function, density estimation of the distance between individual instance in the training dataset and the expectation of the 'z' number of training records. The mean of elements 'u' and 'v' is computed by using the equation (6). To find the normal distribution function, there is a need to compute the mean and deviation of the distance of all instance containing 'c' class labels by using equation (5) respectively. The distribution of the distance is described by mean and deviation of the c[th] class.

Training Algorithm

Input: X^c

Output: $\mu, \sigma, Cov^c, \overline{T^c}$

Procedure:

Step 1: Compute Triangular Area

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

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 μ
Step 11: Compute σ
Step 12: ***Return*** $\mu, \sigma, Cov^c, \overline{T^c}$

The boundary layer for each class label is computed by using equation (10). μ *and* σ Values are taken from the normal distribution function of each class from the training algorithm. α is a constant value varies from 0.5 to1.5. From this variance, prediction of class label varies from 91% to 98% in association with the selection of different values of α. Thus if the distance is calculated for testing instance and the respective value is greater than or lesser than the boundary layer. It is considered as an outlier and the other class label for the appropriate instance is predicted.

Testing Algorithm

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

In the testing phase, the triangular area of 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 (8). The computed distance is compared with each class boundary layer, and the class label is stored in the respective satisfied boundary layer class.

4. Weighing-Based Feature Selection and Monotonic Classification

The proposed Weighing based Feature Selection and Monotonic Classification includes the imputation methods, ordinal feature selection, and monotonic constraints are explained in this section.

A. Imputation of Dataset

The initial process of proposed WFSMC is weighing algorithm consists of two phases namely, imputation of dataset and computation of weight. The new estimated dataset (D) calculated from the original dataset (DS) by using imputation algorithm as follows:

Imputation method

Input: *Original dataset (DS), Imputation method (I)*

Output: *Estimated dataset* (DS_E)

1. Set DS_E=null;
2. **For** each example *eg* in *DS*
3. $eg' = \emptyset$.
4. **For** each feature F_i
5. $eg'(F_i')$=estimate the new value for $eg(F_i)$
6. **end**
7. $DS_E = DS_E \cup \{eg'\}$
8. **end**

Initially, the estimated dataset value is assumed to null. Then, new value (eg') is estimated for each example (eg) by using imputation method. The process of estimation of new values performed until the whole dataset is processed. The estimated new dataset holds the conditioned distributions for specific feature. The non-parametric test is performed to evaluate the changes of features as a similarity measure DS_n of two distributions (F_i', F_i). In general, the non-parametric test for two distribution functions (F_X, F_Y) with the corresponding indicator functions $(I_{X_i \leq x}, I_{Y_i \leq y})$ described by

$$F_X(x) = \frac{1}{N}\sum_{i=1}^{N} I_{X_i \leq x}, F_Y(y) = \frac{1}{N}\sum_{i=1}^{N} I_{Y_i \leq y} \quad (11)$$

The statistic measure (DS_n) between two functions by Kolmogorov – Smirnov statistic such that the degree of superiority (sup) of feature changes is follows:

$$DS_n = \sup_x |F_X - F_Y| \quad (12)$$

$$DS_n^i = \sup_i (F_i', F_i), \quad \forall i, F_i \in DS, F_i' \in DS_E \quad (13)$$

F_i- set of features in original dataset (DS) and

F_i' -set of features in the imputed dataset (DS_E). The values of statistic measure in the ranges from 0 to 1, i.e. $DS_n \in [0,1]$. The feature with low value of DS_n has small influence on similarity function and the feature with high DS_n value has high influence on similarity function. The statistic transformed to weight by using following equation (14)

$$W_i = DS_n^i / \sum_{j=1}^{N} DS_n^j \quad (14)$$

The weight factor computation used to highlight the influence of changed features on distance estimation in the selection process.

B. Ordinal Feature Selection

The margin based feature selection involved in this paper. The distance between the instance and distance border refers to margin. Based on the decision, margin provides the geometric measure of weighting of feature subset as shown in Fig. 1.

