ISSN 2063-5346

APPLICATION OF PROBABILITY BASED SURPRISING MEASURE IN OUTLIER DETECTION

A.M.Rajeswari[1], B.Subbulakshmi[2], M.Nirmaladevi[3] and M.Sivakumar[4]

[1] Associate Professor, Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, India
amrtce@gmail.com

[2] Assistant Professor, Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, India.
bscse@tce.edu

[3] Assistant Professor, Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, India.
nirmaladevi2004@gmail.com

[4] Associate Professor, Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, India.
mskcse@tce.edu.

Article History: Received: 12.05.2023 **Revised:** 25.05.2023 **Accepted:** 05.06.2023

Abstract

Outlier detection was instigated as noise removal technique for enhancing the prediction accuracy of Machine Learning (ML) algorithms. In due course, outlier detection emerged as the phenomenon of mining rare/alarming patterns to assist in decision making process. Outliers can be point or collective types and can be detected by the supervised and unsupervised ML algorithms. Such algorithms have the probability of missing out certain point or collective outliers in high dimensional quantitative data. To overcome the aforementioned issue, this work proposes a Semi Supervised Outlier Detection (SSOD) algorithm with a probability based surprising measure 'Lift' for outlier detection. The performance proposed SSOD is benchmarked with the existing outlier detection ML algorithms. By studying performance of the algorithms, it is understood that the proposed SSOD with Lift measure outperforms the benchmarked ML algorithms.

1. Introduction

Outliers are the rare observations as in the case of weak performers in education data, ozone days in weather data, hike or dip in the shares sold in the stock market data, fatal cases in accident data, malignant cases in cancer data, etc. These rare occurring events always used to tip off the system or domain experts. Identifying such outliers will give an alarm and help in taking the required precautionary measures. Based on the outlier score the rare observations will be weak outliers (noise) or strong outliers (help in decision making). "Usually, Noise possesses less outlier score than the actual Outliers [1]". Hence it makes a clear statement that identifying the outlier score of every observations is a must to segregate the rare cases from the normal observations and noise. Most of the existing works made

use of subjective measures [2], [3], [4], [5], [6] to detect the outliers in various domains, where the researcher considered the nature of the data along with their objective. The subjective measurements [7] require user interaction such as setting the threshold value for the measurement in order to prune outliers and, therefore, will be domain specific and dependent on the domain expert. But the proposed work contribute to fix a common objective measure (based only on the probability of occurrence of data) which can be suitable for pruning the outliers in a wide set of applications without user's interaction. Such a common measure to prune outliers is 'Lift' which is classified as 'Surprising measure' [7]. "Lift is used to measure the independency among the unexpected correlated item. The unexpected correlated items with high Lift value (one and above) represents the rare items (the nature of the outliers)" [8].

Outliers may be point (single occurrences here and there) or collective (occur in small group) outliers. The occurrence of ozone days in weather data, the pre-diabetic conditions of the patients in diabetes data and the hike, dip in the returns of the stock market are example of point outliers [9]. The occurrence of malignant cases in cancer data and weak performers in the examinations of education data falls under the collective outliers. The figure 1 gives the clear picture of these types of outliers.

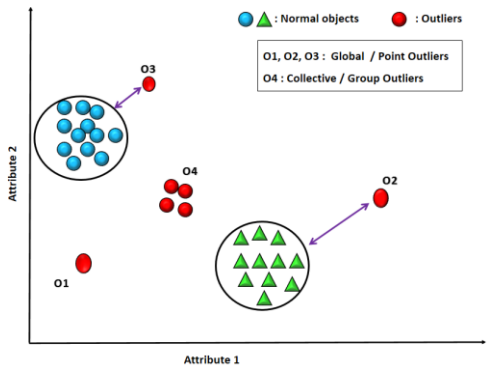


Figure 1. Types of Outliers.

classification (supervised) and clustering (unsupervised) are well suitable for detecting outliers based on single dimension, they will miss certain small groups of collective outliers like 'O4' in figure 1 and point outliers like 'O1', 'O2' and 'O3' of figure 1 in high dimensional data. A best suitable approach for detecting such outliers in high dimensional quantitative data is correlation based technique. A well-known correlation based approach is Association Rule Mining (ARM) [10], [11] an unsupervised learning algorithm.

