DOI: 10.21917/ijsc.2021.0338

PERFORMANCE ANALYSIS OF ANOMALOUS DETECTION SCEHMES BASED ON MODIFIED SUPPORT VECTOR MACHINE AND ENHANCED RELEVANCE VECTOR MACHINE

S. Senthil Kumar[1] and S. Mythili[2]

[1]Department of Computer Science, Kongunadu Arts and Science College, India
[2]Department of Information Technology, Kongunadu Arts and Science College, India

Abstract

Anomalous transactions are common activity happening on the financial oriented transaction. Detecting those anomalous transactions from the financial transaction patterns is the most complex task which is focused in this work. In the existing work it is achieved by introducing the method namely Fuzzy Exception and Fuzzy Anomalous Rule (FEFAR). The accuracy of this existing work FEFAR found to be lesser which is resolved in the proposed research work. There are two research works has been proposed those are namely Rule Pruning based Anomalous Rule Detection Strategy (RPARD) and Lasso Regression based Improved Anomalous Detection Scheme (LR-IADS). Both of these methods attempt to find the anomalous transaction from the given input database by finding the anomalous rules. Each method differ in its methodologies, thus the accuracy of the methods would differ. The main goal of this analysis work is to compare the performance of existing and proposed methodologies based on simulation outcome. This research work aims to highlight the performance variation between the proposed and existing techniques and the best method that can offer accurate anomalous transaction detection. The analysis of the research work is carried out on matlab environment over four databases namely soil, bank, german statlog and auto mpg based on which performance outcome has been given.

Keywords:

Anomalous Transaction, Anomalous Rules, Accuracy, RPARD, LR-IADS, FEFAR

1. INTRODUCTION

The goal of data mining is to discover inherent and previously unknown information from data. When the knowledge discovered is in the form of association rules, the methodology is called association rule mining. An association rule de- scribes a relationship among different attributes. In association rule mining, large number of association rules or patterns or knowledge is generated from the large volume of dataset. But most of the association rules have redundant information and thus all of them cannot be used directly for an application. In order to apply the mining algorithm to various problems, the quantitative attributes should also be appropriately dealt with as well as the Boolean attributes. Especially in manufacturing area, quantitative attributes such as states of a process, conditions of manufacturing, and measured quality of products, are necessary for quality control [1], manufacturing management, planning and decision of management strategy. In order to deal with the quantitative attributes in mining association rules, algorithms based on the generalized association rules that handle the continuous attributes as the Boolean vector by partitioning into several intervals are proposed [2],[3]. Fuzzy association rules mining approaches are proposed to overcome such disadvantages based on the fuzzy set concept [4]-[9]. These approaches are based on fuzzy extensions of the classical association rules mining by defining support and confidence of the fuzzy rule.

The fuzzy association rules mining has a good property in terms of quantization of numerical attributes in database compared with Boolean quantized generalized association rules mining. Though the mining results of fuzzy association rules are easy to understand for corresponding human operators, two drawbacks are still remain for applying such fuzzy approaches to the actual problems. One is the computational time for mining from database, and the other is huge redundant rules extraction as the result of mining. The issue of computational efficiency is critical for fuzzy association rules mining compared with traditional Boolean rules mining, as the number of fuzzy items for mining increases for quantization of numerical attributes.

Exception rules were first defined as rules that contradict the user's common belief [10]. In other words, for searching an exception rule we have to find an attribute that changes the consequent of a strong rule. The direct technique of mining exception rules are in most of the cases highly subjective as the set of user's beliefs is compared to the set of mined rules. The indirect techniques use the knowledge provided by a set of rules (usually strong rules) and then the exception rules are those that contradict or deviate this knowledge. Anomalous rules were first presented in [11] as a set of rules that are in appearance similar to exception rules, but semantically different. An anomalous association rule is an association rule that comes to the surface when we eliminate the dominant effect produced by a strong rule. In other words, it is an association rule that is verified when a common rule fails. The knowledge provided by the exception and the anomalous rules are (semantically) complementary. The authors in [12] analyzed in detail about semantics and formulation of approaches for mining exception and anomalous rules that are defined in terms of association rules.

Rule Pruning based Anomalous Rule Detection method (RPARD) is a technique in which primarily relevant attribute selection performed based on metrics called Gini index and information gain. Lasso, an anomaly detection method, convert the process of anomaly detection into a linear regression model. The detection parameters are used as regression arguments, and the Lasso method is used to establish a parameter model of regression arguments and dependent variables. In anomaly detection, arguments correspond to measurable event parameters, while dependent variables correspond to the classification results of measurable event parameters. They constitute the training data sets for learning model. The key problem of detecting abnormal events is to estimate the consistency between detected events and the parameter model established by Lasso. Lasso Regression-based Detection Scheme (LRDS) approach is introduced to achieve optimal detection in the proposed datasets. Based on the

aforementioned discussions, in this paper, we are interested to give a comparative study on fuzzy exception and fuzzy anomalous rules, rule pruning based anomalous rule detection and Lasso regression-based detection scheme. Firstly, the aim of the research work is to use infrequent rules to detect normal and anomalous patterns automatically. This could be used for obtaining the common customer behavior (association rules) as well as the anomalous deviations (anomalous rules). In order to extract the normal and anomalous patterns in form of fuzzy rules, a new approach that takes advantage of fuzzy techniques for mining rules is proposed. Secondly, the attributes chosen from the database will be previously produced fuzzy anomalous rule. The selection of attributes is predicated on the Gini index, the gain of information and the gain ratio. Then rule pruning has been expected to perform by applying the lasso regression analysis process of creating anomalous rules.

2. RELATED WORKS

In [13], the authors dealt with an improvement of algorithm for extracting fuzzy association rules from a database and also improved the computational efficiency of data mining to reduce the redundant rules extracted for an actual application. They introduced the redundancy of fuzzy association rules and propose an essential algorithm based on the Apriori algorithm for rule extraction considering equivalence redundancy of fuzzy items. Exception rules were first defined by Suzuki in [14]. For extracting them we need a set of two rules noted by (csr,exc) where csr stands for common sense rule which is equivalent to impose that the rule is frequent and confident; and exc represents the exception rule associated to the csr. An easy example is given in [15] and explained in a simple way. In [16] a good overview of methods for exception discovery is done including exception instance discovery, exception rule discovery and exception structured-rules discovery. In [17] the authors categorize exception rules into eleven categories and then propose a unified algorithm for discovering all of them based on their formalization in rules triples (csr, exc, ref) where ref stands for reference rule.

In [15], the authors introduced the notion of fuzzy exception and fuzzy anomalous rule for the recognition of these types of deviations. The deviations are associated to the common pat- terns which usually are hidden in data affected by some kind of fuzziness. A new approach for mining such rules based on a recently proposed model for representing and evaluating fuzzy rules is presented in [15]. Important advantages are obtained more understandable results and that the mining process can be parallelized. An algorithm of the given model is developed and some experiments are performed in data where some numerical attributes have been fuzzified and also in some real fuzzy transactional datasets for testing the algorithm.

Zhang et al [18] suggested a comprehensive evaluation method incorporating all conventional data, like individual socio-demographic information and information on loan applications, as well as data on the dynamic transaction behaviour of the applicant. The reported technique is predicated on Multiple Instance Learning (MIL) Radial Basis Function (RBF), which produces characteristics from the transaction behaviour history of a person. To validate the effectiveness of our suggested solution, five real-world datasets from two major commercial banks in China are used.

