Enhanced Outlier Detection Method Using Association Rule Mining Technique

S.Preetha
M.Phil Scholar
Department Of Computer Science
Avinashilingam University for Women
Coimbatore-43.

V.Radha
Associate professor
Department Of Computer Science
Avinashilingam University for Women
Coimbatore-43

ABSTRACT
In recent years, much attention has been given to the problem of outlier detection, where the aim is to identify data which behave in an unexpected way or have abnormal properties. This paper focuses on the development of non-parametric association based algorithm to effectively identify classes and outliers simultaneously. The proposed algorithm identifies outlier transaction by enhancing an association classification approach using FP-Growth. The algorithm is enhanced by using an automatic procedure for calculating the minimum support and minimum confidence automatically and introduces two new measures called collective support and confidence measure. Using these thresholds, frequent itemsets and association rules are generated. Pruning algorithms and redundant rule identification and removal procedures are used as speed optimizers. An outlier degree using the threshold is called for each association rules, using which high confidence rules are identified. The various experimental results prove that the proposed model are scalable and efficient in terms of outlier detection and classification accuracy and can be used by data mining techniques for accurate and fast knowledge discovery.

Keywords
Outlier, Associative Rule, Itemsets, Confidence, Support, Fp-growth.

1. INTRODUCTION
Data mining is a technology with great potential that predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining is performed in various stages like preprocessing, clustering, classification and prediction. Out of this preprocessing is the focal area of this paper. Preprocessing transforms raw unprocessed data into a format that will be more easily and effectively processed by the subsequent data mining steps. Analyzing data that has not been carefully screened often produce misleading results. Therefore, using preprocessing routines that improves the representation and quality of data is an important task that should be performed before running an analysis [4].

Based on this, many applications are performing preprocessing as a mandatory step during data mining. In general, the preprocessing stage of data mining consists of five stages, namely, data cleaning, data integration, data transformation, data reduction and data discretization. All the above tasks focus on increasing the data quality. Data cleaning processes uses routines that can handle incomplete, noisy and inconsistent data and is the focal point of this paper. Data cleaning is also referred to as outlier detection or noise removal. Even though outlier detection is a task that has been probed by various researchers [3], works related to transaction database is minimal [2]. This study focuses on integrating outlier detection and associative techniques for removing outliers in transaction databases for classification.

In [6] proposed a detection method that detects outliers in transaction databases. They proposed a framework for detecting outlier transactions that behave abnormally compared to others. They modified a frequent pattern mining algorithm to detect outliers and proved that the algorithm is efficient in terms of accuracy and running time. In this algorithm, a crucial step is the selection of support and confidence value. The accuracy of outlier detection has a direct impact on the correct selection of these two value. In this paper, an automatic method for finding these parameters is presented and is combined with the system proposed by [6]**.** The rest of the paper is organized as follows. A brief overview of transactional database and outliers is presented in Section 2. The proposed method is described in Section 3. The results of experimentation and performance analysis of the proposed method are presented in Section 4, while Section 5 concludes the work with future research directions.

2. TRANSACTION DATABASE AND OUTLIERS
A transactional database is defined as a database that consists of one or more data-manipulation statements and queries, each reading and/or writing information in the database. The transactional databases can mine and manipulate tremendous amounts of information about an individual's personal lives, habits and transactions. The most famous example of transaction data is market basket data, where each transaction

corresponds to the set of items bought by a customer during a single visit to a store. With market basket transaction database, a dataset is typically in a table format. Each row is a transaction, identified by a transaction identifier or a TID. A transaction contains a set of items bought by a customer.

An example set is shown in Figure 1.

ItemID	Item
A	Cereals
B	Beer
C	Diaper
D	Egg
E	Milk

(a) Item List

TID	Items
1	{Cereals, Milk}
2	{Beer, Cereal, Diaper, Egg}
3	{Beer, Diaper, Milk}
4	{Beer, Cereal, Diaper, Milk}
5	{Diaper, Milk}

(b) Transaction Database

Figure 1 : Example Transaction Database

To understand the transactions outliers, consider the following. Let 'T' be a transaction database with set of transactions. Let |T| indicate a set cardinality of database T, that is, the number of transactions included in T and let 't' be the set of items in a transaction ($t \in T$). The set cardinality |t| is the number of items for t. Let I be a set of all items that exist in transaction database T and let $X \subseteq I$ be an itemset. X's support sup(X) on T is defined as

$$sup(X) = | \{t / t \in T \text{ for every } X \subseteq t\} | / | T | \quad (1)$$

