Research Journal of Applied Sciences, Engineering and Technology 10(9): 1021-1028, 2015
DOI: 10.19026/rjaset.10.1869
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: February 27, 2015 Accepted: March 25, 2015 Published: July 25, 2015

Research Article

A Review of Outlier Prediction Techniques in Data Mining

[1]S. Kannan and [2]K. Somasundaram
[1]Department of Computer Science and Engineering, Karpagam University, Coimbatore, Tamilnadu 641021, India
[2]Department of CSE, Vel Tech High Tech Dr RR and Dr SR Engineering College, Avadi, Chennai-60062, India

Abstract: The main objective of this review is that to predict the outliers in data mining. In general, the data mining is a process of applying various techniques to extract useful patterns or models from the available data. It plays a vital role to choose, explore and model high dimensional data. Outlier detection refers a substantial research problem in the domain of data mining those objectives to uncover objects which exhibit significantly different, exceptional and inconsistent from rest of the data. The outlier potential sources can be noise and errors, events and malicious attack in the network. The main challenges involved in the outlier detection with high complexity, size and different types of datasets, are how to catch similar outliers as a group by using clustering-based approach. The outlier or noise available in the clustered data is accurately removed and retrieves an efficient high dimensional data. Nowadays, the classification and clustering techniques for outlier prediction are applied in various fields like bioinformatics, natural language processing, military application, geographical domains etc. This study surveys various data classification and data clustering techniques in order to identify the optimal techniques, which provides better outlier predicted data detection. Moreover, the comparison between the various classification and clustering techniques for outlier prediction are illustrated.

Keywords: Data classification, data clustering, data mining, high dimensional data, outlier detection

INTRODUCTION

Data mining is defined as the process that includes the extraction of interesting, interpretable, useful and novel information form of data. It has been used since several years in businesses, scientists and governments to sift through volumes of data like airline passenger records, census data and the supermarket scanner data that produces market research reports (Romero and Ventura, 2010). The major tasks involved in the data mining are.

Classification: A prediction learning function is discovered to classify a data item into multiple pre-defined classes.

Regression: The prediction learning function is discovered to map the data item to a real-value prediction variable.

Clustering: A common descriptive task seeks to identify a finite set of categories or clusters to describe the data.

Summarization: An additional descriptive task is involved in the process of finding a compact description for a subset of data.

Dependency modelling: A local model is found to describe significant dependencies between variables or between the values of a feature in the dataset.

Change and deviation detection: The most significant changes in the dataset are discovered.

The major objective of the data mining is further reduced into prediction and the classification. The prediction process predicts unknown or future values of interest by utilizing some variables or fields in the dataset (Koteeswaran and Janet, 2012). The classification process is used to find patterns by describing the data. In order to execute these processes data mining requires clustering and outlier analysis for reducing and identifying the useful dataset.

Outlier detection is also named as anomaly detection or deviation detection. It is one type of the fundamental tasks of data mining along with a predictive modelling, a cluster analysis and association analysis. These outlier detection is the nearest task to the initial aim behind the data mining. The detection of

Corresponding Author: S. Kannan, Department of Computer Science and Engineering, Karpagam University, Coimbatore, Tamilnadu 641021, India, Tel.: +91422 647 1115

outliers has regained considerable interest in data mining with the realization of outliers from very large databases. The outlier detection is divided into different univariate methods (Williams *et al*., 2002). A Replicator Neural Network (RNN) is one type of univariate approach for outlier prediction. This RNN approach is a non-parametric outlier detection method. In the class of non-parametric outlier detection method can set apart the data mining methods. It is also called as distance-based methods. Usually these distance-based methods are performed depending on local distance measures to handle large databases.