Senthil Kumar and Mythili [19] addressed exception laws and anomaly identified in the literature study. This report gives an overview of anomaly detection studies. A wide variety of tools are available to detect anomalies, exceptions or anomalies: most of them are expert systems, neural networks, clustering methods and association rules. Senthil Kumar and Mythili [20] suggested an Anomalous Rule Detection Strategy (RPARD) focused on Rule Pruning for the detection of fuzzy anomalous rules. The implemented research method's fundamental dedication would be to introduce the structure that can accurately recognize the fake Visa exercises. Through displaying the system called the Stepwise Regression (SWR) test, preclude pruning is transmitted based on those selected values.

Similarly, the discovery of peculiarity is accomplished by methods to use fuzzy special case rules. Contingent on the criteria selected; order is made through methods for the Modified Support Vector Machine (MSVM) methodology over the last assumption test. Due to the difficulty of selecting high dimensional variables in anomaly detection model, an anomaly detection method based on Lasso is presented in [21]. Moreover, smoothly clipped absolute deviation penalty (SCAD) function is added as a constraint term which guarantees the accuracy of lasso solution. Further, experiments are carried out for three data sets, and two sets of evaluating indicator have been used to discuss and analyze proposed method and the results showed that anomaly detection method based Lasso can execute parameter estimation quickly and regression fitting accurately.

3. ANALYSIS OF ANOMALOUS DETECTION TECHNIQUES

In this research work, comparison analysis of the proposed and existing techniques has been carried out. The techniques that are utilized in this work for the comparison analysis are FEFAR, RPARD and LR-IADS. The discussion of these techniques is given in the subsection.

3.1 FUZZY EXCEPTION AND FUZZY ANOMALOUS RULE

Nowadays searching for specific kind of knowledge that deviates from the usual standards is very useful in several domain: network traffic anomalies, anomalous detection, economic analysis or medical diagnosis. Fuzzy association rules have been developed as a powerful tool for dealing with imprecision in databases (that may come from the source, i.e. imprecise measures taken by the machine, or from the human understanding of a concept) and offering a comprehensive representation of found knowledge. In this work fuzzy exception and fuzzy anomalous rules has been introduced for mining such rules based on a recently proposed model for representing and evaluating fuzzy rules. Important advantages are to obtain more understandable results and that the mining process can be parallelized. Authors proposed an algorithm which is a variation of Apriori, modifying it for dealing with a set of items represented by means of BitSets.

This bit-string representation has the advantage of speeding up logical operations such as conjunction or cardinality, which is a fundamental part of our algorithm when computing the frequencies of the Table.4(f). Beside this, the algorithm has also been accordingly adapted for the Apriori philosophy in order to consider the two aspects:

1. For each csr we add additional steps for discovering the exception and anomalous rules
2. The algorithm is executed independently for each level. Furthermore, step 3 complements step 2 to obtain for each discovered rule a summary measure of the accuracy values given in each level.

In this work the java class java.util.BitSet is utilized which contains the implementation of the object BitSet and some useful operations. The BitSet object stores a set of bits (zero or one) in each position. The algorithm processes the database and transforms it into a set of vector of BitSets with size equal to the number of transactions and dimension equal to the number of items. We then obtain a vector of BitSets for each level. For each transaction, a bit in the BitSet takes the value 1 if an item (or itemset when dealing with conjunction of items) is satisfied, or 0 if not.

3.2 RULE PRUNING BASED ANOMALOUS RULE DETECTION STRATEGY

Rule Pruning based Anomalous Rule Detection method (RPARD) is presented in the proposed technique in which primarily relevant attribute selection is performed based on metrics called Gini index and information gain. The Entropy-based discretization is utilized amongst the supervised discretization approach that uses the class info entropy of candidate partitions for choosing the limits for discretization. Class information entropy is an amount of purity and it gauges the volume of information that is required to state to which class an instance have its place. The main contribution of the proposed research method is to introduce the framework which can accurately identify the credit card anomalous activities. Depending upon those chosen attributes, rule pruning is carried out by presenting the technique called Stepwise Regression (SWR) analysis. In addition, by means of utilizing fuzzy exception rules, anomaly detection is carried out. In our proposed research method, for the given input credit card dataset, 100 rules are generated totally. These rules would consists of both relevant and irrelevant rules which needs to be filtered and resulted with the optimal rule set so that better outcome i.e., credit card transaction anomaly can be obtained. Thus, these generated rules are applied with rule pruning technique to avoid the irrelevant rules. Depending upon those chosen rules, classification is carried out for the final prediction outcome by means of Modified Support Vector Machine (MSVM) method.

3.3 LASSO REGRESSION BASED IMPROVED ANOMALOUS DETECTION SCHEME

Lasso Regression-based Improved Anomalous Detection Scheme (LR-IADS) approach is introduced to achieve optimal anomalous detection in this research. The pre-processing area was designed to minimize the number of items to also be evaluated by the algorithm for the association rule extraction and find the items that make up the database's local knowledge. The pre-processing analysis usually uses clustering algorithms in the literature to find local data, bringing together related knowledge. The association rule extraction algorithms will thus consider products with low support and the influencing items will also be broken into all classes and be less dominant. In this work, the attributes chosen from the database will be previously produced fuzzy anomalous rule. The selection of attributes is predicated on the Gini index, the gain of information and the gain ratio. Then rule pruning has been expected to perform by applying the lasso regression analysis process of creating anomalous rules. In this work, from the produced 100 rules, 75 rules are pruned. Eventually, anomalous identification is carried out on the basis of these anomalous rules by initiating the classification process that is carried out using the Association Classifier based on Enhanced Relevant Vector Machine.

4. RESULTS AND DISCUSSION

The performance evaluation of the research work is carried out in the matlab simulation environment under different metrics. In this section discussion of results obtained for both existing and proposed methodologies are given and then numerical evaluation of the research work is carried out. The performance analyses of the proposed and existing methodologies are given in the sub sections. Initially database details that are utilized in this work for analysis has been given.

4.1 DATASET DESCRIPTION

In this work four datasets has been considered for the performance evaluation. Those are soil database, german database, auto-mpg database, and bank database. The details of this datasets are given below:

- **Soil Dataset:** In this work, soil details of soybean plant are considered for the analysis. Here are 19 classes, only the first 15 of which have been used in prior work. The folklore seems to be that the last four classes are unjustified by the data since they have so few examples. There are 35 categorical attributes, some nominal and some ordered. The value “dna" means does not apply. The values for attributes are encoded numerically, with the first value encoded as “0," the second as “1," and so forth. An unknown values is encoded as “?". Dataset has 541 samples and 31 attributes
- **German Dataset:** This dataset classifies people described by a set of attributes as good or bad credit risks. Comes in two formats (one all numeric). Also comes with a cost matrix. Two datasets are provided. The original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes and is in the file “german.data”.

For algorithms that need numerical attributes, Strathclyde University produced the file “german.data-numeric”. This file has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables. Several attributes that are ordered categorical (such as attribute 17) have been coded as integer. This was the form used by StatLog. Dataset has 1000 samples and 23 attributes

- **Auto-MPG Dataset:** This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition. This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute.

The original dataset is available in the file "auto-mpg.data-original". "The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." Dataset has 398 samples and 7 attributes.

- **Bank Dataset:** The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. There are four datasets:

1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g. SVM). The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable *y*). Dataset has 4522 samples and 16 attributes

5. RESULTS ANALYSIS

In this section analysis of the proposed and existing research work is carried out based on each and every step of the proposed algorithm. The performance analysis of the existing and proposed techniques are shown in the pictorial representation which are gathered from matlab.

5.1 DISCRETIZATION

Discretization process is performed for partitioning the numerical data into intervals with crisp boundaries. This usually leads to an over estimation or under estimation of the boundary values. In this work entropy based discretization is utilized for finding boundary values. The Entropy-based discretization [21] is used among the managed discretization approach that utilizes the class information entropy of competitor allotments for picking the cutoff points for discretization.

