IJCTA, 9(36), 2016, pp. 517-531

EMPWC: Expectation Maximization with Particle Swarm Optimization based Weighted Clustering for Outlier Detection in Large Scale Data

J. Rajeswari* and R. Gunasundari**

Abstract : Outlier detection is usually considered as a pre-processing step for locating in a data set, those objects that do not conform to well-defined notions of expected behaviour. It is very important in data mining for discovering novel or rare events, anomalies, vicious actions, exceptional phenomena etc. However, investigation of outlier detection for categorical data sets is especially a challenging task because of the difficulty of defining a meaningful similarity measure. In addition, one might have to determine the optimal value of outliers, that is, how many outliers a data set really has. A possible theoretical approach to this problem is to search for a range of values of outliers (o) and decide on an optimal value of outlier (o) by optimizing certain variational property. Because of this reason, Particle Swarm Optimization (PSO) method is introduced in this paper to search for a range of values of outliers (o). The proposed work consists of three major steps: (*i*) define a function for outlier factor (*ii*) optimization value of outlier and (*iii*) clustering methods for outlier detection. In the first step of the work, define a new concept of entropy that takes both Shannon and Jensen-Shannon Divergence (JSD) into consideration. The second step, PSO is introduced to search outliers. Here, the PSO includes n number of data samples N which are moving around a D-dimensional search space for optimizing a certain variational property. Based on this PSO, define a function for the outlier factor of an object which is solely determined by the object itself and can be updated efficiently. Finally, propose EMPWC outlier detection method which requires no user-defined parameters for deciding whether an object is an outlier. In addition to this EMPWC based outlier detection methods associate a weight from entropy function with each observed dataset samples. Here, introduce the weighted-data Gaussian mixture and EM algorithms. The first one considers a weight for each categorical data attributes. The second one treats each weight and detects outliers. The experiment results on large scale categorical datasets demonstrate that the proposed EMPWC based outlier detection methods can achieve a better tradeoff between Detection Rate (DR), False Alarm Rate (FAR) when compareing to state-of-the-art outlier detection approaches.

Keywords : Outlier Detection, Attribute Weighting, Shannon and Jensen-Shannon Divergence (JSD), Particle Swarm Optimization (PSO) and Expectation Maximization (EM).

1. INTRODUCTION

Outlier detection is an active research area [1-2], refers to the problem of finding objects in a data set that do not conform to well-defined notions of expected behavior. The objects detected are called outliers, also referred to as anomalies, surprises, aberrants, etc. Outlier detection can be implemented as a preprocessing step prior to the application of an advanced data analysis method. It can also be used as an effective tool to discover the interest patterns such as the expense behavior of a to-be bankrupt credit cardholder. Outlier

* Research scholar, Karpagam University, Karpagam Academy of Higher Education, Coimbatore

** Head Dept of Information Technology, Karpagam University, Karpagam Academy of Higher Education, Coimbatore.

detection is an essential step in a variety of practical applications including intrusion detection [3], health system monitoring and criminal activity detection in E-commerce [4], and can also be used in scientific research for data analysis and knowledge discovery in biology, chemistry, astronomy, oceanography, and other fields [5].

A few studies were conducted on outlier detection for large datasets [6]. Many data mining algorithms try to minimize the influence of outliers or eliminate them together. However, this could result in the loss of important hidden information since one person's noise could be another person's signal. In other words, the outliers themselves may be of particular interest, such as in the case of fraud detection, where outliers may indicate fraudulent activity. Outlier detection or outlier mining is the process of identifying outliers in a set of data. The outlier detection technique finds the applications in credit card fraud, network robustness analysis, network intrusion detection, financial applications and marketing. Thus, outlier detection and analysis is an interesting and important data mining task.

According to [1], [2], if the existing methods for outlier detection are classified according to the availability of labels in the training data sets, there are three broad categories: supervised, semi-supervised, and unsupervised approaches. In principle, models within the supervised or the semi-supervised approaches need to be trained before use, while models adopting the unsupervised approach do not include the training phase. Moreover, in a supervised approach, a training set should be provided with labels for anomalies as well as labels of normal objects, in contrast with the training set with normal object labels alone required by the semi-supervised approach. On the other hand, the unsupervised approach does not require any object label information. Thus, the three approaches have different prerequisites and limitations, and they fit different kinds of data sets with different amounts of label information. The three broad categories of outlier detection techniques are discussed below.

The supervised anomaly detection approach learns a classifier using labelled objects belonging to the normal and anomaly classes, and assigns appropriate labels to the test objects. The supervised approach is studied extensively and many methods developed. The work of Barbara´et al.[7] is based on statistical testing and an application of Transduction Confidence Machines, which requires k neighbors. Moreover, variety of methods [8] based on information theory are also proposed. However, this method controls the false positive rate in the novelty detection problem.

The semi-supervised anomaly detection approach primarily learns a model representing normal behavior from the given training data set of normal objects, and then calculates the likelihood of a test objects are generated by the learned model. Zhang et al.[9] propose an adapted hidden Markov model for this approach to anomaly detection, while Gao et al.[10] propose a clustering-based algorithm which punishes deviation from known labels. Methods that assume availability of only the outlier objects for training occur in the data are rare [2], because it is difficult to obtain a training data set which covers all possible abnormal behavior that can occur in the data

The unsupervised anomaly detection approach detects anomalies in an unlabeled data set under the assumption, the majority of the objects in the data set are normal. The techniques in this category make the implicit assumption that normal instances are far more frequent than outliers in the test data. If this assumption is not true then such techniques suffer from high false alarm rate. For example, parametric statistical techniques assume a parametric distribution for one or both classes of instances. Several techniques make the basic assumption that normal instances are far more frequent than outliers. Thus a frequently occurring pattern is typically considered normal when a rare occurrence is an outlier. Local Distance-Based Outlier Detection Factor (LDOF) uses the relative distance from an object to its neighbors to measure how much objects deviate from their scattered neighborhood. The higher outlier factor, it more likely the point is an outlier. It is observed that outlier detection schemes are more reliable when used in a top-n manner. This means that the top n factors are taken as outliers, the n is decided by the user as per the requirements [11]. So, unsupervised method is used as the first step to find a candidate set of outliers, which help experts to build the training data set. The unsupervised approach is the research focus of this work.

In real applications, a large portion or the entirety of the data set is often presented in terms of categorical attributes. Examples of such data sets include transaction data, financial records in commercial banks, demographic data, etc. The problem of outlier detection in this type of data set is more challenging since there is no inherent measurement of distance between the objects. However, they cannot be easily adapted to deal with categorical data. The early research in the field of outlier detection is based on the statistical methods [12]. The literature in this field defines an outlier as an observation, which appears to be statistically inconsistent with the remainder of the data set. The statistical methods are parametric methods that assume a known underlying distribution or statistical model of the given data. According to these methods, outliers are the objects that have low probability of belonging to the statistical model. However, these approaches are not effective even for moderately high dimensional spaces. Also, finding the right model is often a difficult task in its own right.

