

A Novel Approach for Outlier Detection using Rough Entropy

E.N.SATHISHKUMAR, K.THANGAVEL

Department of Computer Science

Periyar University

Salem - 636011, Tamilnadu

INDIA

en.sathishkumar@yahoo.in

Abstract: - Outlier detection is an important task in data mining and its applications. It is defined as a data point which is very much different from the rest of the data based on some measures. Such a data often contains useful information on abnormal behavior of the system described by patterns. In this paper, a novel method for outlier detection is proposed among inconsistent dataset. This method exploits the framework of rough set theory. The rough set is defined as a pair of lower approximation and upper approximation. The difference between upper and lower approximation is defined as boundary. Some of the objects in the boundary region have more possibility of becoming outlier than objects in lower approximations. Hence, it is established that the rough entropy measure as a uniform framework to understand and implement outlier detection separately on class wise consistent (lower) and inconsistent (boundary) objects. An example shows that the Novel Rough Entropy Outlier Detection (NREOD) algorithm is effective and suitable for evaluating the outliers. Further, experimental studies show that NREOD based technique outperformed, compared with the existing techniques.

Key-Words: - Data Mining, Outlier, Rough Set, Classification, Pattern recognition

1 Introduction

Outlier detection refers to the problem of finding patterns in data that are very different from the rest of the data based on appropriate metrics. Such a pattern often contains useful information regarding abnormal behavior of the system described by the data. These inconsistent patterns are usually called outliers, noise, anomalies, exceptions, faults, defects, errors, damage, surprise, novelty or peculiarities in different application domains. Outlier detection is a widely researched problem and finds massive use in application domains such as cancer gene selection, credit card fraud detection, fraudulent usage of mobile phones, unauthorized access in computer networks, abnormal running conditions in aircraft engine rotation, abnormal flow problems in pipelines, military surveillance for enemy activities and many other areas. Outlier detection is most important due to the fact that outliers can have significant information. Outliers can be candidates for abnormal data that may affect systems adversely such as by producing incorrect results, misspecification of models, and biased estimation of parameters. It is therefore important to identify them earlier to modeling and analysis.

With increasing awareness on outlier detection in literatures, more concrete meanings of outliers are

defined for solving problems in specific domains. In [1] Nguyen discusses a method for the detection of outliers, as well as how to obtain background domain knowledge from outliers using multi-level approximate reasoning schemes. Y. Chen, D. Miao, and R. Wang [2] demonstrate the application of granular computing model using information tables for outlier detection. M. M. Breunig proposed a method for identifying density based local outliers [3]. He defines a Local Outlier Factor (LOF) that indicates the degree of outlier-ness of an object using only the object's neighborhood. F. Jiang, Y. Sui and C.Cao [4, 5] propose a new definition of outliers that exploits the rough membership function. Xiangjun Li, Fen Rao [6] propose a new rough entropy based approach to outlier detection.

Rough set theory (RST) is proposed by Z. Pawlak in 1982 [7], which is an extension of set theory for the study of intelligent systems characterized by insufficient, inconsistent and incomplete information. The rough set philosophy is based on the assumption that with every objects of the universe there is related a certain amount of information (data, knowledge), expressed by means of some attributes used for object description. Objects having the same description are indiscernible (similar) with respect to the available