Class data entropy is a measure of virtue and it checks the volume of data that is required to state to which class an occurrence have its place. It takes one major interim containing every known estimation of an element and after that recursively parcels this interim into minor subintervals till certain halting standard, for example Least Description Length (MDL) Principle or a perfect measure of interims is achieved accordingly delivering numerous interims of highlight.

The calculation of Entropy-based Discretization with Inconsistency Checking (EDIC) comprises of two stages. In the primary expression, an underlying discretization plot including a fundamental cut focuses set and a hopeful cut focuses set is discovered utilizing an entropy-based methodology.

5.1.1 Attribute Selection based On Gini Index, Gain Ratio and Information Gain:

Attribute selection is the process of choosing the attributes from the given input database with the concern of improvising overall accuracy. The selected attributes should be capable of generating the proper rules, thus the anomalous patterns can be detected accurately. In this work attribute selection is performed based on three metrics. Those are gini index, information gain and gain ratio. Here existing work FEFAR and proposed work RPARD considers only GINI index and information for the attribute selection whereas gain ratio is also considered by the LR-IADS method.

- ***Gini Index*:** The Gini index estimates the imbalance among estimations of a recurrence conveyance. It is guesstimated as

(1)

Here $p(i)$ is known as the attributes and i is the number of attributes

- **Information Gain:** The Information gain is known as the volume of data that is picked up by perceiving the estimation of the trait that is the entropy of the circulation in advance the split less the entropy of the conveyance after it. The greatest data gain is equivalent to the least entropy.
- **Gain ratio:** Gain ratio considers the result of the number of tuples in addition to the total number of tuples in D for each potential outcome. The ratio of gain is known as

(2)

In the Table.1 to Table.8, Gini index, information gain and gain ratio values measured for the existing and proposed techniques for the four databases has been given. Also, selected attribute names based on those metrics values are depicted.

Table.1. Gini Index, Information Gain and Gain Ratio Measurement for Soil Dataset

Feature number or attribute number	FEFAR		RPARD		LR-IADS		
	Gini Index	Information Gain	Gini Index	Information Gain	Gini Index	Information Gain	Gain Ratio
Date	0.4270	0.0247	0.3036	0.0247	0.3036	0.0630	0.0805
Plant-stand	0.9382	0.0265	0.3662	0.0265	0.3662	0.0706	0.0645
Precip	0.6354	0.0283	0.4253	0.0283	0.4253	0.0817	0.0450
Temp	0.6613	0.0285	0.4270	0.0285	0.4270	0.0823	0.0485
Hail	0.8688	0.0293	0.4468	0.0293	0.4468	0.0624	0.0582
Crop-hist	0.4253	0.0413	0.4960	0.0413	0.4960	0.0547	0.0848
Area-damaged	0.5989	0.0490	0.5172	0.0490	0.5172	0.0558	0.0625
Severity	0.7848	0.0534	0.5404	0.0534	0.5404	0.0639	0.0683
Seed-tmt	0.6438	0.0547	0.5989	0.0547	0.5989	0.0789	0.0474
Plant-growth	0.7556	0.0556	0.6354	0.0556	0.6354	0.0413	0.1164
Leaves	0.6536	0.0558	0.6438	0.0558	0.6438	0.0795	0.0575
Leafspots-halo	0.7710	0.0565	0.6524	0.0565	0.6524	0.0720	0.0451
Leafpots-marg	0.3036	0.0603	0.6536	0.0603	0.6536	0.0699	0.0657
Leafspot-size	0.7038	0.0619	0.6613	0.0619	0.6613	0.0680	0.0689
Leaf-shread	0.9121	0.0624	0.7038	0.0624	0.7038	0.0741	0.0520
Leaf-malf	0.8894	0.0630	0.7104	0.0630	0.7104	0.0534	0.0872
Leaf-mild	0.8877	0.0638	0.7149	0.0638	0.7149	0.0671	0.0410
Stem	0.3662	0.0639	0.7412	0.0639	0.7412	0.0619	0.0429
Lodging	0.9481	0.0653	0.7556	0.0653	0.7556	0.0653	0.0617
Stem-cankers	0.7794	0.0671	0.7573	0.0671	0.7573	0.0283	0.0348
Canker-leison	0.4960	0.0671	0.7710	0.0671	0.7710	0.0265	0.0266
Fruiting-bodies	0.5404	0.0680	0.7794	0.0680	0.7794	0.0490	0.0923
External decay	0.7149	0.0684	0.7848	0.0684	0.7848	0.0684	0.0593
Mycelium	0.6524	0.0699	0.8688	0.0699	0.8688	0.0671	0.0485
Int-discolor	0.7104	0.0706	0.8877	0.0706	0.8877	0.0247	0.0258
Sclerotia	0.7412	0.0720	0.8894	0.0720	0.8894	0.0556	0.0476
Fruit-pods	0.4468	0.0741	0.9121	0.0741	0.9121	0.0565	0.0956
Fruit spots	0.9396	0.0789	0.9382	0.0789	0.9382	0.0638	0.0806
Seed	0.7573	0.0795	0.9396	0.0795	0.9396	0.0603	0.0512
Mold-growth	0.5172	0.0817	0.9481	0.0817	0.9481	0.0293	0.0595
Seed-discolor	0.1254	0.0823	NaN	0.0823	NaN	0.0285	0.0127

Table.2. Selected Attributes based on Gini Index, Information Gain and Gain Ratio for Soil Dataset

FEFAR		RPARD		LR-IADS	
Attribute number	Attribute name	Attribute number	Attribute name	Attribute number	Attribute name
25	Int-discolor	13	Leafspot-marg	13	Leafspot-marg
21	Canker-leison	18	Stem	18	Stem
20	Stem-cankers	6	Crop-hist	6	Crop-hist
31	Seed-discolor	1	Date	1	Date
30	Mord-growth	27	Fruit-pods	27	Fruit-pods
10	Plant-growth	21	Canker-leison	21	Canker-leison
22	Fruiting-bodies	30	Mold-growth	30	Mold-growth
16	Leaf-malf	22	Fruiting-bodies	22	Fruiting-bodies

6	Crop-hist	7	Area-damaged	7	Area-damaged
26	Sclerotia	9	Seed-tmt	24	Mycelium
7	Area-damages	24	Mycelium	14	Leafspot-size
27	Fruit-pods	11	Leaves	25	Int-disoclor
29	Seed	14	Leafspot-size	23	External decay
18	Stem	25	Int-discolor	26	Sclerotia
5	Hail	23	External decay	10	Plant-growth
1	Date	26	Sclerotia	29	Seed
28	Fruit spots	10	Plant-growth	12	Leafspots-halo
8	Severity	29	Seed	20	Stem-cankers
19	Lodging	12	Leafspot-halo	8	Severity
17	Leaf-mild	20	Stem-cankers	5	Hail
24	Mycelium	8	Severity	17	Leaf-mild
14	Leafspot-size	5	Hail	16	Leaf-malf
23	External decay	17	Leaf-mild	15	Leaf-shread
13	Leafspot-marg	16	Leaf-malf	2	Plant-shread
2	Plant-stand	15	Leaf-shread	28	Fruit spots
12	Leafspots-halo	2	Plant-stand	19	Lodging
15	Leaf-shread	28	Fruit spots	31	Seed-discolor
9	Seed-tmt	19	Lodging		
		31	Seed-discolor		

Table.3. Gini Index, Information Gain and Gain Ratio Measurement for German Dataset