The outlier prediction incorporates with both statistical and data mining methods and the datasets. It encompasses the aspects of broad spectrum of techniques. Many techniques are fundamentally identical and employed to detect outliers. It is a critical task in several safety critical environments as the outlier that indicates the abnormal running conditions (Hodge and Austin, 2004). The important exhaustive lists of applications in the outlier prediction are as follows:

- **Fraud detection:** The fraudulent applications are detected for credit cards, state benefits or detecting fraudulent usage of credit cards or mobile phones
- **Intrusion detection:** The unauthorized access in the computer networks is detected
- **Network performance:** The performance of the computer networks is monitored to detect the network bottleneck
- **Fault diagnosis:** The faults in the data are detected by monitoring the process
- **Structural defect detection:** The manufacturing lines are monitored to detect fault production runs
- Detecting mislabeled data in a training dataset

The outlier detection in large data sets is an active research field in the data mining. It contains multiple applications in the domain that leads to illegal or abnormal behavior such as error detection, network intrusion detection and so on. In this survey, various outlier detection techniques in the data mining are clearly stated. The issues are formulated in this review and also the drawbacks of existing techniques for data mining clearly. By using this survey, the better outlier detection technique is selected to provide high outlier detection rate.

MATERIALS AND METHODS

Classification criteria of outlier detection techniques for data mining: A significant amount of work has been performed in the data mining for outlier prediction. Various models are subsequently proposed in the literature survey, including, clustering-based approaches and classification-based approaches. This section identifies various important aspects of the outlier detection techniques in the data mining.

Nature of input data: This act as a key aspect of some outlier detection technique. Generally, an input is a collection of data instances. Each data instance is described by using a group of attributes. These attributes consists of various types such as binary, categorical or continuous. The outlier detection techniques analyze the data by determining the type of input data. Usually, the outlier detection techniques consider two following aspects of sensor data.

Attributes: A measurement of data was identified as outliers when its attributes contains anomalous values. If the outlier is in univariate data with single attribute then it can easily detected. Thus, the outlier detection techniques for data mining have a capability to analyze the multivariate data. These analyzes part improves the accuracy of the outlier detection techniques.

Correlations: Two types of dependencies are available in each sensor nodes, i.e., dependencies among the attributes of the sensor nodes and the dependency of sensor node readings on history and neighboring node readings. The mining accuracy and computational efficiency are enhanced by capturing the attribute correlations. It is mainly used to analyze distinguish between error and events.

Normally, the outliers are classified into two types namely, local outliers and global outliers. The local outlier detection can is used in several event detection applications. Based on the network architecture, the identification of global outlier is performed at various network levels.

Identity of outliers: There are three different types of outlier sources occurred in data mining.

Errors: Generally, an error refers to a noise-related measurement or data from a faulty sensor. These erroneous data is represented as an arbitrary change and is totally varies from the rest of the data. These errors are discarded to save transmission power and have low energy consumption.

Events: An event is defined as a particular phenomenon, which changes the real-world state. This sort of outlier normally lasts for long time period and modifies the historical pattern of sensor data.

Outlier detection techniques in data mining: In this section, we focused on the earlier work carried out for outlier detection techniques in data mining based on performance analysis of each outlier detection technique. Furthermore, we provide deep explanation for each of these techniques.

Supervised methods for outlier detection: Typically, a labeled set of anomalous data instances that cover all different type of anomalous behavior. It has more difficulty than getting labels for normal behavior (Singh and Upadhyaya, 2012). Moreover, these outlier

behavior is either dynamic in nature. Depending on the availability of the label, this outlier detection technique operates in one of the following modes.

Supervised outlier detection technique: In these supervised outlier detection, the availability of a training data set is assumed. It contains labeled instances for normal and also the outlier class. It has a capability to create the predictive models for both normal and outlier classes (Chandola *et al*., 2007). The explicit notion of both normal and outlier behavior is located in the supervised outlier detection techniques in an accurate manner. The main disadvantage involved in this technique is that to label the training data in an accurate manner. Hence, it is more prohibitively expensive to obtain. These labeling is manually performed by a human expert and also needs a lot of effort to obtain the labeled training data set. These types of techniques insert artificial outliers into normal data set to obtain a fully labeled training data set. After that the supervised outlier detection techniques is applied to detect outliers in testing data.