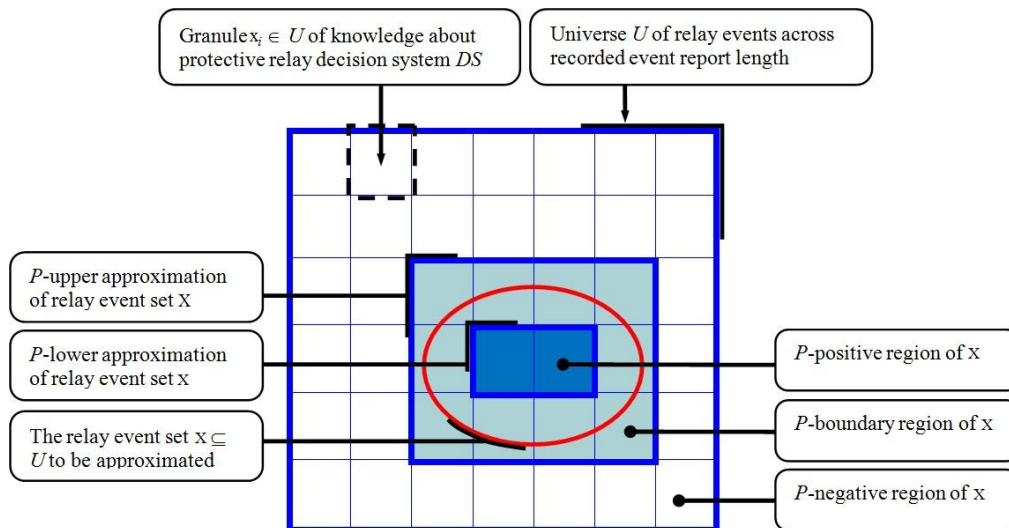


Fig.1 Representation of the data partitioning for a subset X

information. In recent years, there has been a rapid growing interest in this theory. The successful applications of the rough set model in a variety of problems have fully demonstrated its usefulness and adaptability [8, 9].

In this paper, we propose a new method for outlier detection which is based on rough entropy. Representation of the data partitioning for a subset X is shown in figure 1. The basic idea is as follows, For any subset X of the universe and any equivalence relation on the universe, the difference between the upper and lower approximations constitutes the boundary region of the rough set, whose elements cannot be characterized with certainty as belonging or not to X, using the available information (equivalence relation). The information about objects from the boundary region is, therefore, inconsistent or ambiguous.

When given a set of equivalence relations (available information), if an object in X always lies in the lower approximation with respect to every equivalence relation, then we may consider this some of the objects are not behaving normally according to the given knowledge (set of equivalence relations) at hand. We assume such objects may have outliers. Further we study rough entropy measure to discover the outliers from that lower and boundary objects to examine the uncertain information.

The rest of the paper is organized as follows. The basic concepts on rough set and rough entropy are shown in Section 2. Novel Approach using Rough Entropy Outlier Detection Algorithm (NREOD) is introduced in Section 3. In section 4, experimental results are listed. Finally, the conclusion and future work are drawn in Section 5.

2 Methods

2.1 Multivariate Outlier Detection

Multivariate Outlier Detection (MOD) is a classical technique for outlier’s removal based on statistical tails bounds. Statistical methods for multivariate outlier detection often indicate those observations that are located relatively far from the center of the data distribution. Several distance measures are implemented for such a task. The *Mahalanobis* distance is a well-known criterion which depends on estimated parameters of the multivariate distribution. Given n observations from a p -dimensional dataset, denote the sample mean vector by \bar{X}_n and the sample covariance matrix by V_n , where

$$V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n) (X_i - \bar{X}_n)^T \quad (1)$$

The *Mahalanobis* distance for each multivariate data point $i, i = 1, 2 \dots n$, is denoted by M_i and given by

$$M_i = \left(\sum_{i=1}^n (X_i - \bar{X}_n)^T V_n^{-1} (X_i - \bar{X}_n) \right)^{1/2} \quad (2)$$

Accordingly, those observations with a large *Mahalanobis* distance are indicated as outliers [10].

