International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 4 (2015) pp. 8575-8589

http://www.ripublication.com

Ascent-Based Monte Carlo Expectation– Maximization Outlier Detection For Large-Scale Categorical Data

J.Rajeswari[1], Dr. R. Manicka Chezian[2]
[1] Research Scholar, Department of Computer Science Professor,
[2] Associate Professor, Dept of Computer Science (Aided),
[1]Karpagam University,
[1]Coimbatore-641 021, India
[1]rajeswari.scholar@gmail.com

Abstract

Detection of outliers is essential to numerous databases and analytic tasks like fraud detection and customer migration. In this paper formulated a method for investigating outlier detection for categorical data sets. This complication is especially tough owing to the complexity of defining a meaningful similarity measure for categorical data. In order to solve this problem, existing work use a new concept of holoentropy that takes both entropy and total correlation into account. But in the existing methods the entropy methods also have lack of the problem for outlier detection for each attribute and some clustering methods are not used in this work for categorical data. In this paper proposes a novel approach which combines the attributes based Kullback-Leibler divergence (KLD) for attribute weighting process and perform the Ascent-based Monte Carlo expectation–Maximization (AMCEM) methods for outlier detection, in major maximization step KLD based attribute weighting plays major important to detect whether the selected data object is outlier or not. The experimentation analysis of the proposed system is carried out by using real datasets from UCI machine learning repository. The performance comparison results of the proposed AMCEM 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 AMCEM have less NMSE error, FAR, and more Detection Rate (DR) with less time taken to complete the process.

Keywords: Expectation Maximization (EM), Monte Carlo Expectation Maximization (MCEM), Kullback- Leibler Divergence (KLD), Ascent-based Monte Carlo Expectation–Maximization (AMCEM), Clustering, Outlier Detection, Large Scale Data, Attribute Weighting.

Introduction

Outlier detection is an essential step in a variety of practical applications including fraud detection [1], network intrusion [2-3], health system monitoring [4], and criminal activity detection in E-commerce [5], and can also be used in scientific research for data analysis. Data mining techniques that have been developed in earlier work based on both supervised and unsupervised learning to solve outlier detection problem. Unlike supervised learning methods that typically require labeled data to classify rare events [6], unsupervised techniques detect outliers as data points that are extremely different from the majority data based on some pre- specified measure. These methods are typically called outlier/anomaly detection techniques, and their success depends on the choice of similarity measures, feature selection, weighting, and most importantly on an approach used to detect outliers.

Besides, in a supervised approach a training set should be offered with labels for anomalies with labels of standard objects, however training set with normal object labels alone essential for the semi-supervised approach. In order to overcome these problems unsupervised approach does not require any object label information and it is mostly used in earlier work. These unsupervised learning methods in outlier detection have focused on datasets with a specific attribute type, mainly assuming that attributes are only numerical and/or ordinal. In the case of data with categorical attributes, techniques which take numerical data required to initially map the categorical values to numerical values, a task which is not a simple process to a numerical attribute [7]. A second issue is that many applications for mining outliers require the mining of very large datasets [8].

The works carried out these methods either support unsupervised learning for discrete data and the measurement of the attributes consider for entire data without consideration of the attribute value and another one of the main challenges of outlier detection algorithms are data sets with non-homogeneous densities. Clustering and outlier detection are two major data mining tasks. They are extensively employed, for instance, in case of bioinformatics, for the purpose of detecting functionally dependent genes, in case of marketing, for the purpose of customer segmentation, in case of health surveillance, for the purpose of anomaly detection, and so on. Clustering-based outlier detection algorithms cannot properly detect the outliers in case of noisy data and unless the number of clusters is known in advance. The common problem with the existing methods is the lack of a formal definition for the outlier detection problem and doesn't support for categorical data for larger dataset.

In order to solve this problem in the unsupervised learning methods, in this work particularly study Kullback- Leibler divergence (KLD), which captures the distribution of the each and every categorical data attribute weighting to find the similarity of the each and every data object based on this concept, build an Ascent-based Monte Carlo Expectation– Maximization (AMCEM) for outlier detection and propose a criterion for estimating the "goodness" of a subset of objects as potential outlier candidates. Then outlier detection is formulated and number of the outliers in the cluster for particular data object is identified in AMCEM step. Experimentation results shows that the proposed AMCEM based system have high outlier detection

results than the informatics theoretic approaches, high complexity of exploring the whole outlier candidate space.

Background Study

Statistical-model based methods assume that a specific model describes the distribution of the data. Common drawbacks include obtaining the appropriate model for each specific dataset and application, and short of scalability with regard to data dimensionality [7]. Distance-based techniques [9] fundamentally work out distances between data points, accordingly become quickly unfeasible for large datasets. Knorr et al [9] define a point as an outlier if at least p% of the points in the dataset lie further than distance D from it. These techniques demonstrate high computational complexity rendering them unfeasible for really large datasets. Numerous techniques might be employed to make the k-NN queries quicker, like an indexing structure, for instance, KD-tree, X-tree, on the other hand, these structures have been shown to collapse as the dimensionality grows [10].