This paper discusses the fundamental aspects of the outlier detection and also briefs the proposed methodology with the use dimensionality reduction, optimization method for outlier detection for large scale dataset with the use of the entropy. Then, outlier detection is formulated as an optimization problem involving to search for the optimal subset in terms of "goodness" and number of outliers. To solve the optimization problem, derive a new outlier factor function from the weighted entropy and show that computation/updating of the outlier factor can be performed without the need to estimate the joint probability distribution. Here, the PSO includes n number of data samples N that are moving around a D-dimensional search space for optimizing a certain variational property. In addition, it also estimate an upper bound of outliers to reduce the search space. The performance comparison results of the proposed **EMPWC** is measured in terms of the Detection Rate (DR), False Alarm Rate (FAR), time comparison among the number of attributes, number of data objects, Normalized Mean Square Error (NMSE) for error results comparison, Area Under the Curve (AUC). It shows that the proposed **EMPWC** have less NMSE error, FAR, and more Detection Rate (DR) with less time taken to complete the process.

2. RELATED WORK

Outliers can be detected in various fashions such as graphical, statistical, unsupervised, supervised, and semi-supervised methods.

In Wu et al.[13]propose an optimization model of outlier detection with a formal definition of outliers, via a concept of holoentropythat takes the both entropy and total correlation into consideration. Based on this model, define a function for the outlier factor of an object which is solely determined by the object itself and can be updated efficiently. This work also proposes two practical 1-parameter outlier detection methods, named ITB ITB-SS (Information-Theory-Based Step-by-Step) and ITB-SP (Single-Pass) methods, which require no user-defined parameters for deciding whether an object is an outlier. Users need only provide the number of outliers they want to detect. The results of the proposed work show that ITB-SS and ITB-SP are more effective and efficient than mainstream methods and can be used to deal with both large and high-dimensional data sets where existing algorithms fail.

Koufakou et al.[14] propose method, which takes into consideration the sparseness of the dataset, and is experimentally shown as highly scalable with the number of points and the number of attributes in the dataset. The proposed outlier detection for the mixed attribute datasets (ODMAD), at first it identifies outliers based on the categorical attributes, after that focuses on subsets of data in the continuous space by utilizing information about these subsets from the categorical attribute space. The results of this work show that the proposed outlier detection method compares very favorably with other state-of-the art, outlier detection strategies propose in the literature and that the speedup is achieved by its distributed version is very close to linear.

Zhang et al.[15] propose a novel Pattern based Outlier Detection approach (POD) for mixed attribute datasets. Pattern in this work is defined to describe majority of data as well as capture interactions among different types of attributes. In POD, the more does an object deviate from these patterns, the higher is its

outlier factor. This proposed work use slogistic regression to learn patterns and then formulate the outlier factor in mixed attribute datasets. A series of experimental results illustrate that POD performs statistically and significantly better than several classic outlier detection methods using in the mixed attribute datasets.

Zhang et al.[16] studied the problem of projected outlier detection in high dimensional data streams and proposed a new Stream Projected Ouliter Detector (SPOT) technique, to identify outliers embedded in subspaces. A set of subspaces obtained by using Sparse Subspace Template (SST), which is constructed in SPOT to detect projected outliers effectively. Multi-Objective Genetic Algorithm (MOGA) is employed as an effective search method for finding outlying subspaces from training data to construct SST. SST is able to carry out online self-evolution in the detection stage to cope with dynamics of data streams. The results of the proposed method demonstrate the efficiency and effectiveness of SPOT in detecting outliers in high-dimensional data streams.

Pham&Pagh.[17] propose a novel random projection-based technique for large high-dimensional data sets. The proposed method is able to estimate the angle-based outlier factor for all data points in time near-linear in the size of the data. Also, the proposed approach is suitable to perform in parallel environment, to achieve a parallel speedup. Here, this work introduces a theoretical analysis of the quality of approximation to guarantee the reliability of estimation algorithm. The empirical experimentation results on synthetic and real world data sets demonstrate that approach is efficient and scalable to very large high-dimensional data sets.

RELOADED approach [18] trains the classifiers and computes covariance matrices incrementally. Therefore, the decision whether a given point is an anomaly or not it is based only on the previously processed data points. For example, in the RELOADED algorithm [18] for each point in the data set, and for each categorical attribute d of that data point, an appropriate classifier is trained. That classifier, in turn, is used to predict the appropriate value of d. If the prediction is wrong, the count of incorrect predictions is incremented. Next, continuous attributes of the data point are used to incrementally compute the covariance matrix corresponding to the attribute-value pair d. The cumulative violation score of the data point is incremented.

Rousseeuw and Hubert.[19] developed an outlier detection scheme using robust location and scatter estimators for outlier detection in multivariate data. The location refers to the coordinate-wise mean and the scatter refers to the covariance matrix. Statistical measure is computed in three phases namely c-step data iteration, data partitioning, and data nesting [19]. Hido et al.[20] propose a statistical based outlier detection method using the direct density ratio estimation. The major drawback of this statistical method is that most of the statistical tests cannot be applied for the multi-attribute problems. Also, they require the prior knowledge of probability distribution of the data and it is difficult to estimate the real distribution of high dimensional data [20] .

Sugiyama and Borgwardt.[21] developed an unsupervised outlier detection method using sampling-based on the literature and reported that the sampling method outperforms the other method that uses the searching technique using k-Nearest Neighbor (kNN) principle. Present an empirical comparison of various approaches to distance-based outlier detection across a large number of datasets. Report the surprising observation that a simple, sampling-based scheme outperforms state-of-the-art techniques in terms of both efficiency and effectiveness.

Koupaie et al.[22] suggest unsupervised outlier detection to detect the stream data. The Multi-Objective Genetic Algorithm (MOGA) is used to search the outliers from an object space and the *k*-means clustering is used to develop the model in order to detect the outliers. Aim of this study is to present an algorithm to detect outlier in stream data by clustering method that concentrate to find real outlier in period of time. It is considered some outlier that receives in previous time and find out real outlier in stream data. The accuracy of this method is more than other methods. The prime advantage of the unsupervised outlier detection does not require the labelled data since the labelled data are costlier than unlabelled data and it requires special mechanism to label the data. Therefore, this approach is simple and cost-effective than the supervised approach.