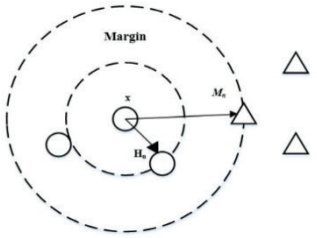

Fig. 1 Large margin

The ordinal based feature selection performed on the basis of monotonic constraints. The ordinal decision system contains finite set of instances X, attributes (At), decisions $d = (d1, \dots, dn))$ and the domain of attributes (D_a). In monotonic model, the domain and attributes follows the structural order. The nested decision structure is described as follows:

$$d_i^{\geq} = \bigcup_{j=i}^{N} d_j \; or \; d_i^{\leq} = \bigcup_{j=1}^{i} d_j \qquad (15)$$

In general, the function refers the monotone for all instances in original (X) and estimated dataset (Y) satisfies the constraint as follows:

$$x \geq y \Rightarrow F(x) \geq F(y) \qquad (16)$$

The weight updating rule in (14) combined with the monotonic constraint of k-nearest neighbours selects the ordinal based feature selection

Ordinal feature selection (With constraint)

For t=1: T
Select the instance from the space in random manner
Compute the nearest hit for each sample $H_{nj-}(x)$ and $H_{nj+}(x)$
Compute the nearest miss for each sample $M_{nj-}(x)$ and $M_{nj+}(x)$
For i=1: I
Update the weight factor by using equation (14)
End
End

Initially, the instance from the space is selected. Then, the nearest hit and miss for each instances are identified. Then the weight factor is updated by using the equation (14) for all the features (I). The computation of distances for each instance x requires the $O\ (mI)$ operations. The operations are doubled $O\ (m^2 I)$ for entire space $(S=m)$. Hence, ordinal based feature selection is more suitable in terms of computational complexity.

C. Monotonic Classification

The monotonic decision system contains N ordered classes. The order of the decision system arranged in the range as $d_1 \prec d_2 \prec \cdots \prec d_N$. For a given instance x, the instance with decision value $d_{k+1}^{\geq}$ provided the better features and the instance with $d_{k-1}^{\leq}$ provided the worse features with respect to attributes. The ordinal decision system contains three subsets. The k nearest misses $(M_{nj-}(x))$ for $d_{k-1}^{\leq}$ and $M_{nj+}(x)$ for $(d_{k+1}^{\geq})$ are computed. Similarly nearest hits $(H_{nj-}(x), H_{nj+}(x))$ are computed for d_k. Hence, there are two margins namely upward and down ward margin is constructed. The decision boundary is defined by using the combination of upward and downward margins as shown in fig. 2. The features with large values are termed as poor estimated features by the imputation method. Higher weight is assigned to it since they are preferred for distance estimation.

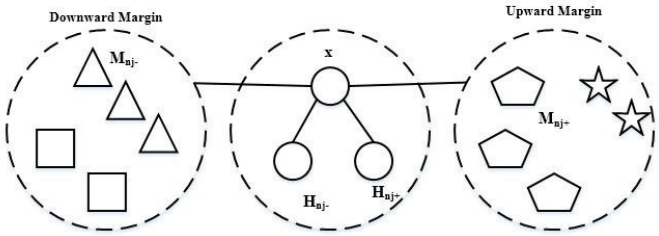

Fig. 2 Margin for monotonic classification

The features with low values are easily estimated by imputation methods. Lower weights are assigned to these features since they are not preferred for distance calculation.

5. Comparative Analysis

This section explains about the performance analysis results of the proposed Weighing based Feature Selection and Monotonic Classification (WFSMC). Our proposed approach is tested by using the WDBC dataset and results are compared with the Triangular Boundary-based Classification (TBC) regarding the parameters of no. of attributes, accuracy, precision, and recall.

A. 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 WDBC dataset are multi-variant dataset that comprises 569 instances and 32 attributes. WDBC describes the characteristics of cell-nuclei in an image.

B. No. of Attributes

The variation of attributes with the boundary-based classification algorithms as in Table 1 shows that the WFSMC utilizes less number of attributes.

The attributes required for TBC algorithm is 15 and for WFSMC 12. The weight computation and the ordinal feature selection in WFSMC reduce the attributes by 20 % compared to TBC.

TABLE 1
No. of Attributes Vs. Classification Algorithms

Algorithm	No. of attributes
TBC	15
WFSMC	12

C. Accuracy

The accuracy variation with the boundary-based classification algorithms as in Table 2 shows that the WFSMC offers an accurate OD.

TABLE 2
Accuracy Vs. Classification Algorithms

Algorithm	Accuracy (%)
TBC	97.54
WFSMC	98.54

The accuracy of TBC algorithm are 97.54 % and for WFSMC 98.54 %. The weight computation and the ordinal feature selection in WFSMC improve the accuracy by 1 % compared to TBC.

D. Precision

The ratio of the number of true positives to the total number of true positives and false positives refers to precision. The status of precision determines the false positive values. The high value of precision indicates the low value of false positives.

TABLE 3
Precision Vs. Classification Algorithms

Algorithm	Precision
TBC	0.9541
WFSMC	0.9631

Table 3 shows the visual representation of precision analysis with the boundary-based classification algorithms. The precision of TBC algorithm are 0.9541 and for WFSMC 0.9631. The weight computation and the ordinal feature selection in WFSMC improve the precision by 0.94 % compared to TBC.