Clustering techniques can be employed to first cluster the data, so that outliers are the points that do not belong to formed clusters. On the whole, the entire clustering-based techniques depend on the clusters to define outliers, as a result major concentration on optimizing clustering, not outlier detection [10]. Outlier labelling techniques, informal tests, produce a space for outlier detection. There are two motivations for using an outlier labelling technique. One is to discover probable outliers as a screening device prior to conducting a formal test. The other is to discover the extreme values away from the majority of the data not considering the distribution. Few extremely common outlier labelling parameters are Z-score, Standard Deviation (SD) technique, Turkey's method, MADe method and Median Rule [11].

Density-based methods by M. M. Breunig et al., [12] assign an outlier score to any given data point, known as Local Outlier Factor (LOF), depending on distances in its local neighborhood. LOF is unable to detect the four outliers for any size of the local neighborhood. Besides, some distance-based outlier detection work has been introduced recently [13-14]. Clearly, distance-based definitions cannot process this category of data. On the other hand, these research attempts do offer valuable thoughts for monitoring outliers. In [12], the authors emphasize that outlying is a relative concept, which should be studied in local area. In [15] and [16], the outliers are mined in subspaces, where only partial attributes are taken into account, with the intention that the curse of dimensionality is partially overcome.

Pang-Ning Tan proposed OutRank-b[17], a graph-based outlier detection algorithm. In this technique the graph representation of data depends upon two approaches- the object similarity and amount of shared neighbours among objects. Besides this a Markov chain scheme is constructed upon this graph, which allocates an outlier score to each object. Agrwal [18] has suggested a local subspace based outlier detection which uses different subspace for different objects. Most of the aforementioned techniques have only concentrated on continuous real-valued data attributes and not applied for categorical data attributes with larger dataset.

Proposed Methodology

In this paper propose a formal optimization-based model of categorical outlier detection, for which a new concept of Kullback- Leibler divergence, which captures the distribution and holoentropy with correlation information of a dataset, is proposed. Then propose an efficient Ascent-based Monte Carlo expectation–Maximization (AMCEM) clustering algorithm for outlier detection. These approaches require only the number of outliers as an input parameter and absolutely dispense with the parameters for differentiating outliers typically required by existing approaches.

Measurement for Outlier Detection

Consider an data be the X containing number of the data objects as n $(x_1, \dots . x_n)$ each x_i for $1 < i < n$ being a vector of categorical attributes $[y_1, y_2, \dots ., y_m]^T$, where m represents the number of categorical and discrete data attributes, y_j indicates the value of the attribute that belongs to either categorical and discrete value represented by $(y_{1,j}, y_{2,j}, \dots . y_{n,j})(1 < j < m)$ and n_j indicates the number of distinct values in attribute y_j. In order to measure the attribute value importance by using the Kullback-Leibler Divergence (KLD) 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 $H_X(Y)$ can be written as follows:

$$H_x(y) = H_x(y_1, y_2, \dots y_m) = \sum_{i=1}^{m} H_x(y_i|y_{i-1}, \dots . y_1) = H_x(y_1) + \cdots . + H_x(y_m|y_{m-1}, \dots y_1) \quad (1)$$

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

The total correlation [19] is defined as the sum of mutual information of multivariate discrete random vectors Y, denoted as $C_x(Y)$,

$$C_x(y) = \sum_{i=2}^{m} \sum_{\{r_{-1}, \dots r_i\} \subset \{1, \dots m\}} I_x(y_{r_1}, \dots . y_{r_i}) \quad (3)$$

$$= \sum_{\{r_1, \dots r_i\} \subset \{1, \dots m\}} I_x(y_{r_1}, ; y_{r_2}) + \cdots + I_x(y_{r_1}, \dots y_{r_m})$$

Where $r_1 \dots r_i$ are attribute numbers chosen from 1 to m. $I_x(y_{r_1}, \dots . y_{r_i}) = I_x(y_{r_1}, \dots . y_{r_{i-1}}) - I_x(y_{r_1}, \dots y_{r_i})$ is the multivariate mutual information of $y_{r_1} \dots y_{r_i}$, where $I_x(y_{r_1}, \dots y_{r_{i-1}}|y_i) = E(I(y_{r_1}, \dots . y_{r_{i-1}})|y_{r_i})$ indicates the conditional mutual information. The holoentropy $HL_x(Y)$ is described as the sum of the entropy and the total correlation of the random vector Y, and can be given by the sum of the entropies on all attributes,

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

Holoentropy allocates equal significance to the entire attributes, while in real applications solved this problem by formulating weighting technique which computes the weights straightforwardly from the data and is stimulated by increased efficiency in practical applications more willingly than by theoretical necessity.

