An Enhanced Classification Based Outlier Detection using Decision Tree for Multi class in Data Stream

R. Sangeetha, S. Sathappan

Abstract :Data Stream has continual, unbound, large and unstable records. The processing in data streams involves extracting significantidea in primary data of the kind static and dynamic with one sweep. In streaming, records are generated by thousands of data sources continuously and simultaneously. These data normally won't have common range. Some data will be deviated from the rest in terms of variant factors. These are considered as outliers and it is tough to find those in data stream as they have multi dimensionality. Outliers, being the most abnormal observations, may include the sample maximum or sample minimum. E-commerce is an application or category of data stream that is generated from millions of sources at a time. It includes multiple products and transactions. Some products are cancelled during transactions and some are infrequent and these are termed as outliers. This paper focus on the challenge in E-Commerce and the objective is to aid the agencies in taking fine decisions in right time by finding the outliers using supervised learning scheme. This work is carried out in two phases. In first phase the outliers are detected and classified as cancelled and delivered products. In the second phase the least transaction is found as an outlier by the enhanced methodology for multi-class classification. The work is implemented in WEKA 3.9.6 and is compared with the existing works with evaluation metrics.

Index terms: DataStream, Outlier, Decision Tree, Random Correction Code, Log loss.

I. INTRODUCTION

Data Stream referred as analyzing high speed enormous data with no limit. It has the feature of continual arrival, multiplicity, rapid, time variant, boundless, unpredictable streams that arise new research problems and challenges. Analyzing this streaming model turn the researchers into a significant attention in the past few years. [5] The challenge includes memory limitation, faster computing in just one pass etc. So the extraction of the hidden knowledge requires automated efficient techniques. However the classic data mining procedures won't fit for stream mining since it involves multiple pass over the data to explore the patterns.The pattern also keeps changing over the time so it should be recorded. As the volume of data is bulky in the case of streaming, the resource allocation is a big challenge. In order to minimize this issue some techniques such as sampling, sliding window are used while analyzing the data.

Revised Version Manuscript Received on Jun 20, 2019.

R.Sangeetha, Research Scholar,Department of Computer science,Erode arts and science college, Erode, Tamilnadu, India.

Dr.S.Sathappan, Associate Professor,Department of Computer science.,Erode Arts and science college, Erode, Tamilnadu, India

These techniques allow the users to fix the amount of data for analyzing at one time. Data preprocessing is also used in which the irrelevant attributes, duplicate instances are removed to reduce the file size and memory space as per the work.For evaluation, learning is parted as'Supervised, Unsupervisedand Semi supervised'.These categoriesexamine the trained set and carry the assessment on the test set.

A. Database Management System versus Data Stream Management System

Table I. Comparison of Database management and Data stream management system

Factors	**DBMS**	**DSMS**
Data access	Random access	Continuous access
Time Requirements	No real time services	Real time Requirements
Rate of updating	Low update rate	Possible multi GB update rate
Nature of Data	Persistent data	Volatile data streams
Query Nature	One time Queries	Continuous Queries
Storage Capacity	Unlimited secondary storage	Limited main memory
Data Type	Data at any granularity	Data at fine granularity

Table 1 shows the comparison of DBMS and DSMS with seven factors.

Database management system defined as a collection of procedures enable for storing, modifying and extracting information from a database. In contrary, data stream management has the real and continuous arrival of items with time stamp. It is unfeasible to control the order of arriving and storing the stream records. To get the part of the stream the concepts such as sliding window, sampling are used.

Features

- Generous amount of uninterruptedfacts.
- Requires rapid real time response.
- Need single scan algorithm.
- Stores only the extract of the record.
- Supportsmulti-dimensional processing.

Applications

- Telecommunication calling records,
- Credit card transactions flow,
- Stock Exchange,
- Weblogs and Webpage click streams,
- A real time stream of actual messages on twitter.
- Market orders or E-commerce,
- Network monitoring and traffic management.

B. Outliers

The dissimilar object which is highly deviated from all the available groupsnormally. [6] Andvaries from the normal one by having peculiar featureswhen compared with the rest. It has high distance with other observations and cannot have similar thing with the data. This detection plays a vital role in mining and aids in developing the accuracy, categorizing and grouping. Also used to find and remove uncommon.