Semi-supervised outlier detection technique: These techniques generally assume the availability of labeled instances for training is not very popular. It is more difficult to gather labels for other class (Bhosale, 2014). Typically, this approach of these techniques is to model only the available class. The major difficulty for their limited popularity is that to obtain a training data set that encloses every possible outlying behavior that can occur in the data. The behavior that does not exist in the training data is harder to detect as outliers.

Unsupervised outlier detection technique: The training data does not needed in this technique that operate in unsupervised mode and thus are most widely applicable (Koupaie *et al*., 2013). The techniques in this category create the implicit assumption that normal instances are more frequent than outliers in the test data. This technique has high false alarm rate.

Statistical-based approaches: It is one of the earliest approaches that deal with the issues of the outlier detection these approaches are essentially model-based techniques. In this approach, a data point is consider as an outlier (Chandore and Chatur, 2013). It also further classified into parametric and non-parametric depends on a probability distribution model.

Parametric approach: In this approach, the knowledge availability about underlying data distribution (Saini *et al*., 2014). Distribution parameters were estimated from the available data namely, Gaussian-based models and non-Gaussian based models. In these Gaussian-based models, the drawback of this technique is that it relies on the select the appropriate values of the threshold. The non-Gaussian based models further detect outliers that deviate unusually from other normal data. It is mainly used for univariate data. Therefore, it was unsuitable for multi-dimensional datasets and also most of the time the real-world data are multivariate with unknown distribution.

Non-parametric based approaches: The availability of the data distribution does not assume in this approaches. The histograms and the kernel density estimator are most widely used categories in this approach. Here, the kernel density estimator uses the kernel functions to evaluate the probability distribution function for the normal instances and it contains low probability (Kumar, 2014). This approach is quite useful when compared with the parametric-based approach. However, this approach has high computational complexity due to high-dimensional data sets.

Clustering-based approaches: In this clustering-based approach, similar data instances in most popular approaches in the data mining community are grouped into clusters with similar behavior.

Similarity based clustering technique: In this technique, the similarities of two sets of clusters are measured by defining a simple formula: Let $C = \{C_1, C_2..C_m\}$ and D = $\{D_1, D_2..D_n\}$ may be considered as a result of the two clustering algorithms on the same data set. Consider *C* and *D* as hard or exclusive clustering algorithms (Torres *et al*., 2009), where these clusters produced are pair-wise disjoint, i.e., each pattern from the dataset corresponds to the cluster. Then the similarity matrix for *C* and *D* is an m x n matrix $S_{C,D}$:

$$S_{C,D} = \begin{bmatrix} S_{11} & S_{12} \ldots\ldots & S_{1j} \ldots\ldots & S_{1n} \\ S_{i1} & S_{i2} \ldots\ldots & S_{ij} \ldots\ldots & S_{in} \\ S_{m1} & S_{m2} \ldots\ldots & S_{mj} \ldots\ldots & S_{mn} \end{bmatrix} \quad (1)$$

where, $S_{ij} = {}^{p}/_{q}$, which is Jaccard's similarity coefficient, p denotes the size of the intersection, q denotes the size of the union of cluster sets C_i and D_j. The similarity of the clustering C and clustering D are then defined as:

$$\text{Similarity (C, D)} = \sum_{i \le m,\ j \le n} {}^{S_{ij}}\Big/_{\max(m,n)} \quad (2)$$

This Jaccard's Similarity is chosen as a better similarity measure technique (Suphakit *et al*., 2013), because it accurately compares the proximity of the data in the data clustering. It has a capability to

Table 1: Computational complexity of clustering algorithm

Clustering algorithm	Computational complexity	Capability of tracking high dimensional data
Similarity-based clustering technique	Less complexity	No
Fuzzy k-means clustering	O (NKd) (time) O (N + k) (space)	No

overcome all the desirable properties such as non-negativity, identity and symmetry and triangle inequality of a distance metric. This technique separates the dataset according to their characteristic. In this clustering approach, the analysis and utilization of the dataset are easy while retrieving and extracting the information from the database compared with a hierarchical clustering technique.