3. PROPOSED METHODOLOGY

Outlier detection methods for categorical data can be characterized by the way outlier candidates are measured with respect to other objects in the dataset. In general, outlier candidates can be assessed based either on represents the value of the attribute that belongs to either categorical and discrete value represented by $(y_{1,j}, y_{2,j}, \ldots y_{n,j})(1 < j < m)$and n_j indicates the number of distinct values in attribute y_j. In order to measure the attribute value importance by using the Shannon, Jensen-Shannon Divergence(JSD) and the holoentropy of the attribute is represented as H_x (), mutual information I_x (), and total correlation C_x ()computed on the set X; *e.g.*, I_x (y_i, y_j) represents the mutual information between attributes y_i and y_j. The holoentropy $H_X(Y)$ is written as follows:

$$\begin{aligned} H_x(y) &= H_x(y_1, y_2, \ldots y_m) = \sum_{i=1}^{m} H_x(y_i \mid y_{i-1}, \ldots . y_1) \\ &= H_x(y_1) + \cdots . + H_x(y_m \mid y_{m-1}, \ldots y_1) \end{aligned} \quad (1)$$

$$H_x(y_m \mid y_{m-1}), \ldots . y_1) = -\sum_{y_m, \ldots . y_1} \frac{p(y_m, y_{m-1}, \ldots y_1)}{\log p(y_m \mid y_{m-1}, \ldots y_1)} \quad (2)$$

The total correlation [23] is defined as the sum of mutual information of multivariate discrete random vectors Y, denoted as $C_x(Y)$

$$\begin{aligned} C_x(y) &= \sum_{i=2}^{m} \sum_{\{r_{-1} \ldots . r_i\} \subset \{1, \ldots m\}}^{m} I_x(y_{r_1}, \ldots . y_{r_i}) \\ &= \sum_{\{r_1 \ldots . r_i\} \subset \{1, \ldots m\}} I_x(y_{r1}, ; y_{r_2}) + \ldots + I_x(y_{r_1}, \ldots y_{r_m}) \end{aligned} \quad (3)$$

Where $r_1 \ldots r_i$ are attribute numbers chosen from 1 to m. $I_x(y_{r_1}, \ldots . y_{r_i}) = I_x(y_{r_1}, \ldots . y_{r\ i-1}) - I_x(y_{r_1}, \ldots y_{r_1})$ is the multivariate mutual information of $yr_1 \ldots y_{r_i}$, where $I_x(y_{r_1}, \ldots y_{r_{i-1}}) \mid y_i) = E(I(y_{r_1}), \ldots . y_{r_{i-1}}) \mid y_{r_i})$ is the conditional mutual information. The holoentropy $HL_x(Y)$ data distribution or attribute correlation, which provides a more global measure. They are also assessed using a between object similarity or local density, which provides a local measure. The goal of this work is two fold. First, deals with the lack of a formal definition of outliers and modeling of the outlier detection problem; second, aims to propose effective and efficient methods that can be used to solve the outlier detection problem in real applications. In this section, first look at how entropy, Shannon, Jensen-Shannon Divergence (JSD) and total correlation is used to capture the likelihood of outlier candidates. The concept of holoentropy is proposed and formulate the outlier detection problem.

4. MEASUREMENT FOR OUTLIER DETECTION

Consider data be the X containing number of the data objects as $n(x_1, \ldots . x_n)$ each x_i for $1 < i < n$ being a vector of categorical attributes $[y_1, y_2, \ldots, y_m]^T$, where m is the number of categorical and discrete data attributes, y_j is defined as the sum of the entropy and the total correlation of the random vector Y, and can be expressed by the sum of the entropies on all attributes holoentropy assigns equal importance to all the attributes, whereas in real applications. Solve this issue proposed weighting method computes the weights directly from the data and is motivated by increased effectiveness in practical applications rather than by theoretical necessity.

$$HL_x(Y) = H_x(Y) + C_x(Y) = \sum_{i=1}^{m} H_x(y_i) \quad (4)$$

$$W_x(y_i) = 2\left(1-\frac{1}{1+\exp(-H_x(y_i))}\right) \quad (5)$$

Even though in the holoentropy function, it sets a minimum value for each attributes and the maximum expected number of attributes value are identified in the Shannon and Jensen-Shannon Divergence (JSD).

Shannon entropy : Shannon entropy is one of the most important metrics in information theory. Entropy measures the uncertainty associated with a random variable, the expected value of the information in the message (in classical informatics it is measured in bits).

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i) = \sum_{i=1}^{n} p(x_i) \log \frac{1}{p(x_i)} \quad (6)$$

Jensen-Shannon Divergence (JSD) is the mean relative entropy between two distributions and the distribution mean [24].

$$JS(y_i|y_j) = \frac{1}{2}\sum_i P(y_i)\ln \frac{P(y_i)}{\frac{1}{2}(P(y_i)+P(y_j))} + \frac{1}{2}\sum_i P(y_i)\ln \frac{P(y_j)}{\frac{1}{2}(P(y_i)+P(y_j))} \quad (7)$$

$$= \frac{1}{2}D(y_i \,||\, M) + \frac{1}{2}D(y_j \,||\, M) = S(M) - \frac{1}{2}S(y_i) - \frac{1}{2}S(y_j) \quad (8)$$

The equation (9) shows probability calculation formula of each firefly for given set of data.

$$p(y_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{y_i - y_m}{h_n}\right) \quad (9)$$

where $\varphi(x)$ defines the window function and n is the total number of data objects, V_n and h_n be the volume and edge length of a hypercube. Once the JSD is calculated then computes the weights directly from the data and it is motivated by increased effectiveness in practical applications rather than by theoretical necessity

$$w_x(y_i) = 2\left(1-\frac{1}{1+\exp(JS(y_i|y_j))}\right) \quad (10)$$

The weighted holoentropy of random vector $W_x(Y)$ is defined as the sum of the weighted entropy on each attribute of the random vector Y.

$$W_x(y) = \sum_{i=1}^{m} W_x(y_i) H_x(y_i) \quad (11)$$

Given a data set X with *n* objects and the number o, a subset Out(o) is defined as the set of outliers if it minimizes J_X (Y; o), defined as the weighted holoentropy of X with o objects removed

$$J_x(Y, O) = W_{x\setminus set(o)}(Y) \quad (12)$$

where set (O) is any subset of o objects from X. In other words

$$Out(O) = \text{argmin} J_X(Y, O) \quad (13)$$

Hence, outlier detection is now formulated and stated as an optimization problem. For a given o, the number of possible candidate sets for the objective function is $C_n^O = \frac{n!}{O!(n-O)!}$, which is very high. Moreover, one might have to determine the optimal value of O, that is how many outliers a data set really has. A possible theoretical approach to this problem is to search for a range of values of O and decide on an optimal value of O by optimizing a certain variational property of J_X (Y,O). Consider this as a proposed direction in this research work direction. For now, it focus is on developing practical solutions to the optimization problem.