Types

The four models are based on their existence in addition th
set may have more than one class.

1) Global or Point

An individual is diverged from the restand considered as th
simplest form.

For example, an entityO_{gb}has the dissimilarity from the others by considering any one subject.

2) Contextual or Conditional

Athingvaries with a particular context. This is relate
with the contextual and behavioural attributes.

For example, an itemO_{co}, deviates significantly based c
selected conditions.

3) Collective

A batch of related instances is separated with respect to entire set. It forms as a subset as a collectionof varied matter that deviates from the whole sets. This is applicable for graph, sequence& spatial categories. It resides mostly with the related instances where the data are bind with each other.

4) Real Outlier

These are the real outlying observations which are lies in the analysis of the system analyst. This does not mean the actual error or anonymous data rather they are the real outliers.

Categories

There are three kinds based on the distribution in the variables:

a. Uni-variate – Exceptions can be found while consideringa single feature space.

b. Bi-variate – Considers two features.

c. Multi-variate – Found in an m-dimensional space.

II. DECISION TREE CLASSIFIER

Classification is a technique that assigns classes to a bulk of data in order to give more precise analysis. It is intended to get the effective analysis of very large file. [9] Initially the outlier detection technique label the whole set as outliers and normal. Then the classification technique is put against the data to form categories. The classifier maps inputs to a class. In the dataset, the records are known as observation, the attributes are explanatory variables, and the possible categories are dependent variables or classes.

DecisionTree

In this structure, leaf symbolizes classes and branches denotes the combined features that lead to the class. Classification tree take a discrete set of values and regression tree take continued values. The determination is to create a layout that predicts the value of a target on the basis ofinpus. Each internal node denotes to one of the input variablesand leafhas the target variable.[10] The full attribute set is divided by partitioning the sourceto subsets based on an attribute test. This is recursive partitioning and is repeated on derived subsettoo. This procedure is ended during the subset node has all the same value.

The mathematical statement is:

$$(M, \mathrm{N}) = (\mathrm{m1}, \mathrm{m2}, \ldots, \mathrm{mk}, \mathrm{N})$$

--- (1)

Where,

(m1,m2,…mk) is the feature set,

N is the target variable.

Advantages

Most convenient to follow and comprehend.

Handle numbered and categorical data.

Lessening the time of data preparation.

Performs well with large datasets.

Robust against co-linearity.

Disadvantages

A small change in the training set deviates highly in the final predictions.

Less accuracy in multi-class classification.

Overfitting.

III. LITERATURE REVIEW

AnuregBeiju*et al*[1] improved the classical pricing strategies by collectingsource from E-Comm and apply methods to extract usefulinfo. The proffered scheme is generated by using trees to extract information regarding

buying attitude of a consumer using if-then rules. Thefacts aretaken from Flipkart. After applying the rules to find out the star rating, classification is applied in a set of objects to categorize them into different groups with similarity.Measures are used to assess the performance of Id3 algorithm with if-then rules.

Cao Lijun *et al* [2] putforth a new detection algorithm with reverse nearest-neighbors. Sliding window model is used and queries are performed for the identification of dissilmiarity in the current window. The insertion/deletion needssole scan which also improves performance. The queries are achieved by Query Manager that capture the concept drift. Experiments are conducted with real,unrealand results showsthe algorithm is effective and efficient.

Kurian. *et al* [7] evaluated the performance of outlier detection with feature selection algorithms. An outfit is an abnormal thingthat must be removed from the massive details. This work aims to control the variation by applying dimensionality reduction which has a less set of features to relate a large set with high dimensions. Since processing with smaller is much faster than largerit diminishes the time. The assessment of the decision tree algorithms were tested with the cancer database. The decision tree algorithm with feature selection algorithm gives high accuracy.

Mani Mehrotra *et al* [8] combined two variant methodsKmeanswith treesin which Euclidean is used to find the adjacent for the data set and then the model is built for each to classify each as normal and abnormal

This paper take 90% data for the training and the balance for the testing to compute theclosest cluster.The number of cluster is defined as twenty after having ten iterations in the set. Experiments are evaluated and show that in training there are six incorrect and five hundred and twenty correct classifications and in test data there are fifty six correct and two incorrect classifications.