2.2 Rough Set Theory

Rough Set Theory approach involves the concept of indiscernibility [11, 12]. Let Information System $(IS) = (U, A, C, D)$ be a decision system data, where U is a non-empty finite set called the universe, A is a set of features, C and D are subsets of A, named the conditional and decisional attributes subsets respectively. The elements of U are called objects, cases, instances or observations. Attributes are

interpreted as features, variables or characteristics conditions. Given a feature a , such that: $a: U \rightarrow V_a$ for $a \in A$, V_a is called the value set of a . Let $a \in A$, $P \subseteq A$, the indiscernibility relation $IND(P)$, is defined as follows:

$$IND(P) = \{(x, y) \in U \times U : \text{for all } a \in P, a(x) = a(y)\} \quad (3)$$

The partition generated by $IND(P)$ is denoted as $U/IND(P)$ or abbreviated to U/P and is calculated as follows:

$$U/IND(P) = \bigotimes \{a \in P / U/IND(\{a\})\} \quad (4)$$

where $A \otimes B = \{X \cap Y / \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\}$ where A and B are families of sets. If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P . The equivalence classes of the P -indiscernibility relation are denoted by $[x]_P$.

2.2.1 Lower approximation of a subset

Let $R \subseteq C$ and $X \subseteq U$, the R -lower approximation set of X , is the set of all elements of U which can be with certainty classified as elements of X .

$$\underline{RX} = \cup \{ Y \in U / R : Y \subseteq X \} \quad (5)$$

According to this definition, we can see that R -Lower approximation is a subset of X , thus $\underline{RX} \subseteq X$.

2.2.2 Upper approximation of a subset

The R -upper approximation set of X is the set of all element of U , which can possibly belong to the subset of interest X .

$$\bar{RX} = \cup \{ Y \in U / R : Y \cap X \neq \emptyset \} \quad (6)$$

Note that X is a subset of the R -upper approximation set, thus $X \subseteq \bar{RX}$.

2.2.3 Boundary Region

It is the collection of elementary sets defined by:

$$BND(X) = \bar{RX} - \underline{RX} \quad (7)$$

These sets are included in R -Upper but not in R -Lower approximations. A subset defined through its lower and upper approximations is called a Rough set. That is, when the boundary region is a non-empty set ($\bar{RX} \neq \underline{RX}$).

2.3 Rough Entropy

Rough entropy is extended entropy to measure the uncertainty in rough sets. Given an information system, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes. For any $B \subseteq$

A , let $IND(B)$ be the equivalence relation as the form of $U/IND(B) = \{B_1, B_2, \dots, B_m\}$.

The rough entropy $E(B)$ of equivalence relation $IND(B)$ is defined by

$$E(B) = - \sum_{i=1}^m \frac{|B_i|}{|U|} \log \frac{1}{|B_i|} \quad (8)$$

where, $\frac{|B_i|}{|U|}$ denotes the probability of any element $x \in U$ being in equivalence class B_i ; $1 \leq i \leq m$. And $|M|$ denotes the cardinality of set M . The relative rough entropy $RE(x)$ of object x is defined by

$$RE(x) = E_x(B)/E(B) \quad (9)$$

Given any $B \subseteq A$ and $x \in U$, when we delete the object x from U , if the rough entropy of $IND(B)$ decreases greatly, then we may consider the uncertainty of object x under $IND(B)$ is high. On the other hand, if the rough entropy of $IND(B)$ varies little, then we may consider the uncertainty of object x under $IND(B)$ is low. Therefore, the relative rough entropy $RE(x)$ of x under $IND(B)$ gives a measure for the uncertainty of x .

In an information system, the rough entropy outlier factor $REOF(x)$ of object x in IS is defined as follows:

$$REOF(x) = \frac{(\sum_{j=1}^k RE_{a_j}(x) \times W_{a_j}(x))}{k} \quad (10)$$

where, $RE_{a_j}(x)$ is the relative rough entropy of object x , for every singleton subset $a_j \in A$, $1 \leq j \leq k$. For any $a \in A$, $W_a : U \rightarrow (0, 1]$ is a weight function such that for any $x \in U$, $W_a(x) = 1 - |x]_a|/|U|$.

Let v be a given threshold value. For any object $x \in U$, if $REOF(x) > v$, then object x is called a RE-based outlier in IS , where $REOF(x)$ is the rough entropy outlier factor of x in IS [6].