The proposed method - SSOD (Semi Supervised Outlier Detection) makes use of the hybrid version of ARM (unsupervised) and classification (supervised) techniques and can be called as Associative Classification technique. The proposed SSOD algorithm make use of the class label only during the rule generation phase and not during the infrequent pattern generation phase. Since, the proposed associative classifier does not use the class label (target attribute) during the pattern generation phase we claim that this is a semi supervised based learning approach. By doing so, the algorithm will not miss the point outliers and also the small groups of collective outliers in multi-dimensional, multi-labeled and imbalanced database like stock market data. The performance of proposed method – SSOD along with the surprising measure 'Lift' in outlier detection is examined along with various conventional supervised and unsupervised outlier detection algorithms. The proposed method - SSOD found to outperform the benchmarked algorithms by detecting more precise outliers.

The rest of the manuscript is structured to have part 2 addresses the related work. Section 3 provides a detailed description of the preliminaries required to implement the proposed methodology. Section 4 deals with the algorithm and Section 5 details

the experiments performed with the proposed methodology - SSOD. Lastly, in Section 6, the proposed work has been summarized.

2. Related Work

When the databases have an adequate amount of normal and outlier samples then the supervised data mining algorithms will be used to detect outliers. In this supervised approach, the model is generated from the existing samples and based on this model the outliers can be identified. The samples if it is not perfect then outlier detection will be at risk. Similarly, to detect outliers, the unsupervised ML algorithms require number of clusters (groups) as user input and a proper measure with accurate threshold value from the domain experts to identify the outliers from the normal objects. For this reason, the proposed SSOD method adopts the semi-supervised associative classification technique and the surprising measure - Lift to extract outliers without any value defined by the user.

The mining of association rules (Apriori algorithm) is the first and most frequent pattern-based mining technique [10], [11]. Initially, this method was introduced to investigate the correlation between elements which frequently occur together. Subsequently, its application concentrates on the study of the correlation between infrequent elements as well. An algorithm called RARM [12] was proposed by Romero et al., to detect infrequent student behaviour and activities in an online learning environment by generating sets of rare items. The authors compared RARM performance with existing algorithms such as Apriori-Infrequent, Apriori-Reverse, and Apriori Rare. Unexpected temporal association rules are generated by an approach based on frequent models to alert equity market systems [3]. In this work, the normal behavior of objects in the stock exchange database is discovered with the help of Temporal Association Rules (TAR). Then, the relationship between the characteristics of these rules over time is identified using quasi-functional dependency and, finally, a predefined measurement called "degree of dependency" is used to extract outliers. Also, the stock splits that occur during a particular period that alert the stakeholders for their investment were identified using TAR with ‘residual leverage’ as the pruning measure [2]. Unexpected episodes to detect the adverse drug reaction in the medical domain were identified using TAR [4], [5], [6]. Preetha et al., came up with a non-parametric FP-Growth algorithm to detect outliers [13]. All of the above work was performed with pre-defined fixed threshold values for all interesting measurements used to extract outliers.

Recent studies show that associative classification is spontaneous, effectual and has good classification accuracy [14], [15]. This method considers the rules with the highest confidence for classification. In imbalanced class distribution, the class rules which cover the minor groups are missed in classification technique due to the presence of least supportable classes. But Associative classification can generate Class Association Rules (CARs) which is supportable within the class rather than in the whole database [15], [16]. Thus it can generate a complete set of rules for classification.

3. Preliminaries

This section describes the basic requirements like the dataset considered for evaluation and the interesting measures that are used to detect the outliers by the proposed method.