Incremental clustering approach: In this technique, the outliers available in the database are detected (Gupta *et al*., 2014). In this incremental algorithm, the incremental part is mined for discovering the frequent pattern. The main goal of the incremental algorithm is to add the new data to the original transaction database. This algorithm not only includes the new item set into a tree, but also disposes the infrequent item set from utility pattern tree structure. Finally, the incremental database is rearranged and the high utility item set is discovered. By using this algorithm, it is possible to separate the dataset into several clusters (Su *et al*., 2009).The major advantage involved in this technique is that there is no need to store the entire dataset in the memory. Because of this, the space and time requirements of this algorithm are very small and are very easy to implement. It has been successfully used in the engineering applications.

Fuzzy k-means clustering technique: It is often used in the modeling such as fuzzy modeling, neural networks and rule-based systems and is a method of clustering (Jose *et al*., 2014). Some difficulties exist in the fuzzy clustering approach are described below: At first, the optimal number of clusters is created to determine the number of cluster and define the apriori and better cluster validity criterion. It is an unsupervised algorithm used in this clustering approach (Rahmani *et al*., 2014). It smartly compares the centroid with the data points depends on their intensity and characteristics on the data points. In this technique, the learning algorithm needs apriori specification of the number of clusters. It may not have the capability to handle the noisy data and the outlier. It does not suitable for non-linear dataset. An outlier prediction is obtained in the clustered data. Table 1 provides the computational complexity of various clustering algorithms.

The high-dimensional data in the database is preprocessed to provide better performance to remove the undesired information in the image with better quality. The preprocessed data is classified according to the characteristics of the data. The classified data is clustered based on the similarity of the data. Finally, the outlier or noise available in the clustered data is accurately removed.

Classification-based approaches: This approach is a major systematic approach in the data mining and the machine learning community. The outcome is predicted by using this classification approaches (Shukla *et al*., 2014). A training set is trained by using this approach that contains a set of attributes and the respective outcomes. It is named as predication attributes. It provides an exact set of outliers by building a classification model to perform classification. In existing outlier detection techniques, the classification-based approaches are further categorized into Support Vector Machine and Bayesian network-based approaches.

Support vector machine: This classifier is mainly used for classification and regression analysis. This classification has been performed well as a computer-aided diagnostic classification mechanism for data. This classifier creates a hyper plane or set of hyper planes in a high or infinite dimensional space for classification. Here, the hyper plane represents the largest separation between the classes. The SVM constructs a separation boundary based on a small portion of training data called support vectors (Peter *et al*., 2013). The main characteristics involved in this classifier are that the SVM uses the non-parametric with binary classifier approach. It efficiently handles multiple input data. The performance and accuracy of the SVM depends on the hyperplane selection and kernel parameter. The hyperplane is selected according to the separation of the classes. It contains a non-linear transformation to provide better generalization capability. An over-fitting problem occurred during the classification is eliminated. It has low computational complexity and is simple to manage decision rule complexity and error frequency. This classifier has high flexibility so as to provide better ability to handle large amount of data sets when compared with the existing techniques. The technique also eliminates spatial correlation of neighboring nodes, which makes the results of local outliers inaccurate.

Bayesian network-based approaches: A typical Bayesian network is for outlier detection that is aggregated with information from different variables. It provides an estimation of the expectancy of the event according to the normal class. A probabilistic graphical model is used by the Bayesian network-based approaches to denote a set of variables and their probabilistic independencies. Depends on the degree of

probabilistic independencies among variables. There are several approaches that are centralized as well as distributed to the training data in the learning phase. The training data is tested until accurate results are determined and then it is used for inference.

Naïve Bayesian models: In the Naïve Bayesian models, the probability of obtained result is evaluated (Tien Bui *et al*., 2012). The classification is performed for each class according to the largest posterior probability. These Naïve Bayesian model is obtained overall classification accuracy in average.

Bayesian Belief Network (BBN) models: In Bayesian Network-Based, when the active molecules contains a high degree of structural homogeneity. Therefore, it substantially outperformed a conventional Tanimoto-based similarity searching system. To overcome the above drawback (Abdo *et al*., 2014), an alternative Bayesian network model is introduced that is named as a Bayesian belief network. These BBN model is proposed for similarity searching (Abdo *et al*., 2010) that is applied for classification of small molecules in the activity classes. These BBN model is act as a useful tool for prediction of unusual activity. This technique determines local outliers via collaboration of neighboring nodes.