3 Novel Approach using Rough Entropy Outlier Detection Algorithm

The proposed NREOD algorithm logically consists of two steps:

- (i) Find class wise certain and uncertain objects based on Rough Set Theory,
- (ii) Compute outlier object from certain and uncertain objects using rough entropy measure.

The overall process of NREOD Algorithm is represented in the figure 2 and the steps involved in NREOD method is described in algorithm 1.

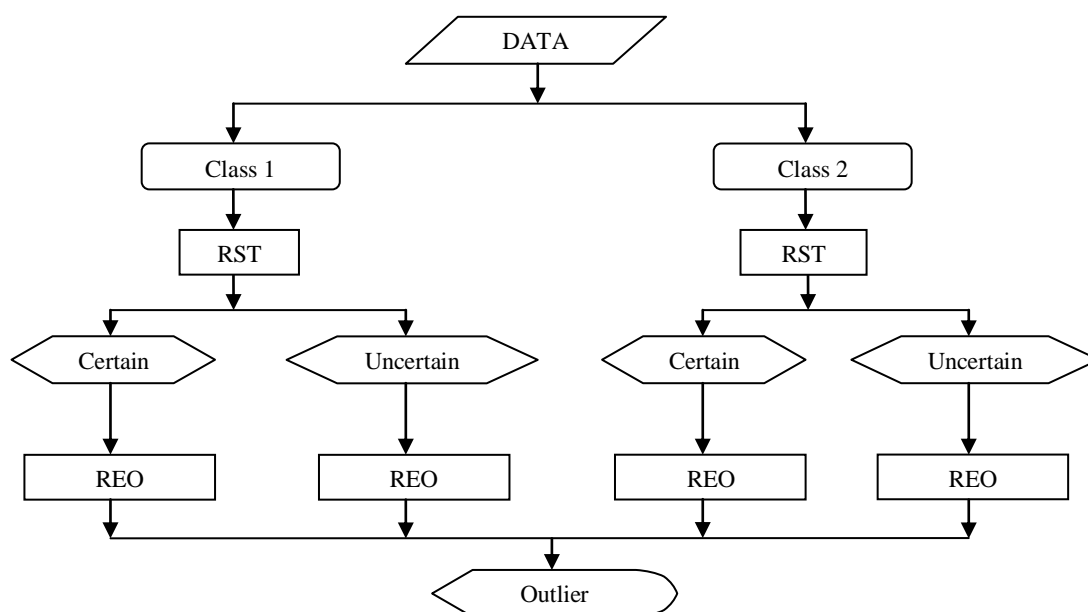


Fig.2 Process of NREOD algorithm

Algorithm 1: NREOD (C, D)

IS = (U, A, C, D) be a decision system data,

C, Conditional attribute,

D, Decision attribute

- (1) Calculate the partition $[x]_C \leftarrow C$
- (2) Calculate the partition $[x]_D \leftarrow D$
- (3) $IND(C) \leftarrow [x]_d$ where, $d=1, 2, \dots, |[x]_D|$
- (4) Calculate the Upper Approximation $\bar{R}X \leftarrow \{x \in U \mid [x]_d \cap X \neq \Phi\}$
- (5) Calculate the Lower Approximation $\underline{R}X \leftarrow \{x \in U \mid [x]_d \subseteq X\}$
- (6) Calculate the Boundary Regions $BND_d(x) \leftarrow \cup \bar{R}X_d - \cup \underline{R}X_d$
- (7) $RLB = \{\underline{R}X_d, BND_d(x)\}$
- (8) For every $S \in RLB$
- (9) For every S, where $S = \{a_1, a_2, \dots, a_m\}$, $|U| = n$ and $|S| = m$; a threshold value v_d .
- (10) For every $a \in S$
- (11) Calculate the partition $U/IND(\{a\})$;
- (12) Calculate the rough entropy $E(\{a\})$, which is the rough entropy of $U/IND(\{a\})$
- (13) end
- (14) For every $x_i \in U$
- (15) For $j = 1$ to m
- (16) Calculate the rough entropy of $E_{x_i}(\{a\})$
- (17) Calculate $RE\{a_j\}(x_i)$, which is the relative rough entropy of x_i ;
- (18) Assign a weight $W\{a_j\}(x_i)$ to x_i ;
- (19) end
- (20) Calculate $REOF_d(x_i)$;
- (21) If $REOF_d(x_i) > v_d$, then $O_d = O_d \cup \{x_i\}$;
- (22) end
- (23) then $Outlier = Outlier \cup O_d$
- (24) end
- (25) $NREO = NREO \cup Outlier$
- (26) end
- (27) Return NREO