3.1. Dataset used

In our previous work [9], we have shown that the detection of outliers by existing supervised and unsupervised machine learning algorithms falls short. Data sets such as Pima Indian Diabetes Dataset [17],

Education Dataset (our Student Department Data), Yahoo Finance Stock Market Dataset [18] are reviewed and evaluated in our work [9**].** The same data set is used for the assessment of the proposed methodology – SSOD.

Table 1. Specifications of the benchmark datasets

Database Name	Size	Dimensions	No of Classes	Classes	Imbalanced distribution of instances	Type of Outliers present
PIMA Indian Diabetes	768	9	2	Yes	500	Point and
				No	178	Point and Collective
Education	155	10	2	Pass	138	Point and
				Fail	16	Point
Stock Market	293	5	5	Volume1	212	Collective
				Volume2	14	Point
				Volume3	24	Point
				Volume4	19	No
				Volume5	24	Point

3.2. Interesting Measures used to extract outliers

The interesting measure used to prune the infrequent patterns from other patterns is 'Support'. Support indicates the number of times the model has been found in the database. Support for the "A" model in the "D" database is computed as in equation 1.

$$Support\ (A) = \frac{P(A)}{|D|} \qquad (1)$$

The measures that are used to extract the exceptional rules, ie., the infrequent CARs are 'Rule Support' and 'Rule Confidence'. The exceptional CAR is of the form

A→C [Support%, Confidence%]

which implies that, the likelihood of occurrence of 'A' set of patterns in class 'C' of the database 'D'. The support and confidence values for the rule A→C are defined as in equations 2 and 3 respectively.

$$Support\ (A \to C) = \frac{P(A \Lambda C)}{|D|} \qquad (2)$$

$$Confidence\ (A \to C) = \frac{P(A \Lambda C)}{P(A)} \qquad (3)$$

Finally the interesting measure that is used to extract outliers from infrequent CAR is 'Lift'. "Rare item sets with low counts (low probability) which per chance occur a few times (or only once) together can produce enormous lift values" [7]. This nature – 'the enormous value' of the Lift measure motivated us to generate the rare item sets by making use of it. The lift value for the infrequent CAR, A→C is defined as in equation 4.

$$Lift\ (A \to C) = \frac{P(A \Lambda C)}{P(A)\ \Lambda\ P(C)} \qquad (4)$$

Outliers extracted by lift measurement from infrequently occurring CARs are assessed by the measurement called Coverage and Accuracy [19]. Let N_{COVERS} be the number of instances in the database 'D' covered by the LHS of exceptional CAR ie., the support(A) and $N_{CORRECT}$ be the number of instances correctly classified by the exceptional CAR. The Coverage and Accuracy of the rule is computed as

per the equation 5 and equation 6 respectively.

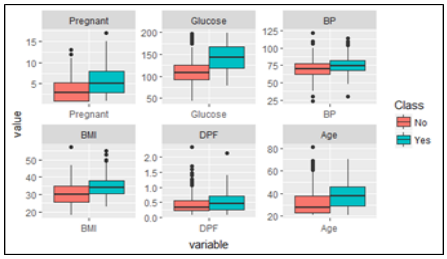
$$Coverage\,(A \rightarrow C) = \frac{N_{COVERS}}{|D|} \qquad (5)$$

$$Accuracy\,(A \rightarrow C) = \frac{N_{COVERS}}{N_{CORRECT}} \qquad (6)$$

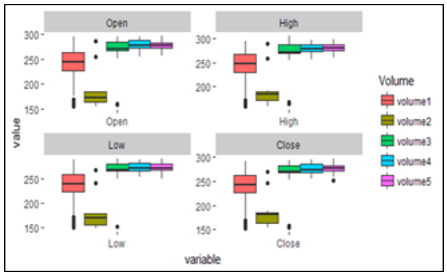
All the association technique based works discussed in the section 2 have used user defined thresholds for all the interesting measures to extract the rules. For assigning the pre-defined thresholds values for the measures, either prior knowledge about the data or domain experts' guidance is required. Also, threshold values are domain specific and vary for every application. Even the size of the data will cause change in the threshold values. Consequently, the proposed method generates threshold values for interesting measurements like Support and Lift dynamically [20] depending on the data distribution. By doing so it is observed that the threshold values of the measures vary for different databases.