Sanizahahmad*et al* [11] introduced a method for the detection of outsiders with regression. It is observed the unmatched data have a considerable influence on the analysis which lead the work to the erroneous conclusions. This paper, has four methods and compared with samples. The methods are 'Pearson residual, Pearson Standardize, Deviance and standardize Deviance residual'. From the analysis it is shown the Deviance and Standard deviances when combined with logistic regression have better assessment while finding outliers.

Victoria J. *et al* [13] evaluated algorithms for classification and outlier detection in temporal data. This work evaluates the algorithms that train and classify and used to incorporate new data regularly. It compares the accuracy of six data mining classification algorithms using a well-known time-series datasets for the detection of outliers.Decision tree performs well but over-fits some data sets. This paper finds outliers for Human activity Recognition dataset but the classification is less due to overfitting.

IV. METHODOLOGY

A. Existing work

The work [1] analyses the e-commerce files by assessing directly the consumers by the internet. The companies can benefitted with the collected information by identifing target consumers, introducinghighly movable products or services to meet the needs. After analyzing the training set, outlier was found in terms of least customer satisfaction based on online product rating. This online rating is given by applying If-Then rules based on the buying behavior of the customer. Finally the data are classified using ID3 algorithm based on the class label A, B, C, D that denotes 2, 3, 4 ,5 star rating respectively.

Methodology: ID3

A simple algorithm that creates a decision tree of given dataset with top-down greedy method to check each attribute. In this approach information gain approach is generally used to have the most suited attribute for splitting each node. So, entropy is calculated first to get the value. The variable with high information or least entropy generates sub tree with another node.

Entropy

It calculates the homogeneity of the instances. If the instances arefully homogeneous the entropy is zero and if it is equally divided it has entropy of one.

It is calculated by,

$$\text{Entropy (J)} = \sum_{j=1}^{n} - p_j \log_2 p_j$$

--- (2)

Here,

p_j is the probability of j.

Information-Gain

This is estimated on the basis of decreasing value in entropy measure to have a better split. The model is constructed with highest information gain. It determines the importance of each attribute in the feature set.

It is calculated by,

$$\text{Information Gain}(J, X) = \text{Entropy}(J) - \text{Entropy}(J|X)$$

--- (3)

Advantage

a. Easy to understand.
b. Gives highest accuracy with labeled data.

Disadvantage

a. Does not guarantee upon local optima.
b. Over fitting.
c. Harder to use on continuous data and missing value.
d. Less accuracy in multi-class classification.

B. Existing work 2

The work [7] did a classification based outlier analysis for outlier detectionwith breast cancer dataset. Feature selection algorithm Gini index is combined with tree to reduce the dimensionality. It consists of Benign and Malignant tumors. The Benign form a separate group and the rest is the outlier. Using decision tree C4.5 this work classifies the outliers which are distinct from the other.

Methodology : C4.5

It is a continuation of the ID3. A tree pruning algorithm is included with the base by using values of only one attribute at a time. It has the same feature for splitting the attribute with entropy and information gain. The modified extension includes tree pruning.

Tree pruning

Pruning decrease the size [4] of the design by removing the branches that give mini knowledge to assort the samples. It reduces the complexity, and thus upgrades the predictionby reducing over-fitting. This can be ofTop-down that traverse and trim by starting at the root and Bottom-up by starting at the leaf.

Reduced error Pruning

It works in bottom up post pruning fashion that every node is displaced with its affordable class. If the correctness is high then the change is preserved. This method gains popularity for its simplicity and speed.

The method is as,

If Parentof error <Child of error then 'Prune'
Else 'Don't Prune'.

Improvements or Advantage from the base ID3

a. Handle continuous value.
b. Handle training data with missing value.
c. Pruning trees after creation thus eliminates the problem of over fitting.

Disadvantages

a. Less accuracy with unlabelled data.
b. Problematic in attribute with large number of values.

c. Less accuracy in multi-class classification.

C. Proposed Methodology

Phase 1 – Filtered Classifier in Decision Tree (FDT)Outlier Detection and classification of Outliers, Inliers.

A new attribute named status is created by the filter and the values are filled automatically based on the attribute Quantity as it has the information of delivered and cancelled products.

Expression 1: Creation of new attribute 'Status' with filled values to indicate delivered and cancelled products.

$$\text{Status} = \sum_{k=1}^{p} (-\text{abs}(A4) * -1/A4)$$

--- (4)

Where,

p - number of sample products.