3.1 Example

Given an information system $IS = (U, A)$, where $U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$, $A = \{a, b, c\}$, as shown in Table 1. Set threshold ν as 0.65. The value of ν has been adopted from [6].

Table 1. An information system

U	Condition			Decision d
	a	b	c	
u_1	5	6	6	Yes
u_2	5	6	6	No
u_3	4	6	5	No
u_4	5	7	6	Yes
u_5	5	6	6	Yes
u_6	5	6	7	Yes
u_7	6	5	5	No
u_8	4	6	5	Yes
u_9	4	6	5	No
u_{10}	6	6	7	yes

R is an equivalence relation, the indiscernibility classes defined by $R = \{a, b, c\}$ are $R = \{u_1, u_2, u_5\}, \{u_3, u_8, u_9\}, \{u_4\}, \{u_6\}, \{u_7\}, \{u_{10}\}$

Calculate the Rough Entropy Outlier Factor for 'decision = yes' Objects:

$$X_1 = \{u \mid d(u) = \text{yes}\}$$

$$\bar{R}X_1 = \{u_1, u_4, u_5, u_6, u_8, u_{10}\}$$

$$RX_1 = \{u_4, u_6, u_{10}\}$$

$$BND(X_1) = \bar{R}X_1 - RX_1 = \{u_1, u_5, u_8\}$$

Lower Approximation:

Find an Outlier from $\bar{R}X_1 = \{u_4, u_6, u_{10}\}$

Table 2. certain objects from class='yes'

$\bar{R}X_1$	a	b	c
u_4	5	7	6
u_6	5	6	7
u_{10}	6	6	7

The partitions induced by all singleton subsets of A are as follows:

$$U/IND(a) = \{\{u_4, u_6\}, \{u_{10}\}\}$$

$$U/IND(b) = \{\{u_4\}, \{u_6, u_{10}\}\}$$

$$U/IND(c) = \{\{u_4\}, \{u_6, u_{10}\}\}$$

From the definition of rough entropy, we can obtain that

$$E(\{a\}) = -(2/3 \log(1/2) + 1/3 \log(1/1)) = 0.2007$$

$$E(\{b\}) = E(\{c\}) = -(1/3 \log(1/1) + 2/3 \log(1/2)) = 0.2007$$

When remove the object u , we can obtain that

$$Eu_4(\{a\}) = Eu_6(\{a\}) = -(1/2 \log(1/1) + 1/2 \log(1/1)) = 0$$

$$Eu_{10}(\{a\}) = -(2/2 \log(1/2)) = 0.3010$$

$$Eu_4(\{b\}) = -(2/2 \log(1/2)) = 0.3010$$

$$Eu_6(\{b\}) = Eu_{10}(\{b\}) = -(1/2 \log(1/1) + 1/2 \log(1/1)) = 0$$

$$Eu_4(\{c\}) = -(2/2 \log(1/2)) = 0.3010$$

$$Eu_6(\{c\}) = Eu_{10}(\{c\}) = -(1/2 \log(1/1) + 1/2 \log(1/1)) = 0$$

Correspondingly, according to the definition of relative rough entropy, we can obtain that