3.3. Understand outliers and their type in benchmark data sets.

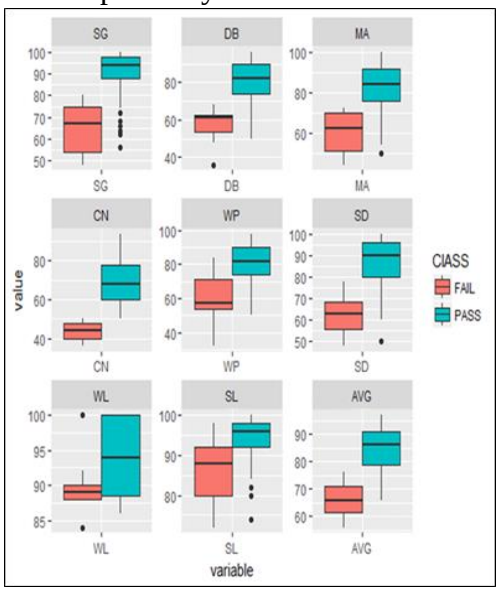
The box plot is the best statistical visualization technique for interpreting outlier presence [21]. Consequently, an analysis of the plots was conducted on the datasets to be assessed. The outlier type present in these data sets is analyzed using the box plot. Figure 2 presents the analytical report for the box plot. In Figure 2, the presence of bubbles in the upper and lower whiskers of the cell indicates the presence of outliers. The emergence of bubbles in a group or points can be interpreted as collective and global outliers respectively.



(a) Pima Indian Diabetes Data



(b) Stock Market Data



(c) Education Data

Figure 2. Analysis of outliers in the datasets examined for evaluation.

4. The Proposed Methodology

The proposed SSOD uses the ARM based associative classification approach for detecting outliers. The method dynamically computes the threshold value for the support measure to generate infrequent patterns. To prune outliers the surprising measures 'Lift' is used. The data set considered are all numeric in nature. Hence the numerical data are categorized before performing the mining process. This data constitutes the input of the SSOD algorithm. Candidate sets are generated, from which infrequent models are pruned by support measurement. The actual support of each candidate set is calculated as per the equation 1.

Then, the infrequent patterns are pruned using the dynamically computed support threshold as in the algorithm. Following the step in the algorithm, the higher level candidate sets are generated and the complete sets of infrequent patterns are pruned by the support threshold. After generating the complete infrequent pattern sets, the infrequent CARs are generated and the one with the high confidence are pruned as rare rules. Lastly, the surprising Lift measurement is used to prune outliers of the rare rules. The exceptional CARs with high confidence are alone considered to extract the outliers

---

Algorithm1: SSOD - Semi Supervised Outlier Detection

Input **:** Database - D, Initial Support - S

Output **:** A set of Temporal Outliers – Outlier

Notations Used **:** σ^{R}_{dynsup} - Max. Support, S_a - Actual Support, R - Rare Pattern sets, C_k: Candidate pattern set of size k, L_k : Infrequent pattern set of size k

Data Preprocessing: */* Categorization of the numerical values */*

1. D(A1, A2, ….An, CLASS) → D^{P}(A1.LOW, A1.MEDIUM, A1.HIGH, ….An.HIGH, CLASS)

Infrequent Pattern Generation: */* Infrequent pattern generation - APRIORI algorithm */*