An itemset X is said to be a frequent itemset (FI), if its support is greater than a user-defined minimum support (min_sup), that is, $sup(X) \geq min\_sup$. An frequent itemset (derived with min_sup), which has no other frequent itemsets as a superset, is called maximal frequent itemset, denoted as maximal FI. For two itemsets X, $Y \in I$ such that $X \cap Y = \varnothing$, then the statement that represents the association between X and Y is terms as an association rule. An association rule is of the form $X \rightarrow Y$ and portrays the relation that whenever X occurs, then Y also occurs. The left part is called the antecedent and the right part Y is called the consequent. From the support value, another important parameter called confidence of an association rule $X \rightarrow Y$ can be calculated and is defined as $conf(X \rightarrow Y) = sup(X \cup Y)/sup(X)$. In general, to select quality rules for analysis, only those rules which have confidence value greater than a minimum confidence (min_conf) are chosen. That is, an association rule is selected only if $conf(X \rightarrow Y) \geq min\_conf$ and these rules are termed as high confidence rules. This research work focus on the high-confidence rules in which the confidence is abnormally high.

From these definitions, a violated transaction rule (outlier transaction rule) denoted as 'Unobserved Rule" is defined as "For an association rule $X \rightarrow Y$, when an itemset $Z \subseteq I$ satisfies $X \subseteq Z$ for every $Y \not\subset Z$, Z is said to violate for $X \rightarrow Y$. Beyond that, such an association rule $X \rightarrow Y$ is called Z's unobserved rule.

3. PROPOSED METHOD

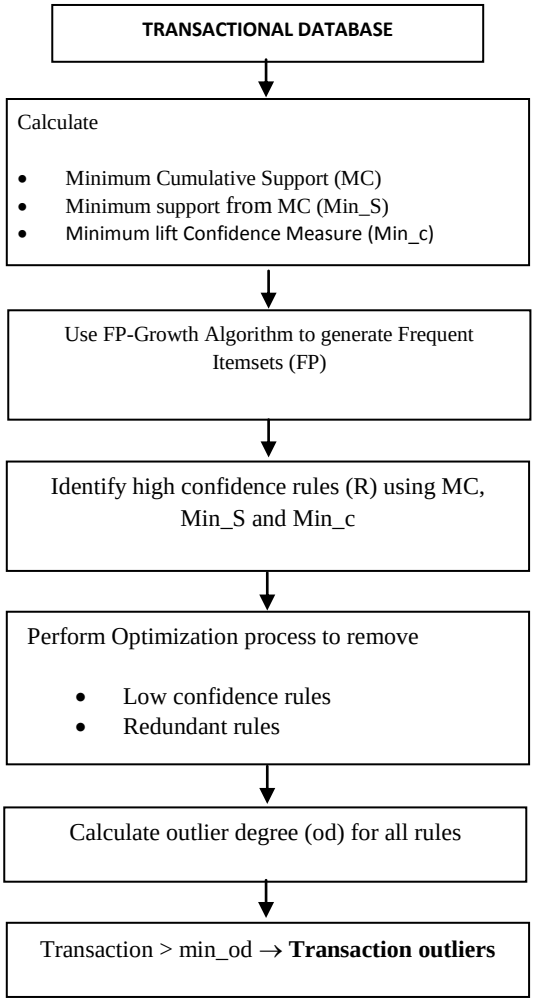

Figure 2: Methodology

The proposed algorithm for detecting outliers in transactional database is shown in Figure 2. The FP-Growth algorithm proposed by [1] is used for frequent pattern mining.

3.1 Automatic Calculation of Minimum Confidence

As the calculation of outlier detection depends on high confidence rules, it becomes imperative that the selection of minimum confidence, which is derived using minimum support, is crucial. The primary deficiency of confidence-based associations is their poor predictive ability, i.e. the confidence measure is unable to capture the real implication. To remedy the above deficiency, two alternative measures have been proposed. They are lift (also known as interest) and conviction. For an association rule $A \Rightarrow B$, the lift is defined using Equation (2). Lift measures the deviation of the rule from independence. The farther the value is from 1, the higher

the dependence will be. Lift values above 1 indicate positive dependence, while those below 1 express negative dependence. The conviction is defined using Equation (3).