$$RE\{a\}(u_4) = RE\{a\}(u_6) = RE\{b\}(u_6) = RE\{b\}(u_{10}) =$$

$$RE\{c\}(u_6) = RE\{c\}(u_{10}) = 0/0.2007 = 0$$

$$RE\{a\}(u_{10}) = RE\{b\}(u_4) = RE\{c\}(u_4) =$$

$$0.3010/0.2007 = 1.5000$$

Calculate the weight $W\{aj\}$ as follows

$$W\{a\}(u_4) = W\{a\}(u_6) = W\{b\}(u_6) = W\{b\}(u_{10}) =$$

$$W\{c\}(u_6) = W\{c\}(u_{10}) = 0.3333$$

$$W\{a\}(u_{10}) = W\{b\}(u_4) = W\{c\}(u_4) = 0.6667$$

Hence, the rough entropy outlier factor is calculated as follows:

$$REOF(u_4) = (0 * 0.3333 + 1.5 * 0.6667 + 1.5 * 0.6667)/3 = 2.0001/3 = 0.6667 > \nu,$$

$$REOF(u_6) = (0 * 0.3333 + 0 * 0.3333 + 0 * 0.3333)/3 = 0/3 = 0 < \nu,$$

$$REOF(u_{10}) = (1.5 * 0.6667 + 0 * 0.3333 + 0 * 0.3333)/3 = 1.00005/3 = 0.3333 < \nu,$$

Similarly find an outlier from Uncertain Objects (Class= 'yes'), $BND(X_1) = \{u_1, u_5, u_8\}$

$$REOF(u_1) \approx 0 < \nu, REOF(u_5) \approx 0 < \nu, REOF(u_8) \approx 0.6667 > \nu,$$

Analogously, we can obtain that $\bar{R}X_2 = \{u_7\}$ and $BND(X_2) = \{u_2, u_3, u_9\}$ for 'decision = no' objects.

$$REOF(u_7) \approx 0 < \nu,$$

$$REOF(u_2) \approx 0.6667 > \nu, REOF(u_3) \approx 0 < \nu,$$

$$REOF(u_9) \approx 0 < \nu,$$

Therefore, $u_2, u_4,$ and u_8 are outlier in IS. Other objects in U are all non outliers.

4 Experimental Results

4.1 Data Set

In this section, we describe the datasets used to analyze the methods studied in sections 2 and 3, which are found in the UCI machine learning repository [13].

4.1.1 Car Evaluation Data Set

The car evaluation dataset was derived from a simple hierarchical decision model originally developed for the demonstration of DEX. The data set contains 1728 instances with 6 attributes. An example in the dataset describes the price and technical features of a car and is assigned one of four classes. The distribution of the examples is heavily weighted towards two classes. There is also an intuitive ordering to the classes, ranging from unacceptable to very good.

4.1.2 Yeast Data Set

Yeast dataset predicting the Cellular Localization Sites of Proteins, it contains 1484 examples. In the yeast dataset, eight features (attributes) are used:

mcg, gvh, alm, mit, erl, pox, vac, nuc. And proteins are classified into 10 classes: cytosolic (CYT), nuclear (NUC), mitochondrial (MIT), membrane protein without N-terminal signal (ME3), membrane protein with uncleaved signal (ME2), membrane protein with cleaved signal (ME1), extracellular (EXC), vacuolar (VAC), peroxisomal (POX), endoplasmic reticulum lumen (ERL).

4.1.3 Breast Tissue Data Set

This is a dataset with electrical impedance measurements in samples of freshly excised tissue from the Breast. It consists of 106 instances. 10 attributes: 9 features+1 class attribute. Six classes of freshly excised tissue were studied using electrical impedance measurements. The six classes namely Carcinoma, Fibro-adenoma, Mastopathy, Glandular, Connective, Adipose.