Dynamic Bayesian Network (DBN) models: The DBN is a hybrid DBN that contains both discrete and continuous variables (Dabrowski and De Villiers, 2015). Various behaviors of vessels in a maritime piracy situation are modelled by the DBN. These behaviors consist of various activities such as sailing, target acquisition and attacking.

The main disadvantage of this network is that the results are similar to those derived from threshold-based systems. It has high computational effort when compared with the SVM classifier.

Decision tree classifier: A decision tree is a flow chart like tree structure: here each internal node denotes a test on an attribute. Each branch denotes an output of the test. Each leaf node in the tree holds a class label. Two phases of the decision tree classifier are growth phase or build phase and pruning phase (Bal *et al*., 2014). A problem of over fitting the data in the decision tree is handled by the pruning phase. Classic decision tree classification belongs to supervised learning methods (Lu and Braunstein, 2014). This phase generates the tree by removing the noise and outliers in the data to improve the classification accuracy. Irrelevant attributes affects construction of a decision tree. The classifications of data are very sensitive to training data.

Decision making technique: This technique is mainly used to analyze and prioritize the issues available in need assessment process. It is a post-processing tool that helps the user to make selection of a better design (Mosavi, 2010). It can select the course of action from multiple possibilities. The Multi-Objective Optimization (MOO) approach is used to find the Pareto solutions. This decision-making procedure is used to select the particular solution from the Pareto solutions set for implementation in Multiple Criteria Decision-Making (MCDM) process. It deals with the problems of the MCDM and obtains the solution that is close to the true optimal solution. The main steps involved in the decision-making algorithm are:

1. Match the new objects feature value with the features of the alternative decision rules generated by the rule extraction algorithm. If the match is satisfactory, go to Step 2; otherwise, go to Step 3.
2. Assign the primary decision to the new object equal to the decision associated with the matching decision rules and go to Step 4.
3. Output "No primary decision is made-More feature values are needed" and go to Step 4.
4. If all new objects have been considered, stop; otherwise, go to Step 1.

The alternative rules used in Step 1 are used to increase the value of Decision Redundancy Factor (DRF). The outcome of this decision-making algorithm does not satisfy the user requirements corresponding to the DRF value. The main reason for this problem is that the lacks of confidence, outliers in the data, missing data and so on. To avoid this application barrier, an orthogonal decision is generated with below confirmation algorithm. To denote this conformation algorithm, an absolute distance measure $d_{i,j}$ between objects i and j is denoted as:

$$d_{i,j} = \sum_{k=1}^{n} \left| f_{ik} - f_{jk} \right| \quad (3)$$

where, f_{jk} the object value of feature (f) and n is the number of specific features. In this technique, group decision quality will reflect several complete informational landscape by capturing the unique information. It has high decision quality while classifying the data in large dataset when compared to the Bayes network.

RESULTS AND DISCUSSION

Various techniques for outlier prediction in data mining are depicted. The results of the survey are shown in Table 2. From this survey, it is evident nature of input data to predict the outlier during data mining. The extracted data from the database is classified by using Support Vector Machine and Decision Making