Particle Swarm Optimization (PSO) : PSO is typically a kind of population independent optimization tool that was initially presented in the form of an optimization scheme for the purpose of real-number spaces. In case of PSO, each particle holds analogy to an individual "fish" existing in a school of fish. Here, in order to choose most optimizing a certain variational property of J_X (Y, O) analysis and optimization for range of values for O. PSO includes n number of data samples N that are moving around a D-dimensional search space for optimizing a certain variational property of J_X (Y, O). The procedure of PSO starts with a population consisting of number of the data objects as $n(x_1,\ldots. x_n)$ each x_i with $r_1 \ldots r_i$ are attribute numbers chosen from 1 to *m* for every data samples and the optimization scheme subsequently looks for the best range of values for O by means of updating the generations continuously. Every data objects (particles) utilize its individual data object which have at the same cluster among data points. The knowledge is attained by the swarm in a complete form to discover optimizing a certain variational property of J_X (Y,O) in a cluster. The location of the i^{th} data samples of cluster particle can be indicated by $l = (l_1,\ldots l_j)$. The velocity corresponding to the ith cluster of data points can be indicated as $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$. The velocities of the datapoints in the cluster are limited inside $[V_{min}, V_{max}]^D$, correspondingly. The best earlier visited location of the i^{th} data points represented its individual best outlier detection results $lbest = (lb_{i1}, lb_{i2},\ldots, lb_{iD})$, a value referred to as $lbest_i$. The best value of the entire individual $lbest_i$ values are indicated the global best position $g$best = $(gb_1, gb_2, \ldots, gb_D)$ and known as $g$best.

At every generation, the position and velocity of the ithevery datapoints in the cluster is revised by $lbest_i$ and gbest in the swarm. It takes place in the space that datapoints discrete problem, with the intention of solving this problem Kennedy and Eberhart proposed PSO (PSO) is employed to discrete binary variables. In case of a binary space, a particle is datapoints in the cluster possibly will make a move to the close corners of a hypercube by means of flipping several numbers of bits. As a result, the particle velocity on an overall might be described through the number of bits transformed per number of processes. In case of PSO, every datapoints for outliers removal are revised in accordance with the equations that follow:

$$v_{id}^{new} = w \times v_{id}^{new} + c_1 \times r_1 \times (\text{lbest}_{id} - O_{id}^{old}) + c_2 \times r_2 \times (g\text{best}_{id} - O_{id}^{old}) \quad (14)$$

$$\text{If} \quad v_{id}^{new} \notin (V_{min}, V_{max}) \text{ then } v_{id}^{new} = \max(\min(V_{mnax}), v_{id}^{new}), V_{min})$$

$$S(v_{id}^{new}) = \frac{1}{1 + e^{-v_{id}^{new}}} \quad (15)$$

$$\text{If} \quad (r_3 < S(v_{id}^{new})) \text{ then } O_{id}^{new} = 1 \text{ else } O_{id}^{new} = 0 \quad (16)$$

Here, w indicates the inertia weight for optimizing a certain variational property of J_X (Y,O) its current dataset samples, r_1, r_2 and r_3 correspond to random numbers between (0,1), and c_1 and c_2 refer to acceleration constants. Velocities v_{id}^{new} new and v_{id}^{old} indicate the new and old velocities for outliers. v_{id}^{old} stands for the current particle position, and v_{id}^{new} stands for the new, updated outlier detection position. In Equation (16) outlier detection position velocities of every dataset samples are attempted to a maximum velocity V_{max}. When the summation of accelerations makes the velocity of that particular dimension to go beyond V_{max}, subsequently the velocity of that specific dimension is restricted to V_{max}. V_{max} and V_{min} specify the constraints. When $S(v_{id}^{new})$ is more than r_3, subsequently its position value of current data point is denoted by {1} elsewhere {0}. Once the outlier values are optimized it becomes very easy to perform clustering process. Finally, clustering is performed using Expectation Maximization (EM) for PSO based Weighted Clustering (EMPWC) algorithm. Preliminary experiment indicates that the performance of exact and approximate outlier factor are very similar. To avoiding the high time complexity of exact factor computation, use the approximate factor J_X (Y,O) to represent the approximate one in this work.

Expectation Maximization (EM) for PSO based Weighted Clustering (EMPWC): Finding significant groups in a set of data points is a central problem in many fields. Consequently, clustering receives a

lot of attention, and many methods, algorithms and software packages are available today. Among these techniques, parametric finite mixture models play a paramount role, due to their interesting mathematical properties as well as to the existence of maximum likelihood estimators based on Expectation-Maximization (EM) algorithms. While the finite Gaussian Mixture (GMM) [25] is the model of choice, it is extremely sensitive to the presence of outliers. Alternative robust models have been proposed in the statistical literature, such as mixtures of skew *t*-distributions [26] and their numerous variants, *e.g*. [27-28].

This work proposes an Expectation Maximization (EM) for PSO based Weighted Clustering (EMPWC) in which variable $W_X(y)$ is used as a weight to account for the reliability of the observed dataset samples X_i and this independently on its assigned cluster. The distribution of $W_X(y)$ is not a gamma mixture anymore but has to depend on i to allow each data point to be potentially treated differently. In this work, introduce the weighted data with Gaussian Mixture Model as two cases,

(*i*) the weights $W_X(y)$ are determined by using factors like holoentropy and JSD, they are fixed, and (*ii*) the weights are modeled as variables and hence they are iteratively updated if the sample changes. Then based on these weights as optimal value of O by optimizing a certain variational property of J_X (Y,O) using PSO and modelled as random variables. Model these variables with gamma distributions and derive a closed-form EM algorithm which will be referred to as the EMPWC. Then M-step and E-step are updated continuously.

This work also proposes to apply the proposed weighted-data robustclustering method to the problem of data clustering and outlier detection. This problem arises when the task is, *e.g*. to detect a sparse datapoints. In this section, presents the formal definition of the proposed EMPWC algorithm. Let $x \in \mathbb{R}^d$ be a random data sample vector following a multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^d$ namely $p(x \mid \theta) = N(x ; \mu, \Sigma)$ with the notation $\theta = \{\mu, \Sigma\}$. Let $W_X(y) > 0$ be a weight indicating the relevance of the attribute value measurement results from dataset samples *x*. Intuitively, higher the weight *w*, stronger the impact of *x* on cluster. In terms of the likelihood function, this is equivalent to raise $p(x ; \theta)$ to the power $W_X(y)$, that is $N(x ; \mu, \Sigma)^{WX(y)}$. It is straightforward to notice that $N(x ; \mu, \Sigma)^{WX(y)} \alpha N(x ; \mu, \Sigma/W_X(y)$. Therefore, $W_X(y)$ plays major role to increase the clustering results and is different for each dataset samples *x*. Subsequently, write:

$$\hat{p}\,(x, \theta, W_X(y)) \;=\; N\left(x ; \mu, \frac{1}{W_X(y)}\Sigma\right) \tag{17}$$

from which derive a mixture model with K components:

$$\hat{p}\,(x, \theta, W_X(y)) \;=\; \sum_{k=1}^{K} \pi_k N\left(x ; \mu_k, \frac{1}{W_X(y)}\Sigma_k\right) \tag{18}$$

where $\Theta = \{\pi_1, \ldots, \pi_k, \theta_1, \ldots \theta_k\}$ are the mixture parameters $\pi_1, \ldots, \pi_k$ are the mixture coefficients satisfying $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, $\theta_k = \{\mu_k, \Sigma_k\}$ are the parameters of the *k*-th component and K is the number of components. It can be refered to the model in (18) as the weighted function derived from equation (11). Let $X = \{x_1, \ldots. x_n\}$ be the observed data and $W_X(y) = \{W_1(y), \ldots., W_n(y)\}$ be the weights associated with X. Assume that the *xi* is independently drawn from (2) with $W = W_i$. The observed-data log-likelihood is:

$$\ln\ \hat{p}\,(x, \pi, W_X(y)) \;=\; \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{n} \pi_k N\left(x_i ; u_k, \frac{1}{W_X(y)}\Sigma_k\right)\right) \tag{19}$$