4.2 Outlier Detection

In this study, we first find the class wise certain and uncertain objects for all dataset based on Rough Set Theory, it identifies group of objects that exhibit same equivalence relation. After that we apply rough entropy measure to discover the outliers from that certain and uncertain objects. Before applying proposed algorithm all the conditional attributes are discretized using K-Means discretization [14, 15, 16]. The numbers of objects in each class of all datasets are tabulated in tables 3, 4 and 5. The number of certain and uncertain objects along with number of selected outliers selected by proposed NREOD method are also tabulated.

Table 3. Car evaluation dataset Class wise NREOD Outliers

S. No	Class	No. of Instances	Certain instances		Uncertain instances		Total Outliers
			Objects	Outliers	Objects	Outliers	
1	unacc	1210	792	12	418	5	17
2	acc	384	42	6	342	12	18
3	good	69	0	0	69	6	6
4	v-good	65	0	0	65	4	4
Total		1728	834	18	894	27	45

The proposed algorithm NREOD has selected eighteen outliers out of 834 certain objects and twenty seven out of 894 uncertain objects in the car evaluation data set. The total number of outliers selected from car evaluation data set is tabulated in table 3.

The proposed algorithm NREOD has selected thirty eight outliers out of 615 certain objects and thirty five out of 869 uncertain objects in the yeast data set. The total number of outliers selected from yeast data set is tabulated in table 4.

Table 4. Yeast dataset Class wise NREOD Outliers

S. No	Class	No. of Instances	Certain instances		Uncertain instances		Total Outliers
			Objects	Outliers	Objects	Outliers	
1	CYT	463	125	5	338	7	12
2	NUC	429	182	7	247	7	14
3	MIT	244	113	7	131	7	14
4	ME3	163	105	7	58	4	11
5	ME2	51	26	4	25	2	6
6	ME1	44	27	3	17	3	6
7	EXC	35	17	1	18	1	2
8	VAC	30	5	1	25	3	4
9	POX	20	12	2	8	1	3
10	ERL	5	3	1	2	0	1
Total		1484	615	38	869	35	73

Table 5. Breast Tissue dataset Class wise NREOD Outliers

S. No	Class	No. of Instances	Certain instances		Uncertain instances		Total Outliers
			Objects	Outliers	Objects	Outliers	
1	Carcinoma	21	18	2	3	2	4
2	Fibro-adenoma	15	1	0	14	2	2
3	Mastopathy	18	3	0	15	2	2
4	Glandular	16	1	0	15	1	1
5	Connective	14	13	3	1	0	3
6	Adipose	22	19	5	3	0	5
Total		106	55	10	51	7	17

The proposed algorithm NREOD has selected ten outliers out of 55 certain objects and seven out of 51 uncertain objects in the breast tissue data set. The total number of outliers selected from breast tissue data set is tabulated in table 5.

4.3 Classification Results

Backpropagation is a neural network learning algorithm. The neural networks field was originally kindled by psychologists and neurobiologists who sought to develop and test computational analogs of neurons. A neural network is a set of connected input/output units in which each connection has a weight associated with it. Back Propagation learns by iteratively processing a data set of training tuples, comparing the network's prediction for each tuple with the actual known target value. The target value may be the known class label of the training tuple (for classification problems) or a continuous value (for prediction). For each training tuple, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual target value. These modifications are made in the "backwards" direction, that is, from the output layer, through each hidden layer down to the first hidden layer (hence the name back propagation) [17]. This classification method has been employed for this study and validates using 10-fold cross validation.

In this section, the NREOD method is compared with the MOD and REOD methods. The BPN classification was initially performed on the unreduced data set, followed by the outlier removed data sets which were obtained by using the MOD, REOD and NREOD methods. Results are presented in terms of classification accuracy. The numbers of selected outliers are tabulated in Table 6.