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

Even though in the holoentropy function thus sets a minimum value for each attributes and the maximum expected number of attributes value are identified in the KL divergence. In this work use both KL and holoentropy measure. Majorly consider the Kullback-Leibler measure through probability function $p(y_i)$ &$q(y_j)$. Kullback-Leibler divergence between two different attributes probability density $p(y_i)$ and $q(y_j)$ for a specified data object x is given by,

$$D_{y_i||y_j} = \sum_{x \subset X} p(y_i) log\left(\frac{p(y_i)}{q(y_j)}\right) \quad (6)$$

The probability values of the $p(y_i)$ and $q(y_j)$ can be determined by using Parzen windows [20]. The equation (7) shows probability calculation formula of each firefly for given set of data.

$$p(y_j) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n} \varphi\left(\frac{y_i - y_m}{h_n}\right) \quad (7)$$

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 KLD is calculated then computes the weights directly from the data and is motivated by increased effectiveness in practical applications rather than by theoretical necessity

$$w_x(y_i) = 2\left(1 - \frac{1}{1+exp\left(-D_{y_i||y_j}\right)}\right) \quad (8)$$

Once the attribute importance value is calculated then perform the clustering Ascent based Monte Carlo expectation–Maximization (AMCEM) to detect the outlier.

Ascent based Monte Carlo Expectation–Maximization (AMCEM) for Outlier Detection

The expectation–maximization (EM) algorithm has become a highly appreciated tool for maximizing probability models in the presence of missing data for outlier's detection. The troubles of E-step possibly will be surpassed by approximating the expectation with Monte Carlo methods [21]. In the MCEM have also some drawbacks to detect the outlier in the data object cannot typically admit both independent sampling and Markov chain Monte Carlo (MCMC) techniques within a common framework. Subsequently, they do not attempt to imitate certain fundamentally appealing properties of the fundamental EM algorithm. To overcome these issues, in this work use Ascent-based Monte Carlo expectation maximization (AMCEM) for this process first need to define data object samples as (9). Let KLY denote a vector of observed KLD data object results for categorical data and U denote a vector of missing attributes data and let λ be a vector of unknown categorical data. Finally, $f_{KLY,U}(kly, u, \lambda)du$ denotes the probability model of the complete data to detect the outlier in the data or clustered group (KLY, U). The objective is to obtain the maximizer $\hat{\lambda}$ of

$$L(\lambda; kly) = \int f_{KLY,U}(kly, u, \lambda)\, du \tag{9}$$

Instead of directly maximizing equation (9), the EM algorithm operates on the so-called Q-function. Let $\lambda^{(t-1)}$ be the current estimate of $\hat{\lambda}$. Then the t^{th} E-step calculates,

$$Q\left(\lambda; \lambda^{(t-1)}\right) = E\left[log\{f_{KLY,U}(kly, u, \lambda)\} \,|kly, \lambda^{(t-1)}\right] \tag{10}$$

Then in the t^{th} M-step for outlier detection require a value $\lambda^{(t)}$ that satisfies $Q(\lambda^{(t)}, \lambda^{(t-1)}) \geq Q\left(\lambda, \lambda^{(t-1)}\right)$ for all λ in the parameter space, it needs to satisfy the following condition,

$$Q\left(\lambda^{(t)}; \lambda^{(t-1)}\right) \geq Q\left(\lambda^{(t-1)}; \lambda^{(t-1)}\right) \tag{11}$$

which yields a generalized EM algorithm. The ascent property is obtained with an application of Jensen's inequality to expression (11), i.e.

$$L\left(\lambda^{(t)}; kly\right) \geq L\left(\lambda^{(t-1)}; kly\right) \tag{12}$$

Approximate the expectation in equation (10) via,

$$\tilde{Q}\left(\lambda; \lambda^{(t-1)}\right) = \frac{\sum_{j=1}^{m_t} w\left(u^{(t,j)}\right) log\{f_{KLY,U}\left(kly, u^{(t,j)}, \lambda\right)\}}{\sum_{j=1}^{m_t} w\left(u^{(t,j)}\right)} \tag{13}$$

Throughout assume that $\{u^{(t,1)}, \ldots, u^{(t,m_t)}\}$ is either,

a) A random sample categorical data selected from $f_{U|KLDY}\left(u|kly, \tilde{\lambda}^{(t-1)}\right)$
b) Sample categorical data is obtained from a candidate $w_x(y_i)$ with associated weight values.
c) Obtained by simulating an ergodic markov chain with invariant density $f_{U|KLDY}\left(u|kly, \tilde{\lambda}^{(t-1)}\right)$

Where the importance weights are calculated from (a) and (c). The MCEM Algorithm uses a Q function with the t^{th} M-step consists of finding a value of $\tilde{\lambda}^{(t)}$ such that

$$\tilde{Q}\left(\tilde{\lambda}^{(t)}, \tilde{\lambda}^{(t-1)}\right) \geq \tilde{Q}\left(\tilde{\lambda}^{(t-1)}, \tilde{\lambda}^{(t-1)}\right) \tag{14}$$