E. Recall

The sensitivity or true positive rate defines the recall. Improving the recall can often decrease the precision, because it gets harder to be precise with the increase in the sample space.

TABLE 4
Precision Vs. Classification Algorithms

Algorithm	Recall
TBC	0.9811
WFSMC	0.9925

Table 4 shows the graphical representation of recall analysis with the boundary-based classification algorithms. The recall of TBC algorithm are 0.9811 and for WFSMC 0.9925. The weight computation and the ordinal feature selection in WFSMC improve the recall by 1.15 % compared to TBC.

6. CONCLUSION

This paper discussed the problems such as diverse large size dataset availability, classifier sensitivity of noise and irrelevant features in the Outlier Detection (OD) techniques. This paper compared the OD performance of Triangular Boundary-based Classification (TBC) and Weighing-based Feature Selection and Monotonic Classification algorithm on WDBC dataset. The ordinal feature selection employment prior to triangular boundary area isolated the relevant features from the irrelevant features. The boundary region analysis by normal distribution function in TBC supported an effective treatment to the patients. Then, the inclusion of monotonic constraints in classification phase improved the accuracy. The comparative analysis between TBC and WFSMC regarding the performance parameters of accuracy, precision, and recall showed the effective OD in real-time data mining applications.

8. REFERENCES

[1] C. C. Aggarwal, "Supervised outlier detection," in Outlier Analysis, ed: Springer, 2013, pp. 169-198.

[2] A. Daneshpazhouh and A. Sami, "Entropy-based outlier detection using semi-supervised approach with few positive examples," Pattern Recognition Letters, vol. 49, pp. 77-84, 2014.

[3] K. Noto, C. Brodley, and D. Slonim, "FRaC: a feature-modeling approach for semi-supervised and unsupervised anomaly detection," Data mining and knowledge discovery, vol. 25, pp. 109-133, 2012.

[4] S. Beniwal and J. Arora, "Classification and feature selection techniques in data mining," International Journal of Engineering Research & Technology (IJERT), vol. 1, 2012.

[5] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining," FSDM, vol. 10, pp. 4-13, 2010.

[6] 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.

[7] 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.

[8] 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.

[9] H. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Outlier detection in arbitrarily oriented subspaces," in IEEE 12th International Conference on Data Mining (ICDM), 2012 2012, pp. 379-388.

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

[11] 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.

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

[13] 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.

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

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

[16] P. Padungweang, C. Lursinsap, and K. Sunat, "A discrimination analysis for unsupervised feature selection via optic diffraction principle," IEEE Transactions on Neural Networks and Learning Systems, vol. 23, pp. 1587-1600, 2012.

[17] Y. Feng, J. Xiao, Y. Zhuang, and X. Liu, "Adaptive unsupervised multi-view feature selection for visual concept recognition," in Computer Vision–ACCV 2012, ed: Springer, 2013, pp. 343-357.

[18] L. Shi, L. Du, and Y.-D. Shen, "Robust Spectral Learning for Unsupervised Feature Selection," in IEEE International Conference on Data Mining (ICDM), 2014 2014, pp. 977-982.

[19] Y. Liu, F. Nie, J. Wu, and L. Chen, "Efficient semi-supervised feature selection with noise insensitive trace ratio criterion," Neurocomputing, vol. 105, pp. 12-18, 2013.

[20] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," IEEE Transactions on Neural Networks and Learning Systems, vol. 26, pp. 252-264, 2015.

[21] N. Sharma and H. Om, "Usage of Probabilistic and General Regression Neural Network for Early Detection and Prevention of Oral Cancer," The Scientific World Journal, vol. 2015, 2015.

[22] S. G. Jacob and R. G. Ramani, "Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multi-class Categorization of Breast Tissue Data," International Journal of Computer Applications (IJCA), vol. 32, pp. 46-53, 2011.

[23] G.-N. Zhu, J. Hu, J. Qi, J. Ma, and Y.-H. Peng, "An integrated feature selection and cluster analysis techniques for case-based reasoning," Engineering Applications of Artificial Intelligence, vol. 39, pp. 14-22, 2015.

[24] Y. Sun, X. Lou, and B. Bao, "A novel relief feature selection algorithm based on mean-variance model," Journal of Information & Computational Science, vol. 8, pp. 3921-3929, 2011.

[25] T. Yilmaz, A. Yazici, and M. Kitsuregawa, "RELIEF-MM: effective modality weighting for multimedia information retrieval," Multimedia systems, vol. 20, pp. 389-413, 2014.