Feature number or attribute number	**FEFAR**		**RPARD**		**LR-IADS**		
	Gini Index	**Information Gain**	**Gini Index**	**Information Gain**	**Gini Index**	**Information Gain**	**Gain ratio**
Status of existing checking account	0.1163	0.0111	0	0.0375	0	0.0861	0.0470
Duration in month	0.1671	0.0122	0.0889	0.0440	0.0889	0.1603	0.1049
Credit history	0.15155	0.0143	0.0987	0.0561	0.0987	0.1824	0.1092
Purpose	0.1622	0.0151	0.1128	0.0610	0.1128	0.0771	0.1240
Credit amount	0.1263	0.0164	0.1333	0.0688	0.1333	0.1788	0.0987
Savings account/ bonds	0.1559	0.0172	0.1371	0.0750	0.1371	0.1744	0.0865
Present employment since	0.1580	0.0173	0.1377	0.0753	0.1377	0.0815	0.0568
Installment rate in percentage of disposable income	0.1617	0.0174	0.1395	0.0753	0.1395	0.0859	0.0467
Personal status and sex	0.1693	0.0174	0.1398	0.0764	0.1398	0.1173	0.0609
Other debtors/ guarantors	0.1533	0.0176	0.1402	0.0769	0.1402	0.1140	0.0853
Presence residence since	0.1568	0.0178	0.1431	0.0771	0.1431	0.0440	0.0493
Property	0.1555	0.0178	0.1432	0.0782	0.1436	0.0784	0.0719
Age in years	0.1609	0.0242	0.1459	0.0782	0.1459	0.0764	0.1322
Other instalment plance	0.1574	0.0242	0.1471	0.0784	0.1471	0.0688	0.0698
Housing	0.0870	0.0245	0.1478	0.0815	0.1478	0.0782	0.3017
Number of existing credits at this bank	0.1707	0.0255	0.1496	0.0859	0.1496	0.0753	0.1081
Job	0.1183	0.0270	0.1511	0.0861	0.1511	0.0769	0.1515
Number of people being liable to provide maintenance for	0.1615	0.0290	0.1513	0.1140	0.1513	0.0375	0.0800
Telephone	0.1711	0.0302	0.1529	0.1173	0.1529	0.0782	0.3017

Foreign worker	0.1712	0.0405	0.1544	0.1603	0.1544	0.0753	0.1081
land	0.1529	0.0486	0.1558	0.1744	0.1558	0.0561	0.0655
Debit amount	0.1575	0.0501	01591	0.1788	0.1591	0.0750	0.1039
Debit history	0.1606	0.0516	0.1668	0.1824	0.1668	0.0610	0.0643

Table.4. Selected Attributes based on Gini Index, Information Gain and Gain Ratio for German Dataset

FEFAR		**RPARD**		**LR-IADS**	
Attribute number	**Attribute name**	**Attribute number**	**Attribute name**	**Attribute number**	**Attribute name**
18	Number of people being liable to provide maintenance for	15	Housing	15	Housing
11	Presence residence since	1	Status of existing checking account	17	Job
21	Land	17	Job	1	Status of existing checking account
23	Debit history	5	Credit amount	20	Foreign worker
14	Other instalment plans	21	Land	7	Present employment since
22	Debit amount	10	Other debtors/guarantors	21	Land
16	Number of existing credits at this bank	12	Property	22	Debit amount
20	Foreign worker	11	Present residence since	13	Age in years
13	Age in years	14	Other instalment plan	23	Debit history
17	Job	22	Debit amount	18	Number of people being liable to provide maintenance for
4	Purpose	7	Presence employment since	11	Present residence since
15	Housing	23	Debit history	4	Purpose
19	Telephone	13	Age in years	8	Instalment rate in percentage of disposable income
12	Property	18	Number of people being liable to provide maintenance for	12	Property
7	Presence employment since	8	Instalment rate in percentage of disposable income	10	Other debtors/guarantors
8	Instalment rate in percentage of disposable income	4	Purpose	14	Other instalment plans
1	Status of existing checking account	2	Duration in month	16	Number of existing credits at this bank
10	Other debtors/guarantors	9	Personal status and sex	19	Telephone
9	Personal status and Sex	16	Number of existing credits at this bank	9	Personal status and sex
2	Duration in month	19	Telephone		
6	Savings account/bonds	20	Foreign worker		
5	Credit amount				
3	Credit history				

Table.5. Gini Index, Information Gain and Gain Ratio Measurement for Auto mpg Dataset

Feature number or attribute number	**FEFAR**		**RPARD**		**LR-IADS**		
	Gini Index	**Information Gain**	**Gini Index**	**Information Gain**	**Gini Index**	**Information Gain**	**Gain Ratio**
Mpg	0.2507	0	0.1896	0	0.2507	0.0461	0.1744
Cylinders	0.1896	0	0.2313	0	0.1896	0.0636	0.0400

Displacement	0.2485	0	0.2485	0	0.2485	0.0282	0.1761
Horsepower	0.2511	0.0282	0.2507	0.0282	0.2511	0.0440	0.1781
Weight	0.2313	0.0440	0.2509	0.0440	0.2313	0	0
Acceleration	0.2514	0.0461	0.2511	0.0461	0.2514	0	0
Model year	0.2509	0.0636	0.2514	0.0636	0.2509	0	0

Table.6. Selected Attributes based on Gini Index, Information Gain and Gain Ratio for Auto mpg Dataset

FEFAR		**RPARD**		**LR-IADS**	
Attribute number	**Attribute name**	**Attribute number**	**Attribute name**	**Attribute number**	**Attribute name**
5	Weight	5	Weight	5	Weight
6	Acceleration	3	Displacement	7	Model year
7	Model year	7	Model year	6	Acceleration
3	Displacement	4	Horse power		
4	Horse power	6	Acceleration		

Table.7. Gini Index, Information Gain and Gain Ratio Measurement for Bank Dataset

Feature number or attribute number	**FEFAR**		**RPARD**		**LR-IADS**		
	Gini Index	**Information Gain**	**Gini Index**	**Information Gain**	**Gini Index**	**Information Gain**	**Gain ratio**
Age	0.8840	0	0.8715	0	0.8715	0.0061	0.0099
Job	0.8893	0	0.8775	0	0.8775	0.0058	0.0064
Marital	0.8844	0	0.8825	0	0.8825	0.0025	0.0019
Education	0.8841	0	0.8830	0	0.8830	0.0064	0.0040
Default	0.8846	0	0.8840	0	0.8840	0	0
Balance	0.8825	0	0.8841	0	0.8841	0.0045	0.0094
Housing	0.9032	0.0025	0.8841	0	0.8841	0	0
Loan	0.9032	0.0025	0.8844	0.0025	0.8844	0	0
Contact	0.9014	0.0025	0.8846	0.0025	0.8846	0	0
Day	0.8715	0.0036	0.8846	0.0036	0.8846	0.0025	0.0021
Month	0.8860	0.0045	0.8860	0.0045	0.8860	0.0057	0.0148
Duration	0.8873	0.0057	0.8863	0.0057	0.8863	0.0060	0.0112
Campaign	0.8846	0.0058	0.8873	0.0058	0.8873	0.0036	0.1117
Pdays	0.8948	0.0060	0.8893	0.0060	0.8893	0.0065	0.0034
Previous	0.8775	0.0061	0.8948	0.0061	0.8948	0	0
Poutcome	0.8830	0.0064	0.9014	0.0064	0.9014	0	0
Deposit	0.8863	0.0065	0.9032	0.0065	0.9032	0	0

Table.8. Selected Attributes based on Gini Index, Information Gain and Gain Ratio for Bank Dataset

FEFAR		**RPARD**		**LR-IADS**	
Attribute number	**Attribute name**	**Attribute number**	**Attribute name**	**Attribute number**	**Attribute name**
5	Default	9	Contact	9	Contact
7	Housing	14	Pdays	14	Pdays
8	Loan	6	Balance	6	Balance
14	Pdays	15	Previous	15	Previous
15	Previous	1	Age	3	Marital
16	Poutcome	3	Marital	5	Default
3	Marital	5	Default	12	Duration
9	Contact	12	Duration	10	Day

12	Duration	10	Day	16	Poutcome
6	Balance	16	Poutcome	2	Job
10	Day	11	Month	8	Loan
2	Job	2	Job	7	Housing
11	Month	8	Loan		
1	Age	7	Housing		
4	Education				
13	Campaign				

In the Table.1, information, Gini index and gain ratio values measured for the proposed and existing methodologies are shown. Based on these values, selected attributed are listed in the Table.2.