$$lift(A \Rightarrow B) = \frac{P(A \cup B)}{P(A)P(B)} = \frac{\sup(A \cup B)}{\sup(A)\sup(B)} = \frac{conf(A \Rightarrow B)}{\sup(B)} \quad \textbf{(2)}$$

$$conv(A \Rightarrow B) = \frac{P(A)P(\neg B)}{P(A \cup \neg B)} = \frac{1 - \sup(B)}{1 - conf(A \Rightarrow B)} \quad \textbf{(3)}$$

which measures the implication strength of the rule from statistical independence. The conviction value of a rule is between 0 and ∞. A larger than 1 value indicates that it is greater than the expected presence. Conviction appears to be preferable to lift in capturing the natural semantics of directed associations because it is directed, i.e. conv(A⇒□B)≠conv(B⇒□ □A), whereas lift is not. Furthermore, conviction has better discrimination power. Using the lift and conviction methods, two thresholds can be derived. They are 'collective support' and 'confidence measure'. To calculate the collective support, the itemsets are first grouped according to their length and then the collective support is calculated from the combination of its individual support from previous level. From the collective support, the confidence measure is calculated using the Lift measure. The Lift measure used in the proposed work is called Confidence Lift, as it is used to measure the level of confidence of an association rule. It is calculated from two measures, namely, the minimum support of the items in the itemset and the collective support from the itemsets previous level. The frequent item sets are then identified using the Collective support and Confidence Lift Support from which the association rules are generated.

3.2 Optimization Process

As mentioned previously, to improve the fastness of the algorithm two steps are introduced. The first step is the pruning of candidates of outlier transactions utilizing maximal frequent itemsets and the second step is concerned with the removal of redundant association rules.

Step 1 Process

During experimentation, it has been observed that **f**or a set of FIs, let R_1 be a set of high-confidence rules derived with min_confidence (min_c_1) and R_2 be that derived with min_c_2. When min_c_1 < min_c_2, then its outliers $od_1(t) \geq od_2(t)$. Based on this the Maximal Associative Closure can be derived as follows. Let M be a set of all maximal FIs derived from a transaction database T. For an itemset t ⊂ I, t's maximal associative closure t^+ max is derived using Equation (4).

$$t^0_{max} = t$$

$$t^{i+1}_{max} = t^i_{max} \cup \{e \in mi \mid mi \in M \text{ for every } mi \cap t^i_{max} \neq \theta\}$$

$$t^+_{max} = t^\infty_{max} \quad \textbf{(4)}$$

It is further observed that for a transaction t ∈ T and a minimal support calculated from cumulative support, a maximal associative closure t^+_{max} equals t's associative closure t^+_0, which is derived with min_c = 0%. Thus, the proposed outlier degree derives an upper bound for a minimal support value. The Upper Bound of Outlier Degrees is defined as follows.

For a transaction t ∈ T, a minimal support from collective support and a minimal lift confidence measure, let t^+_{max} be t's maximal associative closure and let od(t) be t's outlier degree. Then, od(t)'s upper bound odmax(t) is derived as

$$od_{max}(t) = \frac{|t^+_{max} - 1)}{|t^+_{max}|} \quad \textbf{(5)}$$

It should be noted that a transaction t, whose upper bound is less than a given minimal outlier degree, is not regarded as an outlier. Thus the proposed algorithm finds Maximal FIs and generates a complete set of maximal FIs M in the step of frequent itemset mining. Generally, the number |M| is much less than that of high-confidence rules |R|. Thus the maximal outlier degrees of all transactions can be calculated in a faster rate than calculating the outlier degree of all transactions. The pruning process is performed by removing all transactions whose outlier degree upperbounds are less than min_od.

Step 2 Process

This sub-section describes the method for pruning redundant rules. To calculate outlier degrees, initially an associative closure t+ of each transaction t ∈ T using a complete set of high-confidence rules R is created. Practically, however, R has many high-confidence rules that are not needed to calculate associative closures. It is a well-known fact that a set of all high-confidence rules for a minimal confidence has obviously redundant rules for making associative closures. Non-redundant Rules are defined as those rules which has no other association rule Z→W ∈ R and S→V ∈R such that (i) X∪Y = Z ∪W for every X ⊂ Z and (ii) X = S for every Y ⊂ V respectively. A set of all non-redundant rules in R is called the minimal rules set for R. From this definition, a minimal rules set naturally has the two properties listed below.

1. For a set of all high-confidence rules generated from a set of FIs F, a minimal rules set R_{min} derived also from F is uniquely decided and $R_{min} \subseteq R$ thus $|R_{min}| \leq |R|$.

2. Let F be a set of FIs, R be a set of all high-confidence rules generated from F, and R_{min} be the minima rules set for R. For an itemset t, let t^+ be t's associative closure derived by R and t^+_{min} be t's associative closure derived by R_{min}. Then, $t^+ = t^+_{min}$.