If the lower bound of the current unknown data object is positive, then new estimated data object is accepted as cluster and if it is negative, this estimate of $\tilde{\lambda}$ is rejected and considered as outlier for the cluster samples. 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 $AS = \{x_i | \widehat{D}_{y_i||y_j} > 0\}$,the data objects with nonpositive set is considered as,

$$UO = N(AS) = \sum_{i=1}^{n} (\widehat{D}_{y_i||y_j} > 0) \tag{15}$$

$\tilde{\lambda}^{(t-1)}$ denotes the current sample outlier approximation results and that $\{u^{(t,j)}\}_{j=1}^{m_t}$ is the monte carlo sample. In equation (11) the inequality problems occurs it is solved by using the following representation consistently with,

$$\Delta Q\left(\tilde{\lambda}^{(t,m_t)}, \tilde{\lambda}^{(t-1)}\right) \equiv Q\left(\tilde{\lambda}^{(t,m_t)}, \tilde{\lambda}^{(t-1)}\right) Q\left(\tilde{\lambda}^{(t-1)}, \tilde{\lambda}^{(t-1)}\right)$$
$$= \frac{\sum_{j=1}^{m_t} w(u^{(t,j)}) \, log\{f_{KLY,U}(kly, u^{(t,j)}, \tilde{\lambda}^{(t,m_t)} / kly, u^{(t,j)}, \tilde{\lambda}^{(t-1)})\}}{\sum_{i=1}^{n} w(y_i)} \quad (16)$$

Where $w_x(y_i)$ is the importance of the weight value derived from (8) and sampling directly from the $f_{KLY,U}$

$$\Delta\tilde{Q}\left(\tilde{\lambda}^{(t,m_t)}, \tilde{\lambda}^{(t-1)}\right) - \Delta Q\left(\tilde{\lambda}^{(t,m_t)}, \tilde{\lambda}^{(t-1)}\right) \quad (17)$$

Has a limiting normal distribution with mean 0 and a variance σ^2 that depends on the sampling mechanism employed. It is represented as

$$\sqrt{m_t}\{\Delta\tilde{Q}\left(\tilde{\lambda}^{(t,m_t)}, \tilde{\lambda}^{(t-1)}\right) - \Delta Q\left(\tilde{\lambda}^{(t,m_t)}, \tilde{\lambda}^{(t-1)}\right)\} = \sqrt{m_t}\{\Delta\tilde{Q}\left(\tilde{\lambda}^{(t)}, \tilde{\lambda}^{(t-1)}\right) - \Delta Q\left(\tilde{\lambda}^{(t)}, \tilde{\lambda}^{(t-1)}\right)\} \quad (18)$$

Calculate an Asymptotic Standard Error (ASE) for expression (17). Consider z_α be such that $Pr(Z > z_\alpha) = \alpha$ where z is standard normal random variable. Then,

$$\Delta\tilde{Q}\left(\tilde{\lambda}^{(t,m_t)}, \tilde{\lambda}^{(t-1)}\right) - z_\alpha ASE \quad (19)$$

If the asymptotic lower bound (19) is positive, there is sufficient evidence to conclude whether the selected data object is outlier or not that $\tilde{\lambda}^{(t,m_t)}$ increases the likelihood. Thus, $\tilde{\lambda}^{(t,m_t)}, \tilde{\lambda}^{(t-1)}$ is accepted as the t^{th} parameter update, i.e. $\tilde{\lambda}^{(t,)} = \tilde{\lambda}^{(t-1)}$ and $t \to t+1$. If the lower bound is negative for data object, then the estimate of Q is deemed swamped with Monte Carlo error and a larger sample size is needed to estimate Q accurately. In this case, the t^{th} iteration is repeated with a larger sample size.

Independent sampling

If importance sampling is employed an estimate of σ^2 for each categorical data object is given by

$$\hat{\sigma}^2 = m_t \left\{\frac{\sum w\left(u^{(t,j)}\right) \Lambda\left(u^{(t,j)}\right)}{\sum w(u^{(t,j)})}\right\}^2 \quad (20)$$
$$\left[\frac{\sum\left\{w\left(u^{(t,j)}\right) \Lambda\left(u^{(t,j)}\right)\right\}^2}{\{\sum w(u^{(t,j)}) \Lambda(u^{(t,j)})\}^2} - 2\frac{\sum w^2\left(u^{(t,j)}\right) \Lambda\left(u^{(t,j)}\right)}{\{\sum w(u^{(t,j)}) \Lambda(u^{(t,j)})\} \sum w(u^{(t,j)})}\right.$$
$$\left. + \frac{\sum w^2\left(u^{(t,j)}\right)}{\{\sum w(u^{(t,j)})^2\}}\right]$$

Where the sum of all range from $j = 1, \ldots m_t$ *and*

$$\Lambda\left(u^{(t,j)}\right) = log\left\{\frac{f_{KLY,U}\left(kly, u^{(t,j)}; \tilde{\lambda}^{(t,m_t)}\right)}{f_{KLY,U}\left(kly, u^{(t,j)}; \tilde{\lambda}^{(t-1)}\right)}\right\}$$