In Table.2, selected attributes of the soil dataset by using FEFAR, RPARD and LR-IADS methods has been shown. Existing method FEFAR selected 28 attributes from the 31 number of attributes. The proposed method RPARD selects 29 attributes from the 31 number of attributes. And proposed method LR-IADS selects 27 attributes from the 31 number of attributes.

In the Table.3, Gini index, information gain and gain ratio values for the german dataset has been shown.

In the Table.4, selected attribute numbers and their names has been listed. These selected attributes are shown for the FEFAR, RPARD and LR-IADS methods for the german dataset has been listed. In Table.4, selected attributes of the german dataset by using FEFAR, RPARD and LR-IADS methods has been shown. Existing method FEFAR selects all attributes from the 23 number of attributes. Proposed method RPARD selects 20 attributes from the 23 number of attributes. And proposed method LR-IADS selects 19 attributes from the 23 number of attributes.

In Table.5, Gini index, information gain and gain ratio values for the auto mpg dataset has been shown.

In Table.6, selected attribute numbers and their names has been listed. These selected attributes are shown for the FEFAR, RPARD and LR-IADS methods for the auto mpg dataset has been listed. In Table.6, selected attributes of the auto mpg dataset by using FEFAR, RPARD and LR-IADS methods has been shown. Existing method FEFAR selects 5 attributes from the 7 number of attributes. Proposed method RPARD selects 5 attributes from the 7 number of attributes. And proposed method LR-IADS selects 3 attributes from the 7 number of attributes.

In Table.7, Gini index, information gain and gain ratio values for the bank dataset has been shown.

In Table.8, selected attribute numbers and their names has been listed. These selected attributes are shown for the FEFAR, RPARD and LR-IADS methods for the bank dataset has been listed. In Table.8, selected attributes of the german dataset by using FEFAR, RPARD and LR-IADS methods has been shown. Existing method FEFAR selects all attributes from the 16 number of attributes. Proposed method RPARD selects 14 attributes from the 16 number of attributes. And proposed method LR-IADS selects 12 attributes from the 16 number of attributes.

5.2 NUMERICAL ANALYSIS

In this section analysis of the proposed and existing research techniques in terms different performance metrics has been shown. The simulation values obtained for methodologies namely FEFAR, RPARD and LR-IADS for the four datasets has been tabulated and compared with each other based on analysis outcome. The performance metrics that are considered in this work for the comparison analysis are accuracy, precision, recall, f-measure, error rate and number of rules.

In the Table.9, simulation values for the accuracy metric for the methodologies FEFAR, RPARD and LR-IADS for the four datasets has been shown.

Table.9. Accuracy Metric Values

Dataset	**Accuracy (Converted into 100%)**		
	FEFAR	**RPARD**	**LR-IADS**
Soil	56	76	68.5185
Auto MPG	72	92	94.9367
German	72	64	72
Bank	52	76	52
Average	63	77	71.8638

In Table.9, comparison analysis of the proposed and existing research techniques namely FEFAR, RPARD and LR-IADS for the four datasets soil, auto mpg, german and bank has been given. Based on average outcome of four datasets, it is learned that proposed method RPARD tends to have higher accuracy rate than the FEFAR and LR-IADS. RPARD has 5.1362 % higher accuracy than LR-IADS and 14% higher accuracy than FEFAR.

In the Table.10, simulation values for the precision metric for the methodologies FEFAR, RPARD and LR-IADS for the four datasets has been shown.

Table.10. Precision Metric Values

Dataset	**Precision (Converted into 100%)**		
	FEFAR	**RPARD**	**LR-IADS**
Soil	42.5	73.3333	34.2593
Auto MPG	85.4167	53.8095	88.7168
German	24	57.5397	24
Bank	53.8462	57.1429	53.8462
Average	51.44073	60.45635	50.20558

In Table.10, comparison analysis of the proposed and existing research techniques namely FEFAR, RPARD and LR-IADS for the four datasets soil, auto mpg, german and bank has been given. Based on average outcome of four datasets, it is learned that proposed method RPARD tends to have higher precision rate than the FEFAR and LR-IADS. RPARD has 10.25078% higher

precision than LR-IADS and 9.015625% higher precision than FEFAR.

In the Table.11, simulation values for the recall metric for the methodologies FEFAR, RPARD and LR-IADS for the four datasets has been shown.

Table.11. Recall Metric Values

Dataset	Recall (Converted into 100%)		
	FEFAR	RPARD	LR-IADS
Soil	44.4853	75.7353	50
Auto MPG	56.25	66.6667	90.9524
German	33.3333	56.9853	33.3333
Bank	75	87.5	75
Average	52.26715	71.72183	62.32143

In Table.11, comparison analysis of the proposed and existing research techniques namely FEFAR, RPARD and LR-IADS for the four datasets soil, auto mpg, german and bank has been given. Based on average outcome of four datasets, it is learned that proposed method RPARD tends to have higher recall rate than the FEFAR and LR-IADS. RPARD has 9.4004% higher recall than LR-IADS and 19.45468% higher recall than FEFAR. In the Table.12, simulation values for the f-measure metric for the methodologies FEFAR, RPARD and LR-IADS for the four datasets has been shown.

Table.12. F-Measure Metric Values

Dataset	F-Measure (Converted into 100%)		
	FEFAR	RPARD	LR-IADS
Soil	42.8274	73.9583	40.6593
Auto MPG	52.5745	59.3567	89.5116
German	27.9070	57.1429	27.9070
Bank	40.4762	55.3571	40.4768

Table.13. Error Rate Values

Dataset	Error rate		
	FEFAR	RPARD	LR-IADS
Soil	44	24	31.4815
Auto MPG	28	8	5.0633
German	28	36	72.0930
Bank	48	24	59.5238
Average	37	23	42.0404

In Table.12, comparison analysis of the proposed and existing research techniques namely FEFAR, RPARD and LR-IADS for the four datasets soil, auto mpg, german and bank has been given. Based on average outcome of four datasets, it is learned that proposed method RPARD tends to have higher f-measure than the FEFAR and LR-IADS. RPARD has 11.81508% higher f-measure than LR-IADS and 20.50748% higher f-measure than FEFAR.

In the Table.13, simulation values for the error rate for the methodologies FEFAR, RPARD and LR-IADS for the four datasets has been shown.

In Table.13, comparison analysis of the proposed and existing research techniques namely FEFAR, RPARD and LR-IADS for the four datasets soil, auto mpg, german and bank has been given. Based on average outcome of four datasets, it is learned that proposed method RPARD tends to have lesser error rate than the FEFAR and LR-IADS. RPARD has 19.0404% lesser error rate than LR-IADS and 14% lesser error rate than FEFAR.

In the Table.14, simulation values for the number of rules for the methodologies FEFAR, RPARD and LR-IADS for the four datasets has been shown.