A4 - name of the attribute Quantity.

Classification of Outliers based on cancelled and delivered products -

The existing two methods use information gain for best split. The proposed method use Gain ratio measure as the best split for attribute splitting. It corrects the problem of information gain measure by taking the intrinsic information into account while splitting. Tree pruning use sub tree raising method which selects a subtree by replacing a child node.

Gain ratio:

$$\text{Gain Ratio}(Y, a) = \frac{\text{IG}(Y, a)}{\text{Intrinsic Info}\ (Y, a)}$$

--- (5)

Where, IG is the information gain,

Y is set of training observation,

'a'denotes attribute *Procedure FDT (Filter Decision Tree)*

Step 1: Split training and test set.

Step 2: Add expression with a filter to form a new attribute with values by (4) to calculate outliers.

Step 3: Calculate Entropy and information gain of every attribute.

Step 4: Find Gain ratio by (5).

Step 5: Form a decision tree with the attribute that has highest Gain ratio.

Step 6: Recur on subsets using remaining attributes.

Step 7: Prune the branch which does not lead to leaf nodes by sub tree raising in post pruning method.

Phase 2 - Filtered Classifier for Multi class in Decision Tree (FMDT), Removal of Outliers and Classification

The outliers are removed for further analysis so that the data set will contain only the delivered products. Further the dataset is classified based on the number of transaction per country to find out the least transaction country in the week as the outlier. Generally the e-commerce dataset may have numeroustransactions, the class may have multi values but decision tree couldn't handle as it gives high accuracy for binary class. To handle multi-class classification problem, this phase includes 'random correction approach' with the decision tree.

Expression 2: Removing the cancelled products which are considered as outliers.

$$\text{Class index} = \sum_{i=1}^{s} \text{if}(\text{status} < 0) \text{ then remove values.}$$

--- (6)

Where, 's' is the number of instances,

Status implies anattributes which are entered as positive and negative based on the cancelled and delivered products.Here, the negative valued instances are removed as they are cancelled products which are not useful for further analysis.

Classification of Outlier based on Transaction -

Enhanced Multi-class Classifier- Random error correction codes with log loss decoding

Error correction finds errors and recreates the original error-free data with binary coded. It is a category of [12] 'Error Correcting Output codes (ECOC)' where error detection allows detecting errors and while error correction reconstructs the original data in many cases. In this procedure, features are transformed to the set of defined category in where each class is noted by code words. The method improves the performance of the classification by inverting a multiclass into binary sub class and correcting errors while decision making.

Categories:

Exhaustive correction codes – Classes are less in number.

Random Correction code – Classes are large in number.

The proposed methodology use random correction code as the classes are not known in prior since it is a E-commerce data the classes are the name of the country and it varies day to day as per the transaction.

Random Correction code –

Let 'CW' be a codematrix having the size C* L.'C'points the number of class and L the length of code. All element is set to either one or zero randomly. The consideration while designing includes the length (L), distance between rows and columns. This distance is calculated using Hamming distance. Minimum Hamming distance determines error-correcting ability.

Example Codeword matrix:

Table II. Codeword Matrix

Class	Codeword
1	0 0 1 1 1 0 0 1 0 0 0 0 0 0 0
2	0 0 1 0 0 0 1 0 0 0 1 1 0 0 1
3	0 1 1 0 0 1 0 0 0 1 1 0 0 0 1
4	1 1 1 1 0 1 1 1 1 0 1 1 0 0 1
5	0 1 1 0 1 1 0 1 1 0 0 0 0 0 1
6	1 1 1 0 0 1 0 1 1 0 1 1 0 1 0
7	1 1 1 0 0 1 1 0 1 0 0 1 1 0 1
8	1 1 0 0 1 1 0 1 0 0 1 0 0 1 0

Table 2 shows the example codeword matrix for eight classes with the codeword length fifteen. The higher the length gives the optimum result. The 0,1 are assigned randomly for each class. Every class has unique code that differentiates each other.

Hamming distance –

The distant[15] between two binary data strings is the number of positions at which the corresponding character is different. It is used as a metric on the set of words of length.

For distance of two binary numbers, the procedure returns the XOR value for the count of set bits. The longer code words provide optimal code. The minimal distance is used to classify the data correctly. This hamming distance is mostly used in coding, information theory, and cryptography. This metric is used to find which strings are nearer and which are further away.