Table 2: Information about various outlier prediction techniques

Techniques	Author and reference	Year	Performance	Data set	Quality measurement
Statistical-based approaches					
Parametric approach	Singh and Kumar (2013)	2013	In this approach, the outlier is detecting based on the general pattern within data points. It combine a Gaussian mixture model and a supervised method. This approach is can also referred as parametric approach.	• Fisher's iris dataset • Bupa data set	• Number of outliers detected • Time complexity
Non-parametric approach	Fan *et al*. (2006)	2006	In the non-parametric approach, the optimum clustering of a dataset is obtained depends on resolution change. The clusters in the dataset redistribute, when the resolution modifies on a dataset.	• Synthetic datasets • Real life construction equipment dataset	• Top-20 outliers identified by RB-outlier in a 200-tuple synthetic dataset • Top-200 outliers identified by RB-outlier in a 10,000-tuple synthetic dataset
Clustering-based approaches					
Similarity based clustering technique	Torres *et al*. (2009)	2009	It has a capability to overcome all the desirable properties such as non-negativity, identity, symmetry and triangle inequality of a distance metric.	• Portuguese dataset • Iris dataset	• Methodology for calculating similarity measure of clustering the Portuguese dataset • Iris dataset
	Zhang *et al*. (2012)	2012	In this technique, the distributed and online outlier detection techniques are presented. It is further divided to temporal, spatial and spatial-temporal outlier detection based on the usage of temporal and spatial correlations.	• Reference dataset	• Labelled data using running average-based labelling technique at node 29 • Temporal outliers detected at node 29 by Temporal Outlier Detection (TOD) • Number of outliers and events detected at different nodes using TOD • Complexity analysis of our outlier detection techniques for each sensor node
Incremental clustering approach	Gupta *et al*. (2014)	2014	The incremental part in this approach is mined for discovering the frequent pattern. The main objective of the incremental algorithm is to update the new data to the original transaction database.	• Spatio-temporal dataset	• Stream anomaly detection • Distance based outliers for sliding windows
	Koupaie *et al*. (2013)	2013	It mainly represents a cluster-based outlier detection available in the data stream. Here, the incremental clustering algorithm is used to identify the real outlier in the stream data.	• Static dataset	• High accuracy
Fuzzy k-means clustering technique	Rahmani *et al*. (2014)	2014	A learning algorithm requires apriori specification of the number of clusters. It may not have a capability to handle the noisy data and the outlier.	• Geometric structure data sets	• High fuzzy factor • Less clustering time
	Huang *et al*. (2005)	2005	It recovers the clusters in data. The synthetic data experiments contain demonstrated that the weights can effectively differ noise variables from the normal variables.	• Constructed synthetic data set • Australian credit card data set • Heart diseases data set	• The relationship between clustering accuracy and the value of the objective function • Heart diseases data
Classification-based approaches					
Support vector machine	Luo and Li (2014)	2014	It contains a non-linear transformation to provide better generalization capability. An over-fitting problem occurred during the classification is eliminated.	• Reuters-21578 dataset • Small-scale test datasets	• The values of macro-precision, macro-recall, macro-F1 and micro-F1 under different number of features reduced by DF method • The values of macro-precision, macro-recall, macro-F1 and micro-F1 under different number of features reduced by PCA method • Relationship between accuracy and the number of Gibbs sampling iterations on reuters-21578 dataset

Table. 2: Continue

Techniques	Author and reference	Year	Performance	Data set	Quality measurement
	Yin *et al*. (2014)	2014	It is mainly used in machine fault detection. The fundamental principle of SVM is separating dataset into two classes.	• Training dataset	• Results of faults detection using partial least square algorithm • Results of faults detection with one against one classifier • Classification results for the testing dataset by the SVM-based classifier with optimal parameters
Bayesian network-based approaches	Abdo *et al*. (2014)	2014	It provides interesting prediction rates (from 79% accuracy for a high diverse data set to 99% for low diverse data set) with short time calculation for activity prediction.	• Homogeneous data sets • Eight well-known data sets • MDL Drug Data Report (MDLDRR) 1 datasets • MDLDRR 2 dataset • Norine data set	• Sensitivity, specificity, AUC and accuracy rates for the prediction models for different data sets • Sensitivity, specificity, AUC and accuracy rates for the prediction models with the Norine data set
	Tien Bui *et al*. (2012)	2012	It is randomly partitioned into 70% for training the models and 30% for the model validation.	• Training and validation datasets	• Minimum number of instances per leaf versus classification accuracy • Confidence factor used for pruning versus classification accuracy • High accuracy • High prediction capability
Decision tree classifier	Bal *et al*. (2014)	2014	Data-driven systems operates in large data stacks and support the decision making process by using data mining methods. Some of these operation refers as Bayes networks.	• Training datasets • Synthetic data sets • Real and simulated datasets	• Classification accuracies with datasets having 10 samples • High classification accuracy
	Lu and Braunstein (2014)	2014	A decision tree classifier acquires from a training dataset that consists of observations about objects.	• Quantum training dataset • Training dataset	• High classification accuracy
Decision making technique	Mosavi (2010)	2010	A classification task of data mining is act as an effective option for identifying the effective variables of the Multiple Criteria Decision-Making (MCDM) systems.	• Training dataset	• Low growing complexity • High classification accuracy • High decision quality