From the above two properties, while generating high-confidence rules from F, a step to derive R_{min} is included. R_{min} is the set having only non-redundant high life-confidence rules from R.

3.3 Outlier Degree

The method for calculating the outlier degree from association rules for outlier detection is outlined in this section. An association rule X → Y with a high confidence means that when X occurs, then Y occurs with high probability. That is, when X occurs in a transaction, then all items included in Y should also occur in the transaction. Violation of the rule is an indication of the outlier transactions. The outlier degree used in the proposed work is based on associative closure property of a transaction and is described below.

Let 't' be the set of transactions (t ⊆I) with high confidence rules (R), then its associative closure (t^+) is defined as below.

$t^0 = t$

$t^{i+1} = t^i \cup \{e / e \in Y \text{ for every } X \subseteq t^i \text{ for all } X \rightarrow Y \in R\}$

$t^+ = t^\infty$ **(6)**

According to this definition, for a transaction't', the cardinality of itemset t^{i+1} increases, when the itemset t^i has unobserved rules. If t^i has no unobserved rules, then t^{i+1} converge and become t's associative closure t^+. The associative closure, t^+, is unique for each't'. When the number of items with strong dependency is small in a transaction't', then the difference between t and its ideal form t^+ becomes larger. If't' has fewer unobserved rules, then t^+ is similar to t and is not an outlier.

Outlier Degree

Let t^+ be the associative closure of transaction t ∈ T for high-confidence rules set R. Then t's outlier degree is derived by the formula below.

$$od(t) = \frac{|t^+ - t|}{|t^+|} \quad \textbf{(7)}$$

When $t^+ = t$, od(t) = 0. Only when t = ∅, od(t) = 1. Thus, od is always a value between 0 and 1, with lower bound of od(t) is 0 and upper bound is od(t) < 1. Rules with 100% confidence are ignored from outlier checking, since they have items which always have high confidence and cannot be unobservabed rules for any other transactions. From the outlier degree, the outlier transactions can be identified. A transaction 't' (t ∈ T) is said to be an outlier transaction, if od(t) ≥ min_od, where min_od is the minimum outlier detection and is user defined.

3.4 Outlier Transaction Detection Algorithms

Given a transaction database T with't' transactions, I itemsets, min_supp, min_conf, collective support measure, confidence measure and outlier degree, the proposed algorithm for outlier transaction detection algorithm consist of three basic steps.

Step 1: Create frequent itemsset F from all FI on T win min_sup calculated from cumulative support measure

Step 2: Create high confidence rule (R) from F satisfying the confidence measure

Step 3: Identify outlier transactions OT from T using R and min_od. (Detect_Outliers Algorithm)

From the generated rules, the Detect_Outliers algorithm (Figure 3) is used to detect transaction outliers. The algorithm accepts as input t, R, min_od and creates associative closures of each transaction, calculates the outlier degree and returns a set of outlier transactions. The algorithm after inclusion of non-redundant rules is described in detail by [6].

Function getOutliers(database T,rulesset R, threshold min_od)

1. **foreach** $t \in T$
2. $t+=t;$
3. $size=0;$
4. **while** $size<|t^+|$
5. $size = |t^+|, ttmp=\emptyset;$
6. **foreach** $X \rightarrow Y \in R$
7. *if* $X \subseteq t+$ **then** $ttmp= ttmpU\ Y;$
8. $t+= t+\ U\ ttmp=ttmp=\emptyset;$
9. $od(t)= |t^+ - t|/|t^+|;$
10. *if* $od(t) \geq min\_od$ **then** $OT=OT\ U\ \{t\};$

Figure 3: Outlier Detection Algorithm

4. EXPERIMENTAL RESULTS

The efficiency of the proposed algorithm was tested using various test data and performance metrics. During the experiments, two real life dataset and one synthetic dataset are used. The real life datasets selected are Abalone, Credit Approval and Annealing datasets obtained from UCI repository [7]. All the experiments were conducted in Windows environment on a Pentium IV machine with 2 GB RAM. The datasets used are summarized in Table I.

Table I: Characteristics Of Datasets Used During Experimentation

S. No.	Dataset	Type	No. of Instances	No. of dimensions	No. of Classes
1	Abalone	Multivariate	4177	8	2
2	Credit card Approval	Multivariate	10384	14	2
4	Synthetic	Multivariate	Parameter-based Generation		

The efficiency of the proposed method was analyzed using three performance metrics namely, Normalized Root Mean Square Error (NRMSE), classification accuracy, outlier detection rate and speed of the algorithm.