It is well known that direct maximization of the log-likelihood function is problematic in case of mixtures and that the expected complete-data log-likelihood must be considered instead. Hence, introduce a set of n hidden (assignment) variables $Z = \{z_1, \ldots z_n\}$ associated with the observed variables X and such that $z_i = k$, $k \in \{1, \ldots K\}$ if and only if xi is generated by the k^{th} component of the mixture. In the following we first consider a fixed (given) number of mixture components K, and then extend the model to an unknown K, thus estimating the number of the components from the data. Then the complete expected data log-likelihood is:

$$Q_c(\Theta, \Theta^{(r)}) = E_{P(Z \mid X;\, W_X(y),\, \Theta^{(r)})}[\ln P(X, Z, W_X(y), \Theta)] \quad (20)$$

where $E_p[\cdot]$ denotes the expectation with respect to the distribution P. Distribution P. The $(r + 1)$-th EM iteration consists of two steps namely, the evaluation of the posterior distribution is given the current model parameters $\Theta^{(r)}$, the weights W, and the maximization of (20) with respect to Θ(M-step):

$$\Theta^{(r+1)} = \arg\max_{\Theta} Q_c(\Theta, \Theta^{(r)}) \quad (21)$$

E-Step

The posteriors $\eta_{ik}^{(r+1)} = p(z_i = k \mid x_i ; W_X(y), \Theta^{(r)}$ are updated with

$$\eta_{ik}^{(r+1)} = \frac{\pi_k^{(r)} \hat{p}(x, \theta, W_X(y))}{\hat{p}(x, \Theta^{(r)}, W_X(y))} \quad (22)$$

M-Step

$$Q_c(\Theta, \Theta^{(r)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \eta_{ik}^{(r+1)} \ln \pi_k N\left(x ; \mu_k, \frac{1}{W_X(y)}\Sigma_k\right) \quad (23)$$

$$\stackrel{\Theta}{\cong} \sum_{i=1}^{n}\sum_{k=1}^{K} \eta_{ik}^{(r+1)} (\ln \pi_k - \ln |\Sigma_k|^{\frac{1}{2}} \frac{W_X(y)}{2} (x_i - (\mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)) \quad (24)$$

$$\pi_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} \eta_{il}^{(r+1)} \quad (25)$$

$$\mu_{ik}^{(r+1)} = \frac{\sum_{i=1}^{n} W_X(y) \eta_{ik}^{(r+1)} x_i}{\sum_{i=1}^{n} W_X(y) \eta_{ik}^{(r+1)}} \quad (26)$$

$$\Sigma_{ik}^{(r+1)} = \frac{\sum_{i=1}^{n} W_X(y) \eta_{ik}^{(r+1)} (x_i - \mu_{ik}^{(r+1)})(x_i - \mu_{ik}^{(r+1)})^T}{\sum_{i=1}^{n} \eta_{ik}^{(r+1)}} \quad (27)$$

This process is repeated until the lower bound is positive. The upper bound on outliers (UO), the anomaly candidate set (AS), and the normal object set (NS). Thus, the data objects with positive lower bound is considered as AS = $\{x_i \mid \hat{D}_{yj \| yj} > 0\}$, the data objects with nonpositive set is considered as,

$$UO = N(AS) = \sum_{i=1}(\hat{D}_{yi \| yj} > 0) \ \& \ \text{argmax} J_X(Y, O) \quad (28)$$

Algorithm 1: Outlier detection

Input : Dataset X and number of the outlier requested o

Output : Outlier results OS

1. Compute $w_x(y_i)$ for $(1 \le i \le m)$ by (11)
2. Initially set OS = 0
3. for i = 1 to n do
4. Compute Out(O) from (12) and obtain AS by (28)
5. End for
6. If O > UO then
7. O = UO
8. Else
9. Build OS by searching for the o Objects with greatest OF(x_i) in AS
10. End if

5. EXPERIMENTAL RESULTS

In this section, conduct effectiveness and efficiency tests to analyze the performance of the proposed EMPWC method. To test effectiveness, compare the result to the existing methods Information-Theory-Based Step-by-Step (ITB-SS) and Information-Theory-Based Single-Pass (ITB-SP) for synthetic data sets. For the efficiency test, conduct evaluations on synthetic data sets to show how running time increases with the number of objects, the number of attributes and the number of outliers. A large number of public real data sets, most of them from UCI [29], are used in this experiments, representing a wide range of domains in science and the humanities. The data set used is the public, categorical "soybean data" [29], with 47 objects and 35 attributes. This data contains a very small class of 10 objects. Since the data does not have explicitly identified outliers, it is natural to treat the objects of the smallest class as "outliers". The Area Under the Curve (AUC) [30] and significance test are used to measure the performance. The AUC results of different methods and the characteristics of all test data sets, such as the numbers of objects (#*n*), attributes (#*m*) and outliers (#*o*), and the upper bound on outliers (#UO), are summarized in the upper part of Table 2. The results reported in Table 2 warrant a number of comments. These results are evidence of the importance of capturing attribute weights and it is also compared with the existing methods ITB-SS, ITB-SP without weighting and with weighting. Frequent Pattern Outlier Factor (FIB), Common-neighbor-based distance (CNB).

Table 1

AUC Results of Tested Algorithms on the Real dataset

Dataset	*#n*	*#m*	*#o*	*#UO*	*CNB*	*FIB*	*ITB-SP*	*ITB-SS*	*AMCEM*	*EMPWC*
Breast-*c*	495	11	45	125	0.99	0.90	0.991	0.993	0.996	0.997
Credit-*a*	413	17	30	171	0.84	0.92	0.985	0.992	0.995	0.996
Diabetes	768	9	268	340	0.86	0.88	0.75	0.912	0.945	0.945
Ecoli	336	8	9	144	0.89	0.92	0.96	0.99	0.996	0.998

To measure the time consumption with increasing numbers of objects, attributes and outliers. As Figure. 1 indicates, the run times of EMPWC, AMCEM, ITB-SP, ITB-SS, and FIB are almost linear functions of the number of objects. The proposed EMPWC has lower rate than other existing system. From the theoretical analysis, time complexity of CNB [31] increases quadratically with the number of objects, which is confirmed by the experimental data of Figure 1.