Calculating a reasonable Monte Carlo standard error for the outlier detection data object results is more difficult when employ MCMC sampling because of σ^2 under weaker regularity conditions [22]. The above mentioned problem is solved by preferring the following steps,

$$s_r\left(\lambda, \tilde{\lambda}^{(t-1)}\right) = \sum_{j=\tau_{r-1}}^{\tau_{r-1}} log\left\{\frac{f_{KLY,U}\left(kly, u^{(t,j)}; \lambda\right)}{f_{KLY,U}\left(kly, u^{(t,j)}; \tilde{\lambda}^{(t-1)}\right)}\right\} \quad (21)$$

Consistent estimate of the desired asymptotic variance is given by,

$$\hat{\gamma}^2\left(\lambda, \tilde{\lambda}^{(t-1)}\right) = \frac{\sum_{r=1}^{R_t}\left[S_r\left(\lambda, \tilde{\lambda}^{(t-1)}\right) - \left\{\frac{\bar{S}\left(\lambda, \tilde{\lambda}^{(t-1)}\right)}{\bar{N}}\right\} N_r\right]^2}{R_t \bar{N}^2} \quad (22)$$

Now the asymptotic lower bound is changed as,

$$\Delta\tilde{Q}\left(\tilde{\lambda}^{\left(t,\tau_{R_t}\right)}, \tilde{\lambda}^{(t-1)}\right) - z_\alpha \frac{\hat{\gamma}\left(\tilde{\lambda}^{\left(t,\tau_{R_t}\right)}, \tilde{\lambda}^{(t-1)}\right)}{\sqrt{R_t}} \quad (23)$$

Updating the Monte Carlo sample size

With the intention of obtaining computational efficiency and circumventing rigorous inflation of the type 1 error rate of the outlier detection results, the preliminary sample size for each MCEM iteration should be selected, in order to go through the appending process occasionally. For MCEM iteration t, let $m_{t,start}$ be the starting Monte Carlo sample size and $m_{t,end}$ be the ending Monte Carlo sample size across MCEM iterations by taking $m_{t+1,start} \geq m_{t,start}$ assume that,

$$\Delta\tilde{Q}\left(\tilde{\lambda}^{(t+1)}, \tilde{\lambda}^{(t)}\right) \sim N\left\{\Delta Q\left(\tilde{\lambda}^{(t+1)}, \tilde{\lambda}^{(t)}\right), \frac{\hat{\sigma}^2}{m_{t+1}}\right\}$$

$$m_{t+1,start} = max\left[m_{t,start}, \hat{\sigma}^2 (z_\propto + z_\beta)^2 / \left\{\Delta\tilde{Q}\left(\tilde{\lambda}^{(t)}, \tilde{\lambda}^{(t-1)}\right)\right\}^2\right] \quad (24)$$

The validity of equation (24) evidently based on the quality of the normal approximation. A meager approximation largely results in an inflated type 1 error rate for the lower bound. The outlier factor of the specific data object from x_o, denoted as $OF(x_o)$, is defined as,

$$OF(x_O) = \sum_{i=1}^{n} OF\left(Q\left(X_{o,i}\right)\right) \quad (25)$$

Algorithm 1: Outlier Detection using AMCEM

Input: Dataset X and number of the outlier requested o

Output: Outlier results OS

Compute $w_x(y_i)$ for $(1 \leq i \leq m)$ by (8)

Initially set $OS = 0$

for $i = 1\ to\ n$ do

Compute $OF(x_O)$ from (25) and obtain AS by (15)

End for

If $O > UO$ then

$$O = UO$$

Else

Build OS by searching for the o Objects with greatest $OF(x_i)$ in AS

End if

Experimentation Results

In this section, conduct effectiveness and efficiency tests to analyze the performance of the proposed method AMCEM. 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 examination, carry out evaluations on synthetic data sets to demonstrate how running time increases with the number of objects, the number of attributes and the number of outliers. A huge number of public real data sets, most of them obtained from UCI [23], are employed in the evaluation, representing an extensive range of domains in science and the humanities. The data set used is the public, categorical "soybean data" [23], 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) [1], [2] 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; 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

DATA SET	#N	#M	#O	#UO	CNB	FIB	UNW ITB-SP	ITB-SP	UNW ITB-SS	ITB-SS	UNW AMCEM	AMCEM
Breast-c	495	11	45	125	0.99	0.90	0.894	0.991	0.898	0.993	0.995	0.996
Credit-a	413	17	30	171	0.84	0.92	0.98	0.985	0.99	0.992	0.994	0.995
Diabetes	768	9	268	340	0.86	0.88	0.76	0.75	0.84	0.912	0.93	0.945
Ecoli	336	8	9	144	0.89	0.92	0.96	0.96	0.98	0.99	0.994	0.996

The time consumption is measured with increasing numbers of objects, attributes and outliers. As Figure 1 indicates, the run times of AMCEM, ITB-SP, ITB-SS, and FIB are almost linear functions of the number of objects. Proposed AMCEM have lower and FIB has a higher increase rate than ITB-SP and ITB-SS. From the theoretical analysis, time complexity of CNB [24] increases quadratically with the number of objects, which is confirmed by the experimental data of Figure 1.