Table.14. Number of Rules Values

Dataset	Number of rules		
	FEFAR	RPARD	LR-IADS
Soil	98	93	58
Auto MPG	67	82	94
German	75	42	75
Bank	62	73	62
Average	75.5	72.5	72.25

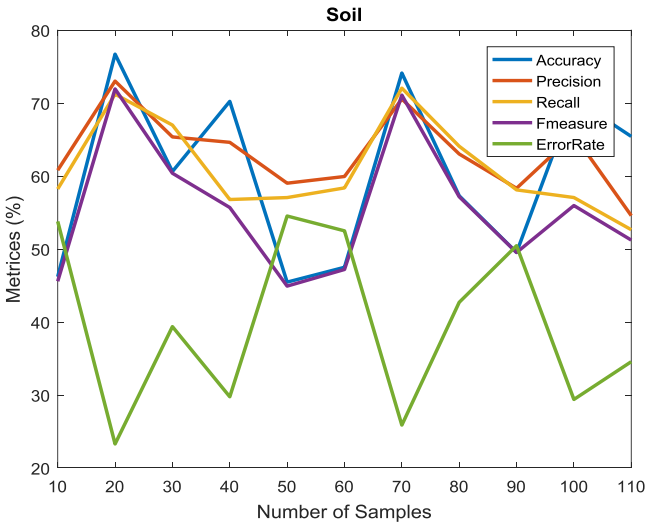
In Table.14, comparison analysis of the proposed and existing research techniques namely FEFAR, RPARD and LR-IADS for the four datasets soil, auto mpg, german and bank has been given. Based on average outcome of four datasets, it is learned that proposed method LR-IADS tends to select lesser number of rules than the FEFAR and RPARD. LR-IADS selects 0.344828% lesser number of rules than RPARD and 4.304636% lesser number of rules than FEFAR.

5.3 GRAPHICAL COMPARISON

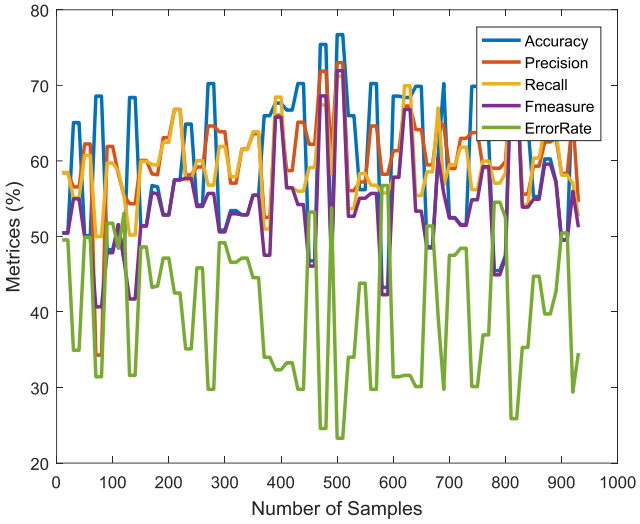
In this section graphical comparison of each performance metric that are discussed in previous section has been shown. Here performance analysis for all performance metrics namely accuracy, precision, recall, f-measure and error rate is shown for the all three methods FEFAR. RPARD and LR-IADS.

5.3.1 Soil Dataset Comparison:

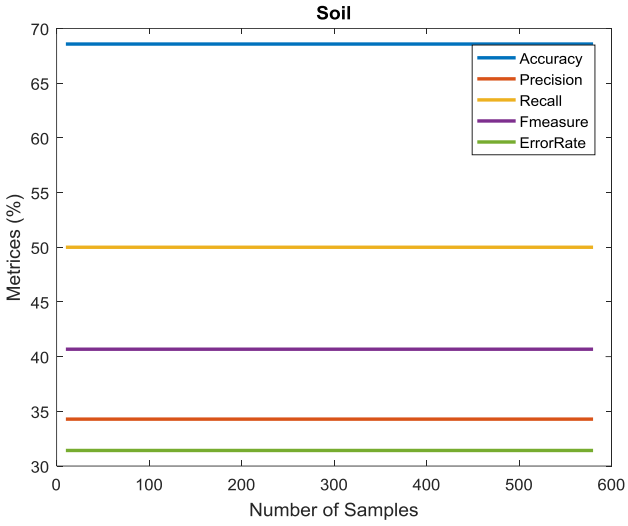
In Fig.1, performance analysis of the methodologies FEFAR, RPARD, and LR-IADS is shown for the soil dataset.



(a) FEFAR



(b) RPARD



(c) LR-IADS

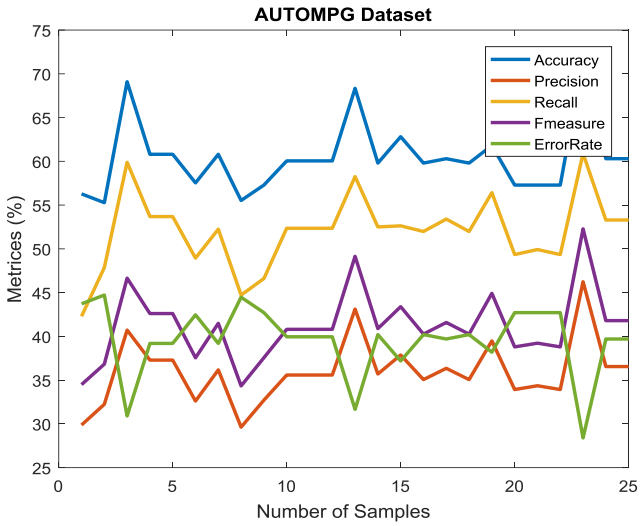
Fig.1. Performance Analysis of Soil Dataset

In terms of accuracy, RPARD seems to have higher accuracy than the FEFAR and LR-IADS where it is 7.4815% higher than LR-IADS and 20% higher than FEFAR method. In terms of precision, RPARD seems to have higher precision where it is 39.074% higher than LR-IADS and 30.8333% higher than FEFAR. In terms of recall, RPARD seems to have higher recall where it is 25.7353% higher than LR-IADS and 31.25% higher than FEFAR.

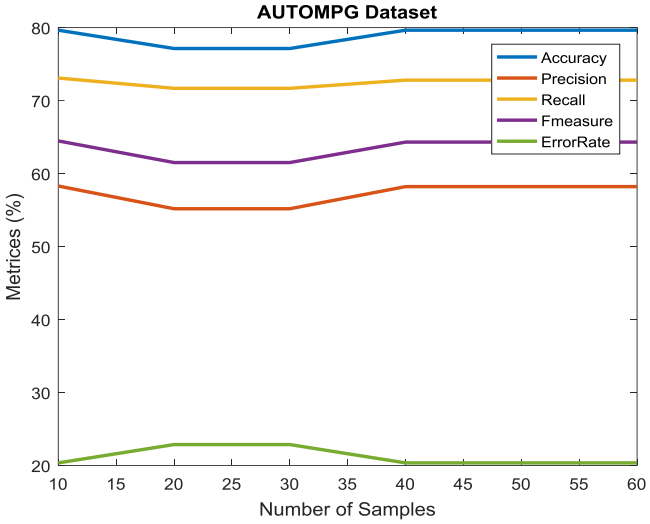
In terms of F-measure, RPARD seems to have higher f-measure where it is 33.2996% higher than LR-IADS and 31.1309% higher than FEFAR. In terms of error rate, RPARD seems to have lesser error rate where it is 7.4815% lesser than LR-IADS and 20% lesser than FEFAR.