2. C_1 = { A1.LOW, A1.MEDIUM, A1.HIGH, ….An.HIGH };
3. L_1 = Set of all C_1 for which S_a (C_1)< S; // Where S=0.2 the Initial $Support_{min}$
4. R= $L_{1;}$
5. for (k= 2; L_{k-1} !=∅; k++) do begin
6. C_k= candidate sets generated from L_{k-1} ;
7. L_k = Set of all C_k for which S_a (C_k)< σ^{R}_{dynsup}
8. R=R∪L_k;
9. end for;
10. return R;

Outlier Pruning: */* Pruning low support with high confidence infrequent CARs with outlier pruning measure */*

11. Outlier = Outlier_Pruning (R);

Post Pruning: /* Redundant Outliers if any are eliminated */

Algorithm2: Outlier_Pruning

Input : D^P – Preprocessed data, R - Rare Pattern sets

Output : A set outliers (exceptional CARs)

Notations Used : R_{Conf} – Rule Confidence, R_{sup} - Rule support, R_{lift} - Rule Lift, L_k - Infrequent pattern set of size k, IRs – Infrequent Rules, ER_{pruned} – Outlier, $Supp_{threshold}$ – Max. Support, $Lift_{threshold}$ – Min. Lift

Infrequent Rule Generation: */* Infrequent CARs from Infrequent Patterns set */*

1. IRs=∅;
2. for all L_k in R do
3. Scan D^P
4. If the instance of D^P matches L_k
5. Then Generate CAR of the form $L_k \rightarrow$CLASS; // Rule of the form A→C
   /* where $L_k \in$ (A1.LOW, A1.MEDIUM, A1.HIGH, ….An.HIGH) */
6. end if;
7. calculate R_{sup} , R_{Conf} and R_{Lift} for CAR;
8. IRs= IRs ∪ CAR ;
9. end for;
10. $Support_{threshold} = Mean(\sum R_{sup}(IRs))$; $Lift_{threshold} = Mean(\sum R_{Lift}(IRs))$;

Outlier Pruning: */* Pruning outliers ie, exceptional CARs from Infrequent CARs */*

11. $ER_{pruned} = \emptyset$;
12. for all rules in IRs do
13. $ER_{pruned} = ER_{pruned} \cup IR$ (with $R_{sup} \leq Supp_{threshold}$, $R_{Lift} \geq Lift_{threshold}$ and $R_{Conf} > 95\%$) ;
14. end for;
15. return ER_{pruned} ;

5. Results and Discussions

To evaluate the proposed method – SSOD, several experiments were carried out on the benchmarked datasets. The performance of SSOD along with the lift measures is evaluated based on various parameters like the time taken to detect outliers, heap space used to detect outliers, the number of outliers detected by the Lift measure from the infrequent patterns. Also, the nature of the outliers extracted was observed. The observed results are tabulated in Table 2. The exceptional CARs with high Lift value (>1) are considered as outliers. The threshold value of Lift measurements is calculated dynamically as indicated in the SSOD algorithm.

Table 2. Performance of the proposed method - SSOD

Evaluation Parameters	**Databases**		
	PIMA Indian Diabetes	**Education**	**Stock Market**
1. User defined thresholds	Initial Support =	Initial Support =0.2	Initial Support =0.2
2. Thresholds	Max Supp. = 0.372	Max Supp. = 0.196	Max sup = 0.291
3. No. of	25	36	94
4. No. of	12	17	2
5. Heap space for outlier	44,507,136	5,908,122	5,571,344
6. Outliers'	0.716	0.568	0.306
7. Outliers'	98.91	98.52	99.92

The results presented in Table 2 show that the number of infrequent models generated and the heap space used differ for each data set. This is because, each dataset vary in size and dimensions. Though the user defined initial support value is kept uniformly, the threshold value of Support and Lift measures computed dynamically varies for each dataset. Also, the number of infrequent CARs generated varies for each dataset varies. Likewise, the number of outliers extracted, the heap space used to extract outliers, and the time required to detect outliers differ for each dataset. The highest coverage value for PIMA Indian diabetes dataset indicates the presence of a greater number of outliers, which may be understood from the box plot analysis in Figure 2. Consistently across the dataset, the detection accuracy for outliers is almost 99%.