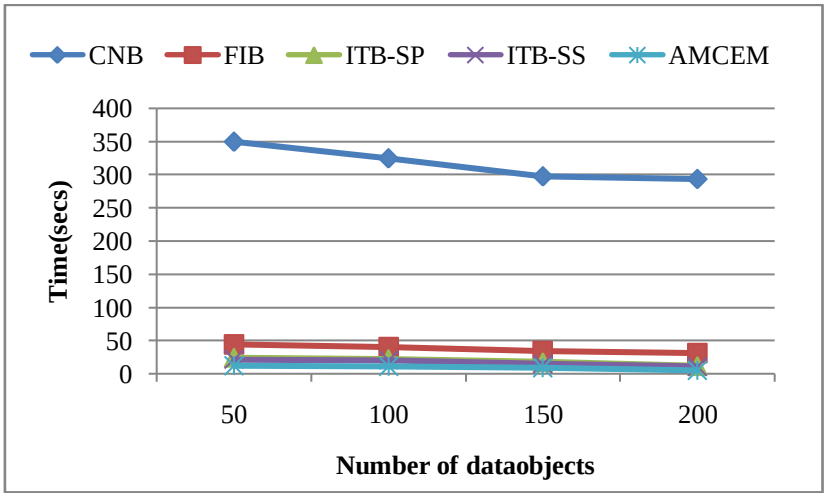

Figure 1: Results of Efficiency Real Data Sets for Data Objects Vs Methods

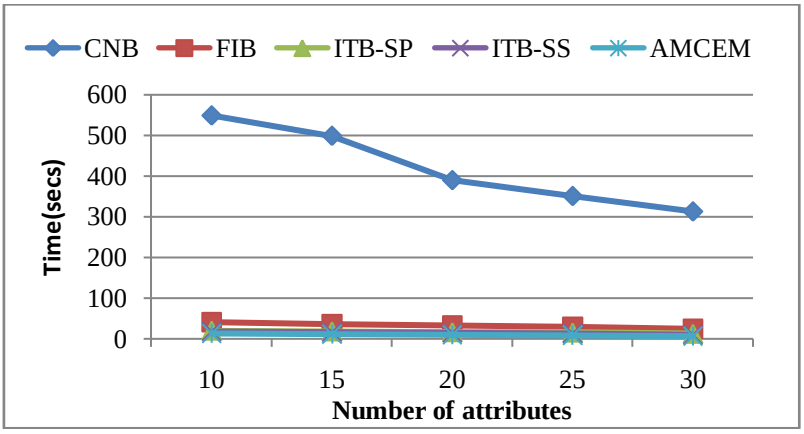

Figure 2: Results of Efficiency Real Data Sets for Data Attributes Vs Methods

For the attributes increasing test, Figure 2 shows that the run times of the AMCEM, increase rapidly with the number of attributes, which closely matches the theory that the time complexities of FIB [25] increase quadratically with the number of attributes. Compared with the time increase of FIB, CNB, ITB-SS, ITB-SP, the increases for the other methods are too small to be noticeable in Figure 2.

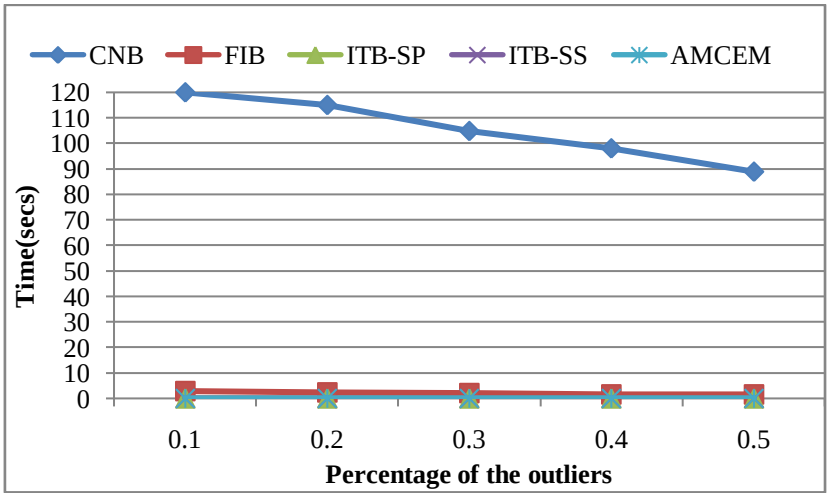

Figure 3: Results of Efficiency Real Data Sets for Percentage of the Outliers Vs Methods

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 AMCEM 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,

$$\text{NRMSE} = \frac{\sqrt{\text{Mean}[(y_{\text{guess}} - y_{\text{ans}})^2]}}{\text{std}[y_{\text{ans}}]} \tag{26}$$