5.3.2 Auto MPG:

In Fig.2, performance analysis of the methodologies FEFAR, RPARD, and LR-IADS is shown for the auto mpg dataset. In terms of accuracy, LR-IADS seems to have higher accuracy than the FEFAR and RPARD where it is 2.9367% higher than RPARD and 22.9367% higher than FEFAR method. In terms of precision, LR-IADS seems to have higher precision where it is 34.9073% higher than RPARD and 3.3001% higher than FEFAR.



(a) FEFAR



(b) RPARD

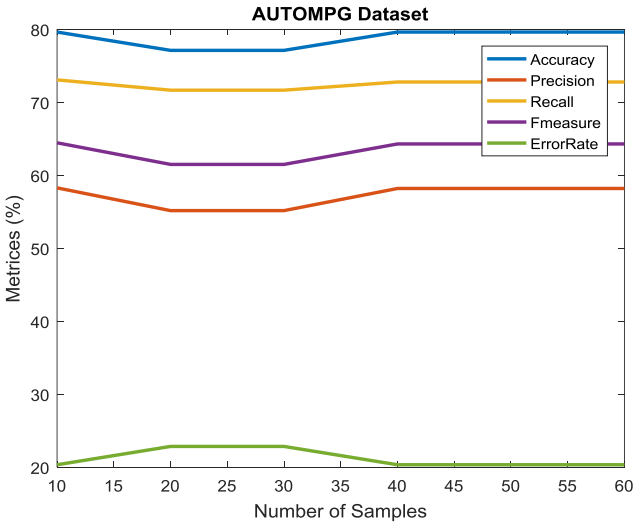


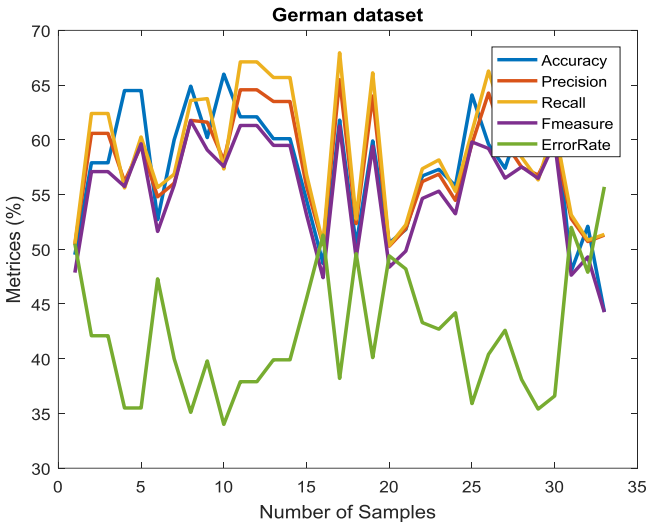
Fig.2(c). LR-IADS

Fig.2. Performance Analysis of Auto mpg Dataset

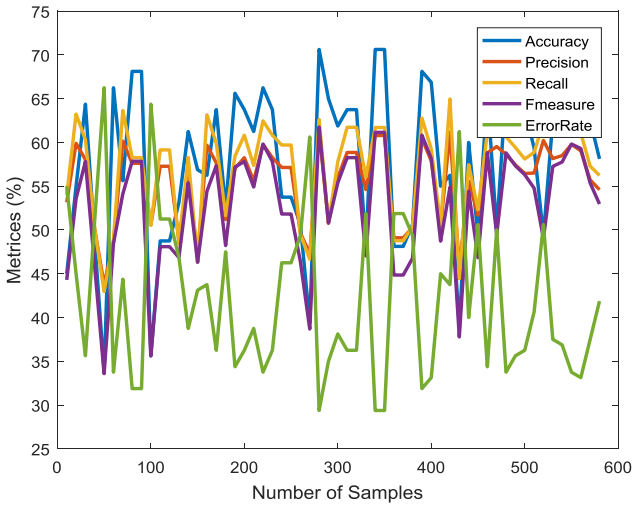
In terms of recall, LR-IADS seems to have higher recall where it is 24.2857% higher than RPARD and 34.7024% higher than FEFAR. In terms of F-measure, LR-IADS seems to have higher f-measure where it is 30.1549% higher than RPARD and 36.9371% higher than FEFAR. In terms of error rate, LR-IADS seems to have lesser error rate where it is 2.9367% lesser than RPARD and 22.9367% lesser than FEFAR.

5.3.3 German:

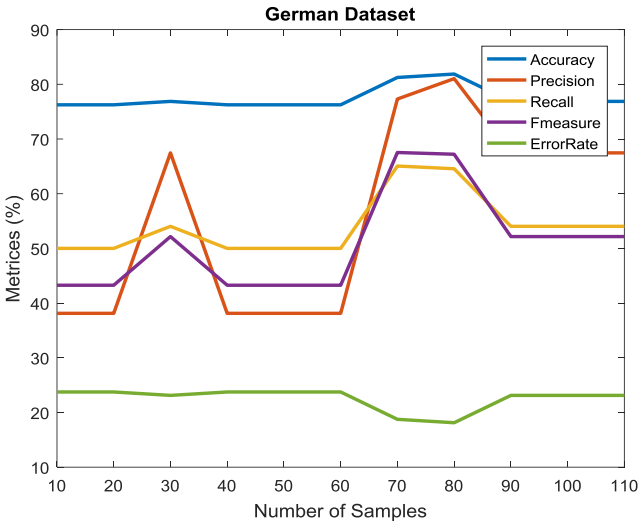
In Fig.3, performance analysis of the methodologies FEFAR, RPARD, and LR-IADS is shown for the german dataset. In terms of accuracy, LR-IADS and FEFAR seems to have higher accuracy which is 8% higher than RPARD method. In terms of precision, LR-IADS and FEFAR seems to have similar and lesser precision where it is 33.5397% lesser than RPARD. In terms of recall, LR-IADS and FEFAR seems to have similar and lesser recall where it is 23.652% lesser than RPARD. In terms of F-measure, LR-IADS and FEFAR seems to have similar and lesser f-measure where it is 29.2359% lesser than RPARD. In terms of error rate, FEFAR seems to have lesser error rate where it is 44.093% lesser than LR-IADS and 8% lesser than RPARD.



(a) FEFAR



(b) RPARD

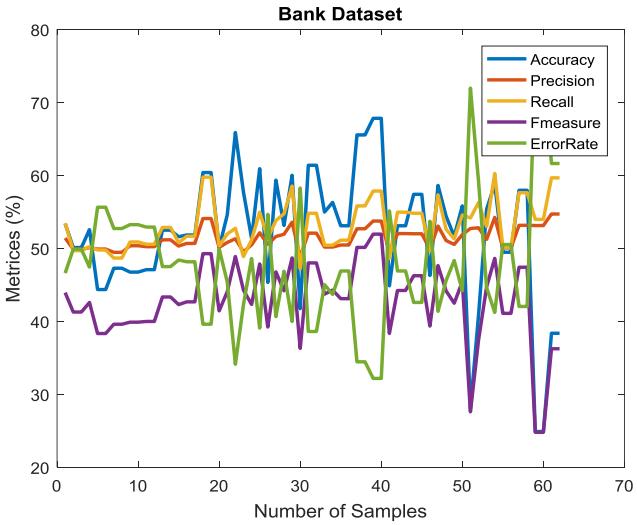


(c) LR-IADS

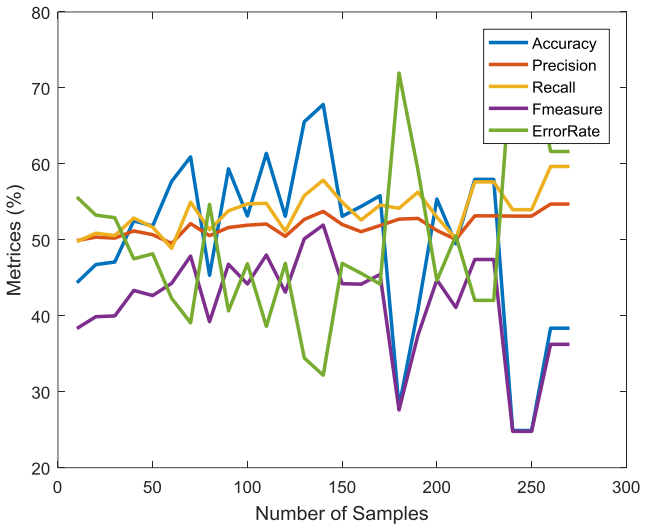
Fig.3. Performance Analysis of German Dataset