5.1. Analysis of the outliers generated

The natures of the outliers/Exceptional CARs generated for the datasets are examined and the inferences of the rules are interpreted. The sample outliers generated for Pima Indian Diabetes dataset, Education dataset and Stock market dataset and the implication of the exceptional CARs are listed in Table 3, 4 and 5 respectively. Based on the coverage value, the types of the outliers are identified. The high coverage value indicates the group / collective outliers and the low coverage values indicate the point outliers.

Table 3. Sample outliers detected from PIMA INDIAN diabetes dataset.

Exceptional CARs	**Support**	**Confidence**	**Lift**	**Coverage**	**Implications**	
					Risk Factors	**Outlier Type**
Age (Above 60) → No	0.003	1.000	1.019	0.214	Age factor	Point
DPF (Above 0.6) ^ BMI (26-30) → No	0.003	1.000	1.529	0.214	BMI level and hereditary factors	Point
DPF (Above 0.6) → No	0.131	0.924	1.396	14.11	Hereditary factors	Collective

Table 4. Sample outliers detected from EDUCATION dataset.

Exceptional CARs	**Support**	**Confidence**	**Lift**	**Coverage**	**Implications**	
					Risk Factors	**Outlier Type**
SG (80-89) → PASS	0.153	0.933	1.029	19.565	Higher marks in SG	Collective
SL (80-89) → PASS	0.196	1.000	1.106	14.304	Higher marks in SL	Collective
SD (50-59) ^ DB (60-69) → PASS	0.011	1.000	1.138	01.087	Border marks in SD	Point

Table 5. Sample outliers detected from STOCK MARKET dataset.

Exceptional CARs	**Support**	**Confidence**	**Lift**	**Coverage**	**Implications**	
					Risk Factors	**Outlier Type**
Close (210.33 -233.8) → Volume1	0.173	0.955	1.013	18.107	Close price is high	Collective
Low (177.08 -218.68) → Volume1	0.171	0.955	1.006	07.489	Low price is considerably low	Collective
Open (183.33 – 211.36) → Volume2	0.003	1.000	1.066	00.290	Open price is considerably low	Point
High (172.34 -183.36) → Volume2	0.008	1.000	1.025	00.246	High price is considerably low	Point

5.2. Performance of SSOD against conventional outlier detection methods

The performance of the proposed SSOD method is compared to existing classical outlier detection methods. The calibrated methods are Naive Bayes, Multi-class Classifier, K-Nearest Neighbor (KNN), Class outlier factor (COF), Distance based algorithm, Density based algorithm, and Local outlier factor (LOF). The performance of the above mentioned algorithms are evaluated on the basis of assessment measures such as Precision, Recall and F-Score. The results are plotted, as shown in Figures 3, 4 and 5 respectively.

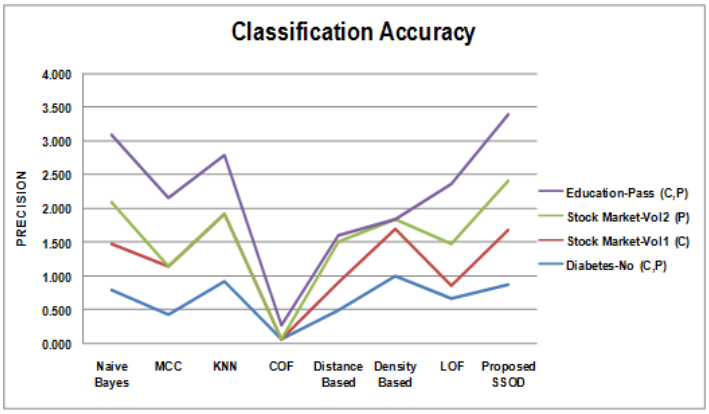


Figure 3. SSOD vs classical methods Accuracy of classification.