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

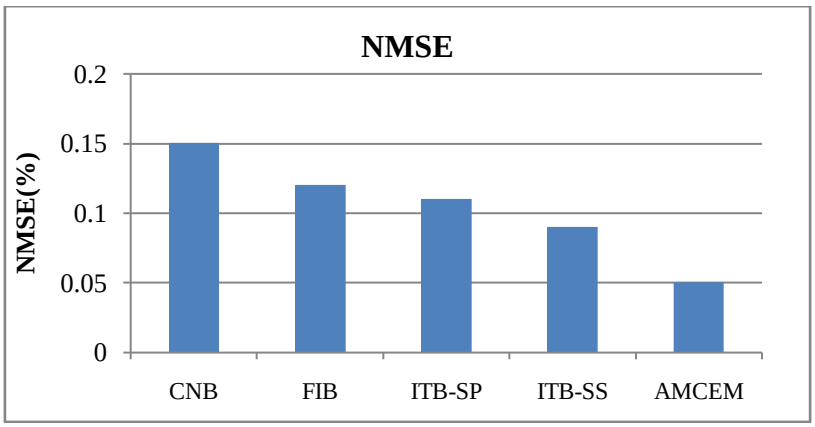

Figure 4: NMSE for Real Datasets Vs Methods

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 AMCEM algorithm, the NMSE value of the proposed AMCEM algorithm have less NMSE when compare to existing methods.

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

$$CDR = \frac{No\ of\ outliers\ correctly\ detected\ as\ outlier}{Total\ no\ of\ outlier\ in\ dataset} \tag{27}$$

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

$$FA = \frac{No\ of\ outliers\ incorrectly\ detected\ as\ outlier}{Total\ no\ of\ normal\ points\ in\ dataset} \tag{28}$$

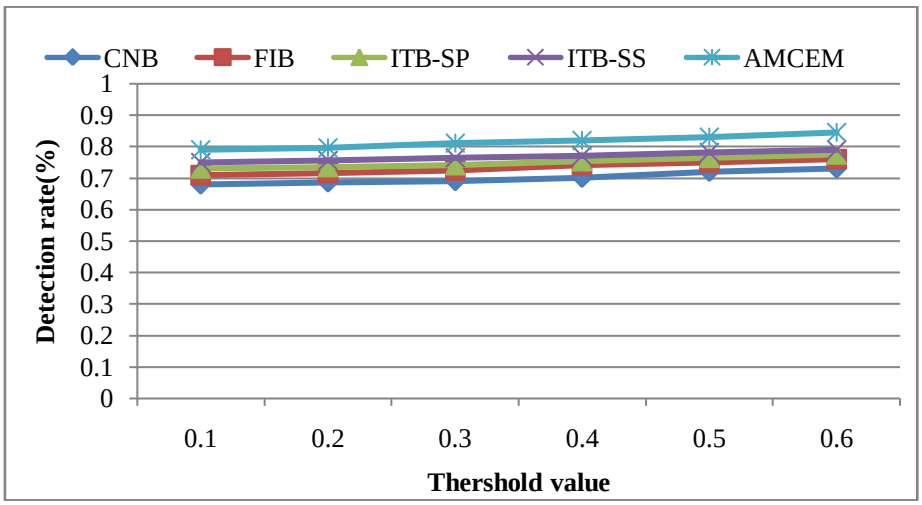

Figure 5: Detection Rate for Real Data Sets Vs Methods

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 AMCEM algorithm between the threshold value of the KLD function for each attribute. Detection Rate (DR) value of the proposed AMCEM algorithm have more DR when compare to existing methods.

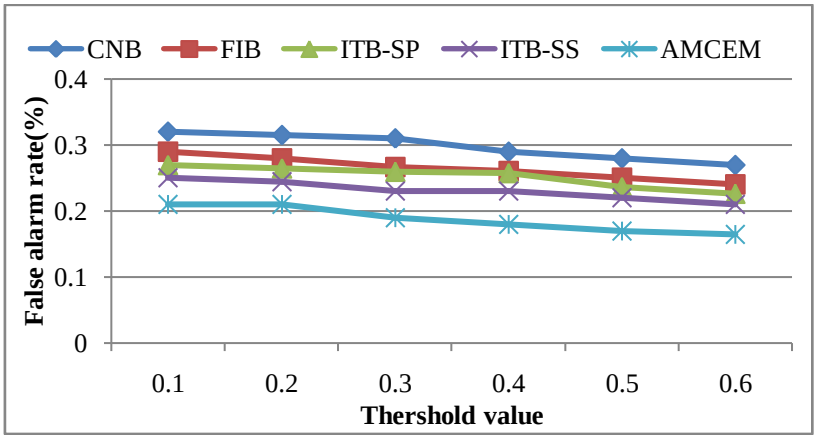

Figure 6: False Alarm Rate for Real Datasets Vs Methods

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 AMCEM algorithm among the threshold value of the KLD function for each attribute. False Alarm Rate (FAR) value of the proposed AMCEM algorithm have less FAR when compare to existing methods.