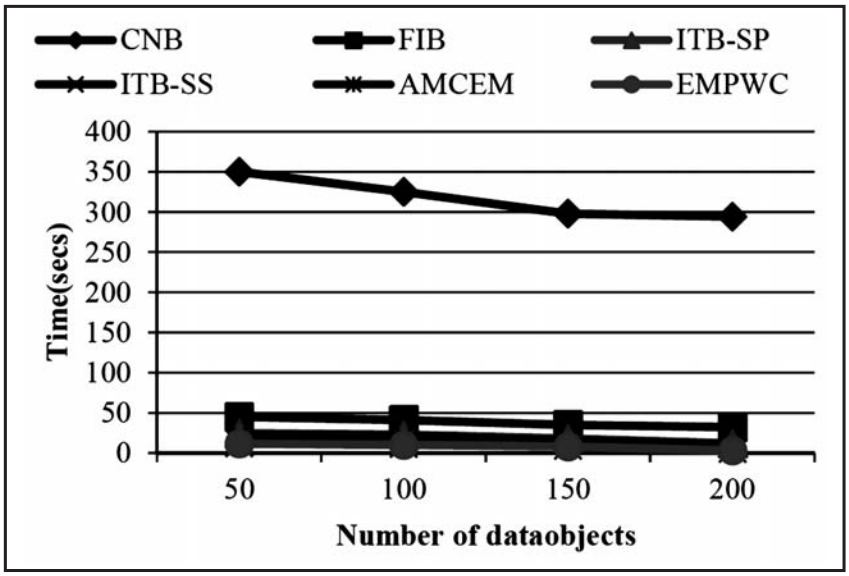

Figure 1: Results of efficiency real data sets for data objects *vs* methods

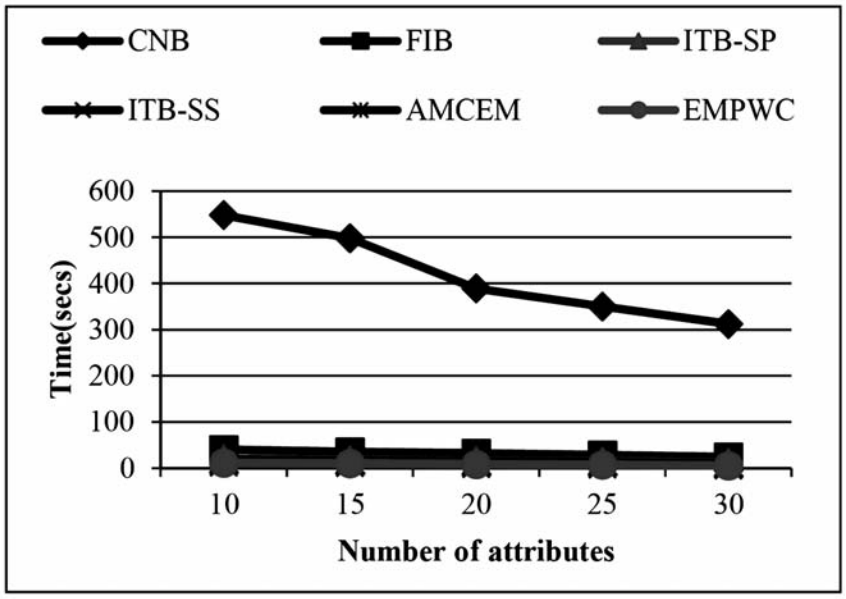

Figure 2: Results of efficiency real data sets for data attributes *vs* methods

For the attributes increasing test, Figure 2 shows that the run times of the EMPWC, increase rapidly with the number of attributes, which closely matches the theory that the time complexities of FIB [32] increase quadratically with the number of attributes. Compared with the time increase of FIB, CNB, ITB-SS, ITB-SP, AMCEM the increases for the other methods are too small noticeable in Figure 2.

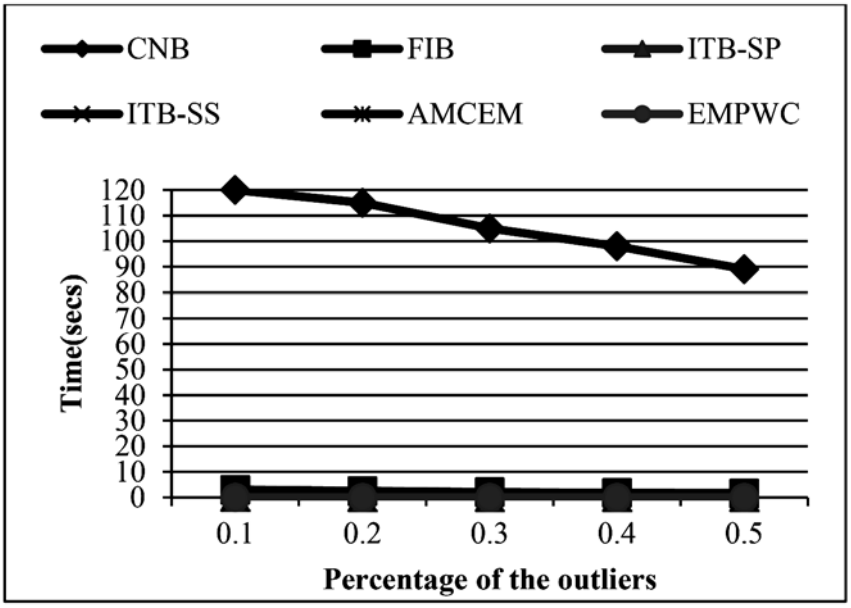

Figure 3: Results of efficiency real data sets for percentage of the outliers vs methods

Figure 3 illustrates the run time as a function of the percentage of "outliers" in the data set each method is asked to search for. The time axis is in the log (10) scale. The run times of CNB and FIB remain almost fixed with the "outlier percentage." Those of ITB-SP and ITB-SS methods increase linearly, and the proposed EMPWC increases highly but remain much lower than those of other methods even for very high "outlier percentages."

The Normalized Root Mean Square Error (NRMSE) is defined as,

$$\text{NRMSE} = \frac{\sqrt{\text{Mean}[(y_{\text{guess}} - y_{\text{ans}})^2]}}{\text{std}[y_{\text{ans}}]} \tag{29}$$

where y_{guess} and y_{ans} are vectors whose elements are the estimated values and the known answer values respectively, for all the data objects in the cluster s. The mean and the standard deviation are calculated over outlier data in the entire matrix.