technique provide better result than the existing technique such as Bayes Network. Also, a similarity based clustering can efficiently cluster the data to predict the outliers in the data in. Moreover, the surveyed result evidently proves that the Incremental classification approach is used to predict the outliers in the data with better quality.

CONCLUSION

Outlier prediction is a major problem with direct application in a wide variety of domains. A key observation with outlier detection is that it is not a well-formulated problem. In this survey, different ways in which the problem has been formulated in literature are discussed. The shortcomings of existing techniques for data mining clearly define for enhancing outlier detection techniques. This outlier prediction is performed in different approaches such as statistical-based approaches, clustering-based approaches and classification-based approaches. Table 2 describes the description, quality measurement and dataset of each approach. In statistical-based approach, the comparison is made between the parametric approach and the non-parametric approach. The similarity-based clustering and the incremental clustering techniques are chosen as a better choice while compared with the fuzzy k-means clustering. In classification-based approaches, the comparison is made between support vector machine-based approaches, Bayesian network-based approaches, decision tree classifier and decision making technique.

REFERENCES

Abdo, A., B. Chen, C. Mueller, N. Salim and P. Willett, 2010. Ligand-based virtual screening using bayesian networks. J. Chem. Inf. Model., 50(6): 1012-1020.

Abdo, A., V. Leclère, P. Jacques, N. Salim and M. Pupin, 2014. Prediction of new bioactive molecules using a Bayesian belief network. J. Chem. Inf. Model., 54(1): 30-36.

Bal, M., M.F. Amasyali, H. Sever, G. Kose and A. Demirhan, 2014. Performance evaluation of the machine learning algorithms used in inference mechanism of a medical decision support system. Sci. World J., 2014(2014): 15, Article ID 137896.

Bhosale, S.V., 2014. Holy grail of outlier detection technique: A macro level take on the state of the art. Int. J. Comput. Sci. Inform. Technol., 5(4): 5872-5874.

Chandola, V., A. Banerjee and V. Kumar, 2007. Outlier detection: A survey. ACM Comput. Surv., pp: 1-83.

Chandore, P. and P. Chatur, 2013. Hybrid approach for outlier detection over wireless sensor network real time data. Int. J. Comput. Sci. Appl., 6(2): 76-81.

Dabrowski, J.J. and J.P. De Villiers, 2015. Maritime piracy situation modelling with dynamic bayesian networks. Inform. Fusion, 23: 116-130.

Fan, H., O.R. Zaïane, A. Foss and J. Wu, 2006. A nonparametric outlier detection for effectively discovering top-n outliers from engineering data. In: Ng, W.K., M. Kitsuregawa and J. Li (Eds.), PAKDD, 2006. LNAI 3918, Springer-Verlag, Berlin, Heidelberg, pp: 557-566.

Gupta, M., J. Gao, C. Aggarwal and J. Han, 2014. Outlier detection for temporal data. Synthesis Lect. Data Mining Knowl. Discov., 5(1): 1-129.

Hodge, V.J. and J. Austin, 2004. A survey of outlier detection methodologies. Artificial Intell. Rev., 22(2): 85-126.

Huang, J.Z., M.K. Ng, R. Hongqiang and L. Zichen, 2005. Automated variable weighting in k-means type clustering. IEEE T. Pattern Anal., 27(5): 657-668.

Jose, A., S. Ravi and M. Sambath, 2014. Brain tumor segmentation using k-means clustering and fuzzy c-means algorithms and its area calculation. Brain, 2(3): 3496-3501.