The performance of the classifier when treated with parametric existing outlier detection algorithm and non-parametric proposed outlier detection algorithm with respect to Normalized Root Mean Square Error (NRMSE) and Classification accuracy is shown in Table II.

Table II: Normalized Root Mean Square Error (Nrmse)

Datasets	Outlier (%)	Existing Method		Enhanced Method	
		NRMSE	Accuracy	NRMSE	Accuracy
Abalone	0	0.5377	82.45	0.4816	83.14
	10	0.5163	86.12	0.4506	86.44
	20	0.5121	85.64	0.4392	85.71
	30	0.5119	87.00	0.4643	87.43
	40	0.5139	88.46	0.4682	89.88
CreditApproval	0	0.5723	83.69	0.5428	84.17
	10	0.5381	85.15	0.4823	85.46
	20	0.5466	88.96	0.4834	89.03
	30	0.5598	90.55	0.5180	90.81
	40	0.5582	91.08	0.5260	92.52
Synthetic	0	0.7591	80.97	0.7216	81.36
	10	0.7396	83.19	0.6904	84.93
	20	0.7358	85.62	0.6926	87.67
	30	0.7219	87.09	0.7016	89.94
	40	0.7335	88.01	0.7062	90.77

From the above table, it is clear that the proposed non-parametric outlier detection algorithm perform in an improved manner than parametric outlier detection algorithm. This is evident from the huge difference between the NRMSE values before and after introducing outliers in the real-life datasets and synthetic datasets. Further, it is also clear that the classification accuracy has improved after the removal of outlier transactions, which is evident from the increase in classification accuracy when compared with 0% outlier insertion. Zero percent indicates the original transaction database with outlier transactions. It could also be seen that the classification accuracy increases when more number of outliers are detected and removed, which is evident from the increasing trend observed in accuracy with varying outlier percentage.

The outlier detection rate of the proposed and base models was shown in Table III . The minimum support and minimum confidence thresholds for the base model was set to 65% and 90% during experimentation.

The results with respect to outlier detection rate again prove that the proposed system is an advanced version of the base model in detecting outlier transactions. Further, it can also be observed that the performance of outlier detection increases with the amount of outlier transaction present in the **transaction database.**

Table III: Outlier Detection Rate (%)

Datasets	Outlier (%)	Existing Method	Enhanced Method
Abalone	10	78.81	84.34
	20	79.22	87.44
	30	80.16	88.46
	40	81.94	90.19
CreditApproval	10	83.42	94.43
	20	79.69	85.73
	30	80.25	87.40
	40	82.58	89.90
Synthetic	10	74.56	82.54
	20	76.70	85.14
	30	79.11	88.52
	40	82.01	91.92

5. CONCLUSION

The method proposed algorithm in this paper identifies outlier transaction by enhancing an association classification approach using FP-Growth. The various experimental results projected prove that the proposed model are scalable and efficient in terms of outlier detection and classification and can be used by data mining techniques for accurate and fast knowledge discovery. The detection time of the proposed model can be reduced, if the processes can be parallelizing. This is feasible, by identifying operations that are independent to each other and propose a parallel architecture to improve the performance. Moreover, the amount of memory used during the detection and classification processes is another area which can be analyzed in future. Pruning algorithms can be developed for this purpose.

REFERENCES

[1] Borgelt, C. (2005) An implementation of the FP-growth algorithm, Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations (OSDM '05). ACM, New York, NY, USA, Pp. 1-5.

[2] Dash, M. and Lie, N.W. (2010) Outlier detection in transactional data, Journal Intelligent Data Analysis, Vol. 14, Issue 3, Pp. 283-298.

[3] Dhaliwal, P., Bhatia, M.P.S. and Bansal, P. (2010) A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner), Journal of Computing, Vol. 2, Issue 2, Pp. 1-8.

[4] Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. (2006) Data Preprocessing for Supervised Leaning, International Journal of Computer Science, Vol 1 No. 2, Pp 111-117.

[5] Meratnia, Y.Z. and Havinga, P.N., (2010) Outlier Detection Techniques for Wireless Sensor Networks: A Survey, IEEE Communication Surveys and Tutorials, Vol. 12, Issue 2, Pp. 159-170.

[6] Narita, K. and Kitagawa, K. (2008) Outlier Detection for Transaction Databases using Association Rules, The Ninth International Conference on Web-Age Information Management, IEEE Computer Soeciety, Pp. 373-380.

[7] Newman, D.J., Hettich, S., Blake, C.L. and Merz, C.J. (1998) UCI repository of machine learning databases.