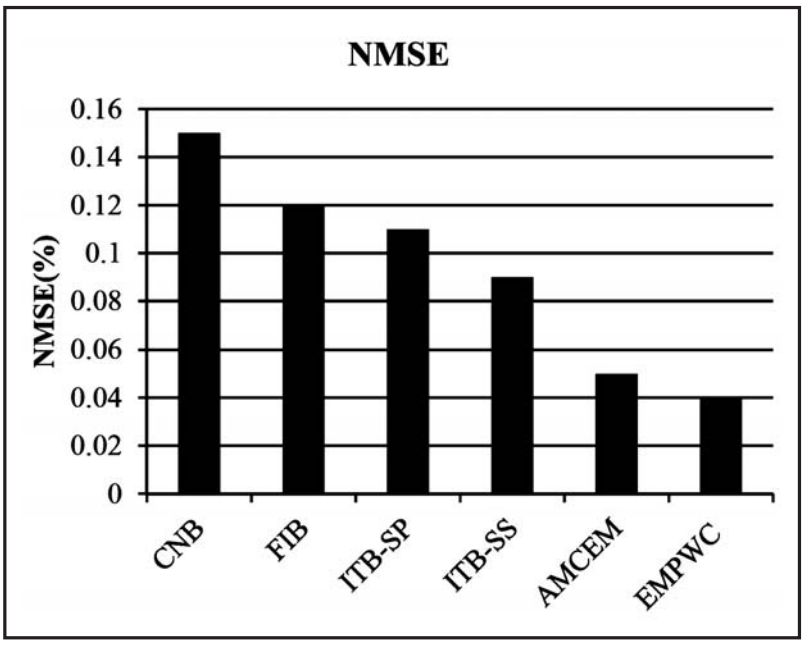

Figure 4: NMSE for real datasets *vs* methods

In Figure 4, shows the performance comparison results of the NMSE for the existing methods such as CNB, FIB, ITB-SP , ITB-SS and proposed EMPWC algorithm,the NMSE value of the proposed EMPWC algorithm have less NMSE when comparing to the existing methods .

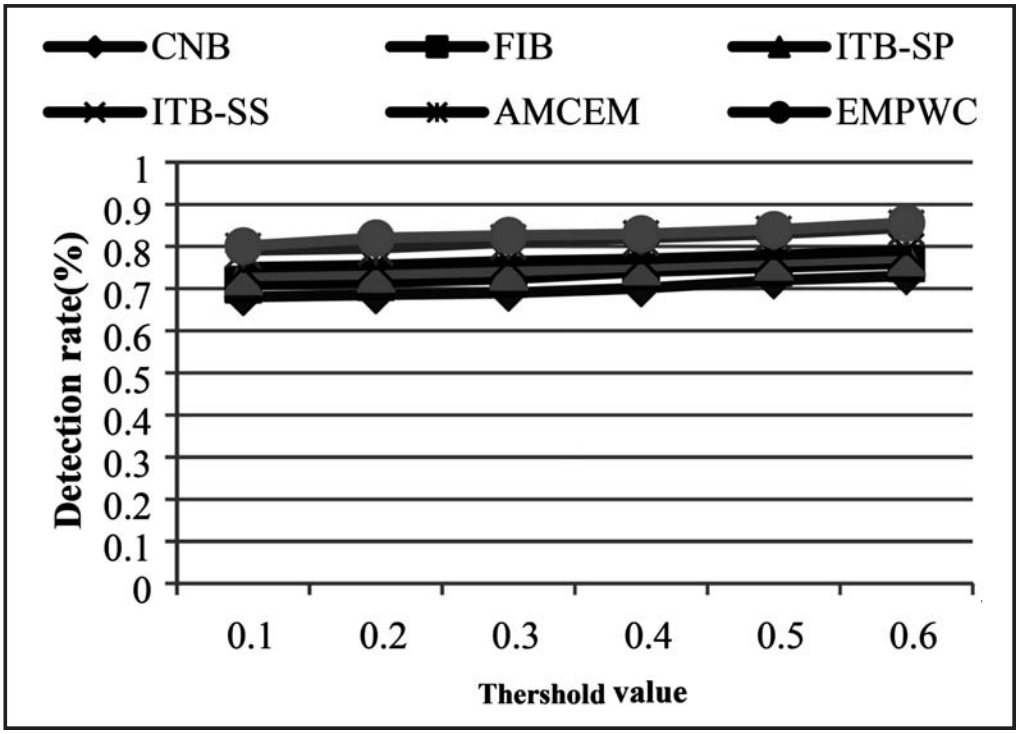

Figure 5: Detection rate for real data sets *vs*. methods

Correct detection rate, which is the number of outliers correctly identified by each approach as outliers:

$$\text{CDR} = \frac{\text{No.of outliers correctly detected as outlier}}{\text{Total no.of outlier in dataset}} \tag{30}$$

False alarm rate, reflecting the number of normal points erroneously identified as outliers

$$FA = \frac{\text{No. of outliers incorrectly detected as outlier}}{\text{Total no. of outlier in dataset}} \quad (31)$$

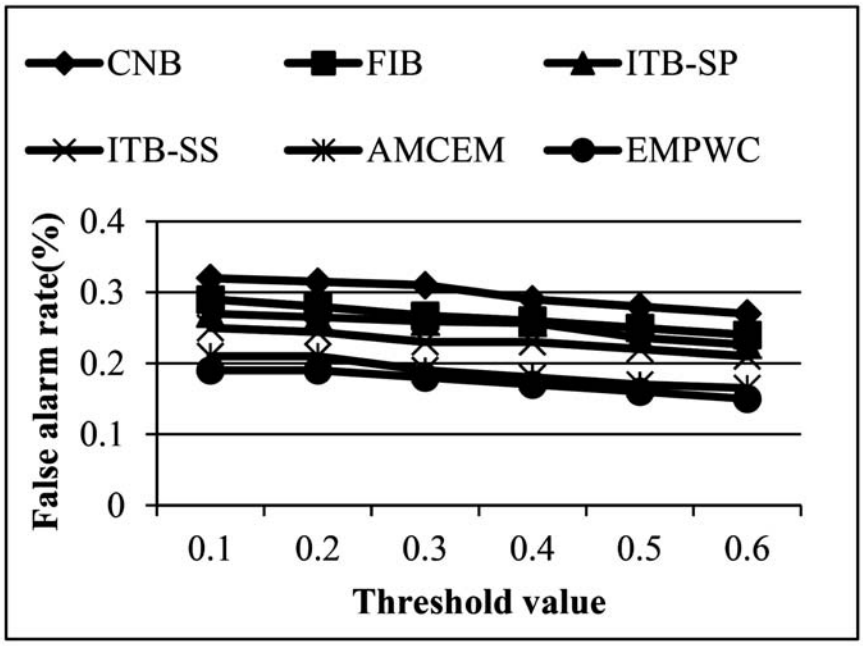

Figure 6: False alarm rate for real datasets *vs* methods

In Figure 5, shows the performance comparison results of the outlier Detection Rate (DR) for the existing methods such as CNB, FIB, ITB-SP, ITB-SS and proposed EMPWC algorithm. Detection Rate (DR) value of the proposed EMPWC algorithm have more DR when comparing to the existing methods.

In Figure 6, shows the performance comparison results of the False Alarm Rate(FAR) for the existing methods such as CNB, FIB, ITB-SP, ITB-SS and proposed EMPWC algorithm. False Alarm Rate (FAR) value of the proposed SAVF algorithm have less FAR when compare to the existing methods .