Procedure Hamming distance
Hamming distance (D1, D2)
If len(D1) != len(D2)
Count++;
return count.

Example
The Hamming distance between:

a. "karolina" and "kathrina" is 3.
b. "karolini" and "kerstini" is 3.
c. "1011100" and "1001000" is 2.
d. Log loss decoding –
e. Log loss function measures the [14] uncertainty of the probabilities of the model by comparing them with the true labels. It upgrades the performance of a classifier by minimizing false classification. A completely perfect model has a log loss of 0. In other words, minimizing the Loss valuewill maximizes the accuracy. For multi- class classificationa separate loss function for each class label per observation is calculated and the result is summed.
f. It is written as,
g. $\text{Logloss} = -\frac{1}{C}\sum_{i=1 \text{ to } C}(x_i (\log p_i) + (1 - x_i) \log (1 - p_i)$
h. --- (7)
i. Where,
j. x_i is the actual value of the class label.
k. C is the number of classes.
l. p_i is the predicted probability of the value regarding the class.
m. Procedure Filter Multi-class Decision Tree
n. Step 1: Remove the outlier using (6).
o. Step 2: Classify the data using random correction code
p. (Table 1.) and log loss decoding using (7).
q. Step 3:Form a decision tree with the attribute that has
r. highest Gain ratio.
s. Step 4: Recur on subsets using remaining attributes.
t. Step 5: Prune the branch which does not lead to leaf nodes
u. by sub tree raising in post pruning method.
v. Advantages of the proposed method
w. Detect the outlier efficiently with simple method.
x. Handle Multi-class classification.
y. Gives highest accuracy.

V. EXPERIMENTAL RESULTS

Data set –
The online retail dataset is mined from the UCI repository. The dataset has the information about one week online ordering details with eight attributes and 9974 instances in UK and is a registered online retail store. The store uniquely sells all occasional gifts. Many customers are wholesalers.The attributes are Invoice no., Stockcode, Name of the product, Ordered/Cancelled products, Date, Price of the product per unit, Customer Identity and Country.

Pre processing –
Correlation subset feature evaluation[3] with greedy search is carried out for in preprocessing to remove the irrelevant, redundant data to ease the process in mining.It increases the speed and accuracy of mining algorithm.
This is a similarity measures to gain information between two variables.
Correlation coefficient = ±1 – Linearly dependent.
Coefficient = 0 - Uncorrelated.
Five attributes are selected after preprocessing. The attributes are Stock code, Ordered/Cancelled Quantity, Unit price, Customer Identity, Country.

Phase 1 (A) Outlier Detection

Viewer

Relation: online retail with cancel csv-weka.filters.unsupervised.attr

No.	1: Scode	2: Uprice	3: Cusid	4: Quantity	5: Country	6: **Status**
	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal
...	2121...	0.55	1430...	-12.0	United...	-1
...	2121...	1.45	1430...	-12.0	United...	-1
...	2291...	4.95	1430...	8.0	United...	1
...	2225...	1.65	1430...	3.0	United...	1
...	1623...	0.21	1430...	3.0	United...	1
...	2271...	0.42	1430...	8.0	United...	1
...	2281...	1.95	1430...	36.0	United...	1
...	8434...	2.55	1430...	-12.0	United...	-1
...	2158...	2.55	1430...	25.0	United...	1

Figure 1. New attribute with filled values

The Figure 1shows the output screen of a new attribute notes as 'Status'. Its values are filled by -1 and 1 by the eq (4) based on the attribute quantity. The quantity attribute has the clue of delivered and cancelled products. This attribute separate those information and placed as nominal values for classification as the numeric values in the quantity attribute could not be classified.
Here, -1 indicates the cancelled products and 1 indicates the delivered products.

Phase 1 (B) Classification of outliers

```
== Confusion Matrix ==
   a    b   <-- classified as
1149    0 |    a = -1 -> Cancelled Products
   0 8825 |    b = 1  -> Delivered Products
```

Figure 2. Classification based on cancelled and delivered products

The figure 2 shows the confusion matrix in which the data are classified as cancelled and delivered products. The cancelled products are the outliers which is to be removed.