Koteeswaran, S. and P.V. Janet, 2012. A review on clustering and outlier analysis techniques in data mining. Am. J. Appl. Sci., 9(2): 254-258.

Koupaie, H.M., S. Ibrahim and J. Hosseinkhani, 2013. Outlier detection in stream data by clustering method. Int. J. Adv. Comput. Sci. Inform. Technol., 2(3): 25-34.

Kumar, M., 2014. Evaluating the existing solution of outlier detection in WSN system. Int. J. Adv. Res. IT Eng., 3(6): 16-25.

Lu, S. and S.L. Braunstein, 2014. Quantum decision tree classifier. Quantum Inf. Process., 13(3): 757-770.

Luo, L. and L. Li, 2014. Defining and evaluating classification algorithm for high-dimensional data based on latent topics. PloS One, 9(1): 1-9.

Mosavi, A., 2010. Multiple criteria decision-making preprocessing using data mining tools. Int. J. Comput. Sci. Issues (IJCSI), 7(2): 26-34.

Peter, T., Z. Michael and U. Stan, 2013. Value-at-risk support vector machine: Stability to outliers. J. Comb. Optim., 28: 218-232.

Rahmani, M.K.I., N. Pal and K. Arora, 2014. Clustering of image data using k-means and fuzzy k-means. Int. J. Adv. Comput. Sci. Appl. (IJACSA), 5(7): 160-163.

Romero, C. and S. Ventura 2010. Educational data mining: A review of the state of the art. IEEE T. Syst. Man Cy. C, 40(6): 601-618.

Saini, A., K.K. Sharma and S. Dalal, 2014. A survey on outlier detection in WSN. Int. J. Res. Aspects Eng. Manage., 1(2): 69-72.

Shukla, D.S., A.C. Pandey and A. Kulhari, 2014. Outlier detection: A survey on techniques of WSNs involving event and error based outliers. Proceeding of Innovative Applications of Computational Intelligence on Power, Energy and Controls with their impact on Humanity (CIPECH), pp: 113-116.

Singh, G. and V. Kumar, 2013. An efficient clustering and distance based approach for outlier detection. Int. J. Comput. Trends. Technol. (IJCTT), 4(7): 2067-2072.

Singh, K. and S. Upadhyaya, 2012. Outlier detection: applications and techniques. Int. J. Comput. Sci. Issues (IJCSI), 9(1): 307-323.

Su, X., Y. Lan, R. Wan and Y. Qin 2009. A fast incremental clustering algorithm. Proceeding of the International Symposium on Information Processing (ISIP'09), pp: 175-178.

Suphakit, N., S. Jatsada, N. Ekkachai and W. Supachanun, 2013. Using of jaccard coefficient for keywords similarity. Proceeding of the International Multi Conference of Engineers and Computer Scientists, Vol. 1.

Tien Bui, D., B. Pradhan, O. Lofman and I. Revhaug, 2012. Landslide susceptibility assessment in vietnam using support vector machines, decision tree and naive bayes models. Math. Probl. Eng., 2012(2012): 26, Article ID 974638.

Torres, G.J., R.B. Basnet, A.H. Sung, S. Mukkamala and B.M. Ribeiro, 2009. A similarity measure for clustering and its applications. Int. J. Electr. Comput. Eng. Syst. (IJECES), 3(3): 164-170.

Williams, G., R. Baxter, H. He, S. Hawkins and L. Gu 2002. A comparative study of RNN for outlier detection in data mining. Proceeding of IEEE 13th International Conference on Data Mining, pp: 709-709.

Yin, S., X. Gao, H.R. Karimi and X. Zhu, 2014. Study on support vector machine-based fault detection in Tennessee eastman process. Abstr. Appl. Anal., 2014(2014): 8, Article ID 836895.

Zhang, Y., N.A. Hamm, N. Meratnia, A. Stein, M. van de Voort and P.J. Havinga, 2012. Statistics-based outlier detection for wireless sensor networks. Int. J. Geogr. Inf. Sci., 26(8): 1373-1392.