6. CONCLUSION AND FUTURE WORK

The effectiveness of proposed EMPWC outlier detection method requires attribute frequency based results from a new concept of weighted entropy optimization that considers both the data Shannon and Jensen-Shannon Divergence (JSD) to measure the likelihood of outlier candidates, while the efficiency of proposed algorithms results from the outlier factor function derived from the entropy. The outlier factor of an object is solely determined by the object and its updating does not require estimating the data distribution. In this work, present a weighted-data clustering for outlier detection. While the first algorithm appears as a straightforward generalization of standard EM for Gaussian Mixtures. The second one is performed based on weight computation results. In this model, weight values are derived from entropy measures. In addition this work , range values of outilers (o) are optimized using particle swarm optimization (PSO). Also estimate an upper bound for the number of outliers and an anomaly candidate set. This bound, obtained under a very reasonable hypothesis on the number of possible outliers, allows to further reduce the search cost. Based on this PSO method, the data clustering results are increased and hence the algorithm is extremely efficient. In particular, the proposed EMPWC algorithm can deal with large number of data sets most efficiently than the existing methods. In future work, evaluation can be done to small real data set and a bundle of synthetic data sets show that the proposed algorithms do tend to optimize the selection of candidates as outliers. Moreover, the experiments on real and synthetic data sets in comparison with other algorithms confirm the effectiveness and efficiency of the proposed algorithms in practice.

7. REFERENCES

1. Chandola V., Banerjee A., and Kumar V. 2012. Anomaly Detection for Discrete Sequences: A Survey, IEEE Trans. Knowledge and Data Eng., Vol. 24, No. 5: 823-839.
2. Takeuchi J., and Yamanishi K. 2006. A Unifying Framework for Detecting Outliers and Change Points from Time Series, IEEE Trans. Knowledge and Data Eng., Vol. 18, No. 4: 482-492, Apr.
3. Leckie T., and Yasinsac A. 2004. Metadata for Anomaly-Based Security Protocol Attack Deduction, IEEE Trans. Knowledge and Data Eng., Vol. 16, No. 9: 1157-1168.
4. Aleskerov E., Freisleben B., and RaoCardwatch B. 1997. A Neural Network Based Database Mining System for Credit Card Fraud Detection, Proc. IEEE/IAFE Computational Intelligence for Financial Eng. Conf., (CIFEr '97).
5. Hodge V.J., and Austin J. 2004. A Survey of Outlier Detection Methodologies, Artificial Intelligence Rev., Vol. 22, No. 2: 85-126.
6. Aggarwal C.C., and Yu S.P. 2005. An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, , Vol. 14: 211–221.
7. Barbara D., Domeniconi C., and Rogers J.P. 2006. Detecting Outliers Using Transduction and Statistical Testing, Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '06).
8. Filippone M., and Sanguinetti G. 2010. Information Theoretic Novelty Detection, Pattern Recognition, Vol. 43: 805-814.
9. Zhang D., Gatica-Perez D., Bengio S., and McCowan I. 2005. Semi Supervised Adapted HMMs for Unusual Event Detection, Proc. IEEE CS Conf. Computer Vision and Pattern Recognition (CVPR '05).
10. Gao J., Cheng H., and Tan P.N. 2006. Semi-Supervised Outlier Detection, Proc. ACM Symp. Applied Computing (SAC '06).
11. Zhang K., Hutter M., and Jin H. 2009. A new local distance-based outlier detection approach for scattered real-world data. In Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, 813-822.
12. Chandola V., Banerjee A., and Kumar V. 2009. Anomaly Detection: A Survey. ACM Computing Surveys, Vol. 41, No. 3: 15:1-15:58.
13. Wu S., and Wang S. 2013. Information-theoretic outlier detection for large-scale categorical data. Knowledge and Data Engineering, IEEE Transactions on, Vol. 25, No. 3: 589-602.
14. Koufakou A., and Georgiopoulos M. 2010. A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Mining and Knowledge Discovery, Vol. 20, No. 2: 259-289.
15. Zhang K., and Jin H. 2010. An effective pattern based outlier detection approach for mixed attribute data. In AI 2010: Advances in Artificial Intelligence, Springer Berlin Heidelberg, 122-131.
16. Zhang J., Gao Q., Wang H., Liu Q., and Xu K. 2009. Detecting projected outliers in high-dimensional data streams. In Database and Expert Systems Applications , Springer Berlin Heidelberg, 629-644.
17. Pham N., and Pagh R. 2012. 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 , ACM, 877-885.
18. Otey M.E., Parthasarathy S., and Ghoting A. 2005. Fast lightweight outlier detection in mixed-attribute data. Technical Report OSU-CISRC- 6/05-TR43. Department of Computer Science and Engineering, The Ohio State University, Ohio, United States.
19. Rousseeuw P.J., and Hubert M. 2011. Robust statistics for outlier detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 1, No. 1: 73-79.
20. Hido S., Tsuboi Y., Kashima H., Sugiyama M., and Kanamori T. 2011. Statistical outlier detection using direct density ratio estimation. Knowledge and Information Systems, Vol. 26, No. 2: 309-336.
21. Sugiyama M., and Borgwardt K. 2013. Rapid distance-based outlier detection via sampling. In Burges C.J.C., Bottou L., Welling M., Ghahramani Z., and Weinberger K.Q. (Eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc, 467–475.
22. Koupaie M.H., Ibrahim S., and Hosseinkhani J. 2014. Outlier detection in stream data by clustering method. International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 3: 25-34.

23. Srinivasa S. 2005. A Review on Multivariate Mutual Information, Univ. of Notre Dame, Notre Dame, Indiana, Vol. 2: 1-6.

24. Lin J. 1991. Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory, 37:145–151.

25. Reynolds D. 2015. Gaussian mixture models. Encyclopedia of Biometrics, 827-832.

26. Lin T.I., Lee J.C., and Hsieh W.J. 2007. Robust mixture modeling using the skew t distribution. Statistics and Computing, Vol. 17, No. 2: 81-92.

27. Forbes F., and Wraith D. 2014. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering, Statistics and Computing, Vol. 24, No. 6: 971– 984.

28. Lee S., and McLachlan G. 2014. Finite mixtures of multivariate skew t distributions: some recent and new results, Statistics and Computing, Vol. 24, No. 2: 181–202.

29. UCI Machine Learning Repository, http://www.ics.uci.edu/ learn/MLRepository.html, 2011.

30. Bolton R., and Hand D. 2002. Statistical fraud detection: A review, Statistical Science. Vol. 17, No. 3: 235-255.

31. Li S., Lee R., and Lang S. 2007. Mining Distance-Based Outliers from Categorical Data, Proc. IEEE Seventh Int'l Conf. Data Mining Workshops (ICDM '07).

32. He Z., Xu X., Huang Z.J., and Deng S. 2005. FP-Outlier: Frequent Pattern Based Outlier Detection," Computer Science and Information Systems, Vol. 2: 103-118.