4.2.4.4. Bank Dataset

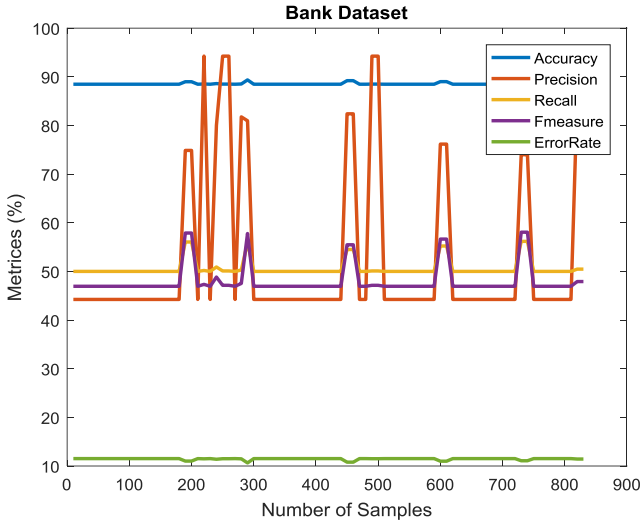
In Fig.4, performance analysis of the methodologies FEFAR, RPARD, and LR-IADS is shown for the bank dataset. In terms of accuracy, RPARD seems to have higher accuracy than the FEFAR and LR-IADS where it is 24% higher than LR-IADS and 24% higher than FEFAR method. In terms of precision, RPARD seems to have higher precision where it is 3.2967% higher than LR-IADS and FEFAR. In terms of recall, RPARD seems to have higher recall where it is 12.5% higher than LR-IADS and FEFAR. In terms of F-measure, RPARD seems to have higher f-measure where it is 14.8803% higher than LR-IADS and FEFAR. In terms of error rate, RPARD seems to have lesser error rate where it is 35.5238% lesser than LR-IADS and 24% lesser than FEFAR.



(a) FEFAR



(b) RPARD



(c) LR-IADS

Fig.4. Performance Analysis of Bank Dataset

6. CONCLUSION

The main goal of this analysis work is to compare the performance of existing and proposed methodologies based on simulation outcome. This research work aims to highlight the performance variation between the proposed and existing techniques and the best method that can offer accurate anomalous transaction detection. The analysis of the research work is carried out on matlab environment over four databases namely soil, bank, german statlog and auto mpg based on which performance outcome has been given.

REFERENCES

[1] T. Watanabe, A. Kitamura, K. Higuchi and H. Ikeda, "Intelligent Manufacturing Database Techniques for Quality and Process Design of Steel Plate", *Proceedings of IEEE Conference on Emerging Technologies and Factory Automation*, pp. 596-603, 2003

[2] R. Srikant and R. Agrawal, "Mining Generalized Association Rules", *Proceedings of IEEE Conference on Very Large Data Bases*, pp. 407-419, 1995.

[3] R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", *Proceedings of IEEE Conference on Management of the Data*, pp.1-12, 1996.

[4] G. Chen and Q. Wei, "Fuzzy Association Rules and the Extended Mining Algorithms", *Information Sciences*, Vo1. l47, No. 1-4, pp. 221-228, 2002.

[5] H. Ishibuchi and T. Yamamoto, "Fuzzy Rule Selection by Data Mining Criteria and Genetic Algorithms", *Proceedings of Annual Conference on Genetic and Evolutionary Computation*, pp. 399-406, 2002.

[6] Y. Hu, R. Chen and G. Tzeng, "Discovering Fuzzy Association Rules Using Fuzzy Partition Methods", *Knowledge-Based Systems*, Vol. 16, No. 3, pp. 137-147, 2003.

[7] T. Watanabe and N. Nakayama, "Fuzzy Rule Extraction Based on the Mining Generalized Association Rules", *Proceedings of IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance*, pp. 2690-2695, 2003.

[8] M. Delgado, N. Marin, D. Sanchez and M.A. Vila, "Fuzzy Association Rules: General Model and Applications", *IEEE transactions on Fuzzy Systems*, Vol 11, No. 2, pp. 214-225, 2003.

[9] M. Delgado, N. Marin, M.J. Martin Bautista, D. Sanchez and M.A. Vila, "Mining Fuzzy Association Rules: An Overview", *Proceedings of IEEE International Conference on Soft Computing for Information Processing and Analysis*, pp. 351-373, 2006.

[10] E. Suzuki, "Discovering Unexpected Exceptions: A Stochastic Approach", Proceedings of IEEE International Conference on Rough Sets, Fuzzy Sets, and Machine Discovery, pp. 225-232, 1996.

[11] F. Berzal, J.C. Cubero, N. Marn and M. Gamez, "Anomalous association rules", *Proceedings of IEEE International Conference on Alternative Techniques for Data Mining and Knowledge Discovery*, pp. 1-8, 2004.

[12] M. Delgado, M.D. Ruiz and D. Sanchez, "New Approaches for Discovering Exception and Anomalous Rules", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 19, No.2, pp. 361-399, 2011.

[13] T. Watanabe and R. Fujioka, "Fuzzy Association Rules Mining Algorithm Based on Equivalence Redundancy of Items", *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pp.1960-1965, 2012.

[14] E. Suzuki, "Autonomous Discovery of Reliable Exception Rules", *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 159-176, 1997.

[15] M.D. Ruiz, D. Snchez, M. Delgado and M.J. Martin Bautista, "Discovering Fuzzy Exception and Anomalous Rules", *IEEE Transactions on Fuzzy Systems*, Vol. 24, No. 4, pp. 930-944, 2016.

[16] E. Suzuki, "Data Mining Methods for Discovering Interesting Exceptions from an Unsupervised Table", *Journal of Universal Computer Science*, Vol. 12, No. 6, pp. 627-653, 2006.

[17] E. Suzuki and J.M. Zytkow, "Unified Algorithm for Undirected Discovery of Exception Rules", *International Journal of Intelligent Systems*, Vol. 20, No. 7, pp. 673-691, 2005.

[18] T. Zhang, W. Zhang, X.U. Wei and H.A.O. Haijing, "Multiple Instance Learning for Credit Risk Assessment with Transaction Data", *Knowledge-Based Systems*, Vol. 161, pp. 65-77, 2018.

[19] S. Senthil Kumar and S. Mythili, "Survey on Exception Rules and Anomaly Detection", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, Vol. 2, No. 6, pp.521- 525, 2017.

[20] S. Senthil Kumar and S. Mythili, "Accurate Fuzzy Anomalous Rule Identification Using Classification Algorithms", *Journal of Advance Research in Dynamical and Control Systems*, Vol. 11, No. 5, pp. 241-262, 2019.

[21] S. Chen, M. Peng, H. Xiong and S. Wu, "An Anomaly Detection Method based on Lasso", *Cluster Computing*, Vol. 22, No. 3, pp.5407-5419, 2019..