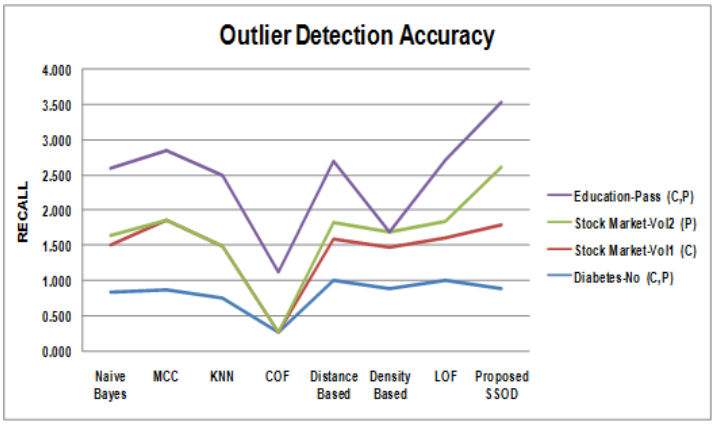


Figure 4. SSOD outlier detection accuracy compared to conventional methods.

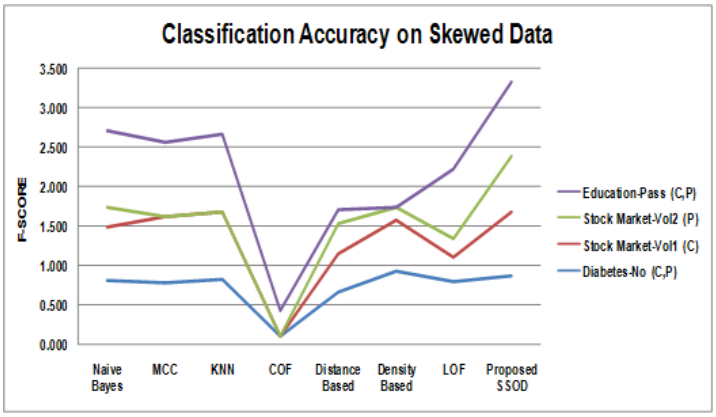


Figure 5. Efficiency of SSOD methods relative to conventional methods.

For simple understanding the alarming target classes with outliers are alone is considered to construct the graphs. For example, in the Education dataset, the 'Pass' target class with collective and point outliers is considered from which the risk factors identified will be useful for improving the results. From the graphs, it is understood that COF algorithm perform very poor in predicting both the collective and point outliers. Though the algorithms like KNN, Naive Bayes and Distance

based performs better in few aspects, the proposed SSOD surpasses all the algorithms in all the aspects. Hence, we claim that the proposed SSOD algorithm along with the surprising measure 'Lift' is the better choice of identifying both the type of outliers.

6. Conclusion and Future Work

An overview of type of outliers and their presence in the databases has been discussed. The need for the surprising measure in outlier detection is explained. The limitations of supervised and unsupervised ML algorithms in outlier detection are discussed and how the proposed SSOD algorithm overcomes these limitations is also explained. The surprising measure Lift along with SSOD algorithm have been experimented with the databases like Pima Indian Diabetes dataset, Education dataset and Stock market dataset where point and collective outliers are present. Based on the results obtained, it is understood that the lift measure is capable of detecting both the point and the collective outliers fruitfully. Also, the performance of the SSOD with Lift as the outlier pruning measure has been examined against the existing conventional outlier detection algorithms like Naive Bayes, Multi-class Classifier, K-nearest neighbor, Class outlier factor (COF), Distance based algorithm, Density based algorithm, Local outlier factor (LOF) and found that the proposed method – SSOD, outperforms the benchmarked algorithms by precisely detecting more accurate outliers. Therefore, we conclude that the surprising Measure Lift is adapted to extract collective and point outliers from different multidimensional, multi-labeled and asymmetric data sets.