Conclusion and Future Work

Outlier Detection techniques for categorical datasets have employed using hybrid Expectation Maximization methods which combines the procedure of the Ascent Monte Carlo method so it is named as Ascent-Based Monte Carlo Expectation–Maximization(AMCEM) to identify those points containing irregular patterns. The proposed weighted KLD measure the attribute value with maximum likelihood of outlier candidates, while the efficiency of the algorithms results from the outlier factor function. The outlier factor of an object is solely determined by the object and its updating does not require estimating the data distribution. The proposed method is specifically applied for UCI machine learning repository. The proposed AMCEM also estimate an ascent property for the number of outliers and an anomaly candidate set. This bound, acquired under an extremely practical hypothesis on the number of feasible outliers, permits to additionally reduce the search cost. Future research includes additionally enhancing the speed and extending for distributed datasets. In this paper, the datasets on which the proposed approach is evaluated are of integer or real type. As a result, in future it can be extended to work other type datasets.

References

[1] Bolton, R., and Hand, D., 2002, “Statistical Fraud Detection: A Review”, Statistical Science, 17(3), pp. 235-255, 2002.

[2] Dokas, P., Ertoz, L., Kumar, V., Lazarevic, A., Srivastava, J., and Tan, P., 2002, "Data Mining for Network Intrusion Detection", Proc. NSF Workshop on Next Generation Data Mining.

[3] Leckie, T., and Yasinsac, A., 2004, "Metadata for Anomaly-based Security Protocol Attack Deduction", IEEE Trans. Knowledge and Data Eng., 16(9), pp. 1157-1168.

[4] Hodge, V.J., and Austin, J., 2004, "A Survey of Outlier Detection Methodologies", Artificial Intelligence Rev., 22(2), pp. 85-126.

[5] Aleskerov, E., Freisleben, B., and Rao Cardwatch, B., 1997, "A Neural Network based Database Mining System for Credit Card Fraud Detection", Proc. IEEE/IAFE Computational Intelligence for Financial Eng. Conf.,.

[6] Joshi, M., Agarwal, R., Kumar, V., and Nrule, P., 2001, "Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Indection", Proceedings of the ACM SIGMOD Conference.

[7] Otey, M., Ghoting, A., and Parthasarathy, S., 2006, "Fast Distributed Outlier Detection in Mixed-Attribute Data Sets", Data Mining and Knowledge Discovery, 12(2), pp. 203-228.

[8] Hays, C., "What Wal-Mart Knows about Customers Habits", The New York Times, 2004.

[9] Knorr, E., Ng, R., and Tucakov, V., 2000, "Distance-based Outliers: Algorithms and Applications", VLDB Journal, International Journal on Very Large Data Bases, 8(3), pp. 237-253.

[10] Bay, S., and Schwabacher, M., 2003, "Mining Distance-based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule", Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 29-38.

[11] Seo, S., 2002, "A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets", BS, Kyunghee University.

[12] Breunig, M. M., Kriegel, H. P., Ng, R. T., and Sander , J., 2000, "Lof: Identifying Density-based Local Outliers", Proceedings of ACM SIGMOD on Management of Data, pp. 386-395.

[13] Ghoting, A., Parthasarathy, S., and Otey, M. E., 2008, "Fast Mining of Distance-based Outliers in High-Dimensional Datasets", Data Mining and Knowledge Discovery, 16, pp. 349-364.

[14] Angiulli, F., and Fassetti, F., 2007, "Very Efficient Mining of Distance-based Outliers", Proc. of 16th ACM Conf. on Information and Knowledge Management.

[15] Knorr, E., and Ng, R., 1999, "Finding Intensional Knowledge of Distance-based Outliers", Proc. of 25th Int'l Conf. on Very Large Data Bases, pp.211–222.

[16] Aggarwal, C., and Yu, P., 2001, "Outlier Detection for High Dimensional Data", Proc. of ACM SIGMOD Int'l Conf. on Management of Data, ACM Press, pp.37–47.

[17] Moonesinghe, H. D. K., and Tan, P. N., 2007, "Outrank: A Graph-based Outlier Detection Framework using Random Walk", International Journal on Artificial Intelligence Tools, 100(10), pp. 1-18.

[18] Agrwal, A., 2009, "Local Subspace based Outlier Detection", IC3 2009, CCIS 40, pp. 149–157.

[19] Srinivasa, S., 2005, "A Review on Multivariate Mutual Information", Univ. of Notre Dame, Notre Dame, Indiana, Vol. 2, pp. 1-6.

[20] Nakariyakul, S., and Casasent, D., 2008, "Improved Forward Floating Selection Algorithm for Feature Subset Selection", IEEE Int. Conf. Wavelet Analysis and Pattern Recognition, Vol. 2, pp. 793–798.

[21] Shi, J.Q., and Copas, J., 2002, "Publication Bias and Meta-Analysis for 2×2 Tables: An Average Markov Chain Monte Carlo EM Algorithm", J. R. Statist. Soc. B, Vol. 64, pp. 221–236.

[22] Jones, G.L., Haran, M., and Caffo, B.S., 2004, "Output Analysis for Markov Chain Monte Carlo Simulations", Technical Report. School of Statistics, University of Minnesota, Minneapolis.

[23] UCI Machine Learning Repository, http://www.ics.uci.edu/learn/ML Repository. html, 2011.

[24] 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).

[25] 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, pp. 103-118.