Phase 2 (A) Removal of outliers

Viewer

Relation: online retail with cancel csv-weka.filters.unsupervised.attri

No.	1: Scode	2: Uprice	3: Cusid	4: Quantity	5: Country	6: **Status**
	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal
...	2291...	4.95	1430...	8.0	United...	1
...	2225...	1.65	1430...	3.0	United...	1
...	1623...	0.21	1430...	3.0	United...	1
...	2271...	0.42	1430...	8.0	United...	1
...	2281...	1.95	1430...	36.0	United...	1
...	2158...	2.55	1430...	25.0	United...	1

Figure 3. Outlier Removal

Figure 3 shows the output screen of removing the outliers(-1 value)by using eq (6) which indicates cancelled products. They are removed for further analysis. Out of 9974 instances 1149 products are removed and are considered as outliers.

Phase 2 (B) classification based on number of transactions done by the country after outlier removal

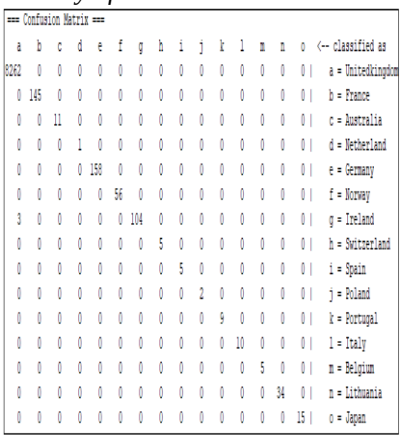

Figure 4. Transaction based classification

Figure 4 shows the output screen of transaction based classification. The diagonal element shows the classified instances based on the transaction done by variant countries on a particular week. In the above classification the least transaction is done by five countries which have minimumfivetransactions on that week and is considered as outlier as compared with other countries.

Performance Analysis

A. Comparison of existing and proposed method

Table III. Performance analysis

Metrics	Existing method 1 (ID3)	Existing method 2 (C4.5)	Proposedmethod Phase 1Phase 2 (FDT)(FMDT)	
Precision	87.72	96.64	100	99.99
F-Score	87.23	96.34	100	99.83
Sensitivity	86.92	96.12	100	98.99
Specificity	85.12	94.65	100	98.72
Error rate	13.52	4.75	0.00	3.01
Accuracy	86.47	95.25	100	99.96

Table 3 shows the performance analysis of existing and proposed method with six metrics. The analysis shows the proposed method 1 and proposed 2 out performs the existing methods. *B. Performance Chart*

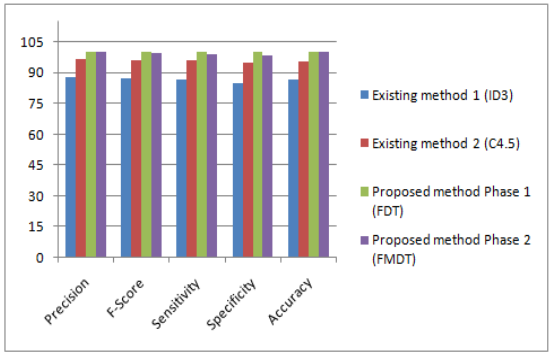

Figure 5. Performance analysis

Figure 5 shows the performance analysis of existing and proposed method with six metrics. The analysis shows the proposed method 1 and proposed 2 out performs the existing methods.

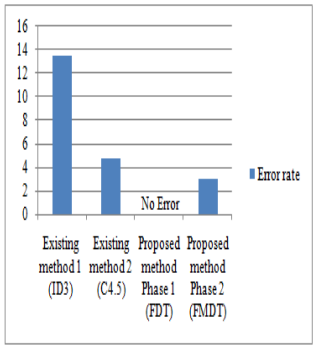

Figure 6. Error rate

Figure 6 shows the error rate for existing and proposed method. The analysis shows the proposed method 1 has no error as it is naturally a binary classification and proposed 2 less error than the existing as per the enhancement done for multi-class classification.

V. CONCLUSION AND FUTURE WORK

Data stream has a continuous, rapid stream of information. E-Commerce is a category of data stream where the products are sold in online. So it has highly varying entries for each countries and finding the pattern among them is a tedious job. Outliers are the entries which deviate from the rest by having peculiar pattern. The E-commerce set normally combined with delivered and cancelled products. This research work analyses the set and separates the outliers with the filtered technique. Then it classifies the data based on the country transaction with an enhancement to handle multi-class classification to find out the least transaction country. The work is implemented in WEKA 3.9.6 and is compared with the existing works with evaluation metrics.