For future research the work can be examined by various outlier pruning measures to improve the accuracy level. Also, the algorithm can be tested on Big data.

References

[1] C.C. Aggarwal. An introduction to outlier analysis. In Outlier Analysis, Springer, Cham, (2017). 1-34.

[2] G. Bruno, and P.Garza. TOD: Temporal outlier detection by using quasi-functional temporal dependencies. Data & Knowledge Engineering (2010), 69(6), 619-639.

[3] A.M. Rajeswari, C. Deisy, V. Abirami Nachammai, and G.V. Aishwarya. Temporal Outlier Detection on Quantitative Data using Unexpectedness Measure', 12th International Conference on Intelligent System Design and Applications, IEEE Proceedings, (2012), 420-424.

[4] H. Jin, J. Chen, H. He, G. J. Williams, Chris Kelman, and M. O. Christine Keefe. Mining Unexpected Temporal Associations: Applications in Detecting Adverse Drug Reactions. IEEE Transactions On Information Technology In Biomedicine, (2018), 12(4), 488-500.

[5] H. Jin, J. Chen, H. He, C. Kelman, D. McAullay, and C.M. O'Keefe. Signaling potential adverse drug reactions from administrative health databases. IEEE Transactions on knowledge and data engineering, (2010), 22(6) 839-853.

[6] Y. Ji, H. Ying, P. Dews, A. Mansour, J. Tran, R.E. Miller, and R.M. Massanari. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance, IEEE Transactions on Information Technology in Biomedicine, (2011), 15(3), 428-437.

[7] M. Hahsler. A probabilistic comparison of commonly used interest measures for association rules. (2015) Available online at http://michael. hahsler.

net/research/association_rules/measures. Html.

[8] A.M. Rajeswari, and C. Deisy. "Prediction of risk factors for pre-diabetes using a frequent pattern-based outlier detection." International Journal of Biomedical Engineering and Technology (2020), 34(2), 152-171.

[9] A.M.Rajeswari et al., A comparative evaluation of supervised and unsupervised methods for detecting outliers. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), (2018), 1068-1073, IEEE.

[10] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of ACM SIGMOD, Washington DC,USA. (1993), 207– 216.

[11] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases. (1994), 478–499.

[12] C. Romero, J. R. Romero, J. M. Luna and S. Ventura. Mining rare association rules from e-learning data. in Educational Data Mining. (2010), 171-180.

[13] S. Preetha, and V. Radha. Enhanced outlier detection method using association rule mining technique. International Journal of Computer Applications. (2012), 42 (7), 1-6.

[14] H. Deng, G. Runger, E. Tuv, and W. Bannister. CBC: An associative classifier with a small number of rules. Decision Support Systems. (2014), 59, 163-170.

[15] L.T. Nguyen, B. Vo, T.P. Hong, and H.C. Thanh. CAR-Miner: An efficient algorithm for mining class-association rules. Expert Systems with Applications. (2013), 40 (6), 2305-2311.

[16] J. Alcala-Fdez, R. Alcala, and F. Herrera. A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Transactions on Fuzzy Systems. (2011), 19 (5), 857-872.

[17] Pima Indian Diabetes database. (2022). Available online at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[18] State Bank of India (SBIN.NS). (2022) Available online at: https://in.finance.yahoo.com/quote/SBIN.NS/history

[19] J. Han, J. Pei, and M. Kamber. 2011, 'Data mining: concepts and techniques. Elsevier. (2011).

[20] C.K. Selvi, and A. Tamilarasi. Mining association rules with dynamic and collective support thresholds. International Journal of Engineering and Technology. (2009), 1 (3), 236-240.

[21] M. Krzywinski and N. Altman. Points of significance: visualizing samples with box plots. Nature methods. (2014), 11 (2), 119-120.