In future, the work can be extended by analysing other attributes and can be employed unsupervised technique also.

REFERENCES

1. Anurag Bejju, " Sales Analysis of E-CommerceWebsites using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 133, No.5, January 2016.
2. Cao Lijun 1 , Liu Xiyin 2 , Zhou Tiejun 1 , Zhang Zhongping 3, Liu Aiyong, "A Data Stream Outlier Delection Algorithm Based On Reverse K Nearest Neighbors", 2010 International Symposium on Computational Intelligence and Design IEEE.
3. DASH, M., & LIU, H (1997) Feature selection for classification. Intelligent Data Analysis, 131- 156.
4. Ding Xiang-wu and Wang Bin, "An Improved Pre-pruning Algorithm Based on ID3," JisuanjiYuxiandaihua,Vol.9, pp. 47,2008.
5. Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining: concepts and techniques", Morgankaufmann, 2006.
6. Kurian M.J and Gladston Raj S,(2015) " Outlier Detection in Multidimensional Cancer Data Using Classification Based approach " International Journal of Applied Engineering Research ,Vol.10, No.79,pp. 342-348 , 2015.
7. Kurian M.J., Gladston Raj S., PhD, " An Analysis on the Performance of a Classification based Outlier Detection System using Feature Selection", International Journal of Computer Applications", Volume 132 – No.8, December2015 .
8. Mani Mehrotra1, Nakul Joshi, "Anomaly Detection in Temporal data Using Kmeans Clustering with C5.0", The International Journal of Engineering and Science (IJES) Volume 6, Issue 5, PP 77-81, 2017.
9. Quinlan J. R. (1986). "Induction of decision trees. Machine Learning," Vol.1-1, pp. 81-106.
10. Rutkowski, L. Pietruczuk, P. Duda, and M. Jaworski,z"Decision Trees for Mining Data Streams Based on the McDiarmid's Bound," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 6, pp. 1272-1279, 2013.
11. Sanizahahmed, Norazon Mohamed Ramli, Habshah midi, , "Outlier detection in logistic regression and its application in medical data analysis", in: 2012 IEEE colloquium on humanities, science and engineering.
12. Somkidamornsamankul, jairajpromrak, pawalaikraipeerapun, "solving multiclass classification problems using combining complementary neural networks and error-correcting output codes", international journal of mathematics and computers in simulation, issue 3, volume 5, 2011.
13. Victoria J. Hodge and Jim Austin, "An Evaluation of Classification and Outlier Detection Algorithms", archive.org, 2018.
14. https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation.
15. https://en.wikipedia.org/wiki/Hamming_distance.

AUTHORS PROFILE

R.Sangeetha received the MCA degree from Bharathidasan University, India in the year 2007 and M.Phil degree in Computer Science from Bharathiar University, India in the year 2013 respectively. Currently, she is a Part-Time Ph.D., Research Scholar of Computer Science, Erode Arts and Science College, Erode, affiliated to Bharathiar University. She also worked as an Assistant Professor with the total experience of 8 years. She has published one papers in International IEEE Conferences and one paper in UGC Journal. Her area of interest is Data Mining. Email: geethasanmca@gmail.com

Dr.S.Sathappan received the M.Sc., degree in Applied Mathematics from Bharathidasan University, India in the year 1984, the MPhil and PhD degrees in Computer Science from Bharathiar University, India in the year 1996 and 2012 respectively. Currently, he is an Associate Professor of Computer Science, Erode Arts and Science College, Erode, Affiliated to Bharathiar University, Coimbatore. He served as a Syndicate Member of Bharathiar University from June 2015 to June 2018. He also served as a Senate Member of Bharathiar University2005-2008 and 2015-2018. He has been a supervisor for several students of MCA/MPhil., programs over the past several years. Also, he has guided 8 Ph.D Scholars and guiding 4 Ph.D Scholars . He has a total experience of over 33 years. He has published 46 papers in Conferences and 46 papers in Journal. His areas of interest include computer simulation, image processing and data mining. Email: devisathappan@gmail.com