Table 6. Selected Outliers

Dataset	MOD Outliers	REOD Outliers	NREOD Outliers
Car Evaluation	28	30	45
Yeast	25	59	73
Breast Tissue	16	22	17

Table 7. BPN 10-Fold Validation for entire Car Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	173 to 1728	1 to 172	172/172	100
2	1 to 172, 345 to 1728	173 to 344	166/172	96.51
3	1 to 344, 517 to 1728	345 to 516	160/172	93.02
4	1 to 516, 689 to 1728	517 to 688	137/172	79.65
5	1 to 688, 861 to 1728	689 to 860	137/172	79.65
6	1 to 860, 1033 to 1728	861 to 1032	143/172	83.14
7	1 to 1032, 1205 to 1728	1033 to 1204	113/172	65.70
8	1 to 1204, 1377 to 1728	1205 to 1376	130/172	75.58
9	1 to 1376, 1549 to 1728	1377 to 1548	133/172	77.33
10	1 to 1548	1549 to 1728	121/180	70.35
Mean Accuracy				82.09

Table 8. BPN 10-Fold Validation for MOD Outliers Removed Car Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	171 to 1700	1 to 170	166/170	97.64
2	1 to 170, 341 to 1700	171 to 340	163/170	95.88
3	1 to 340, 511 to 1700	340 to 510	143/170	84.11
4	1 to 510, 681 to 1700	511 to 680	137/170	80.58
5	1 to 680, 851 to 1700	681 to 850	164/170	96.47
6	1 to 850, 1021 to 1700	851 to 1020	141/170	82.94
7	1 to 1020, 1191 to 1700	1021 to 1190	129/170	75.88
8	1 to 1190, 1361 to 1700	1191 to 1360	138/170	81.17
9	1 to 1360, 1531 to 1700	1361 to 1530	140/170	82.35
10	1 to 1530	1531 to 1700	103/170	60.58
Mean Accuracy				83.76

The computational results of car evaluation data set by applying BPN with 10 fold cross validations are tabulated in table 7. The mean accuracy of classification result is 82.09% before removing outliers.

Table 9. BPN 10-Fold Validation for REOD Outliers Removed Car Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	172 to 1716	1 to 171	165/171	96.49
2	1 to 171, 343 to 1716	172 to 342	133/171	77.78
3	1 to 342, 514 to 1716	343 to 513	163/171	95.32
4	1 to 513, 686 to 1716	514 to 685	135/171	78.95
5	1 to 685, 857 to 1716	686 to 856	160/171	93.57
6	1 to 856, 1028 to 1716	857 to 1027	141/171	82.46
7	1 to 1027, 1199 to 1716	1028 to 1198	129/171	75.44
8	1 to 1198, 1370 to 1716	1199 to 1369	132/171	77.19
9	1 to 1369, 1541 to 1716	1370 to 1540	146/171	85.38
10	1 to 1540	1541 to 1716	129/176	73.30
Mean Accuracy				83.59

The computational results of car evaluation data set by applying BPN with 10 fold cross validations are tabulated in table 8 and 9. The mean accuracy of classification result is 83.76% and 83.59% obtained by applying existing MOD and REOD methods.

Table 10. BPN 10-Fold Validation for NREOD Outliers Removed Car Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	169 to 1683	1 to 168	168/168	100
2	1 to 168, 337 to 1683	169 to 336	159/168	94.64
3	1 to 336, 505 to 1683	337 to 504	165/168	98.21
4	1 to 504, 673 to 1683	505 to 672	141/168	83.93
5	1 to 672, 841 to 1683	673 to 840	158/168	94.05
6	1 to 840, 1009 to 1683	841 to 1008	152/168	90.48
7	1 to 1008, 1177 to 1683	1009 to 1176	130/168	77.38
8	1 to 1176, 1345 to 1683	1177 to 1344	149/168	88.69
9	1 to 1344, 1513 to 1683	1345 to 1512	141/168	83.93
10	1 to 1512	1513 to 1683	125/171	73.10
Mean Accuracy				88.44

The computational results of car evaluation data set by applying BPN with 10 fold cross validations are tabulated in table 10. The mean accuracy of classification result is 88.44% obtained by applying proposed NREOD method.

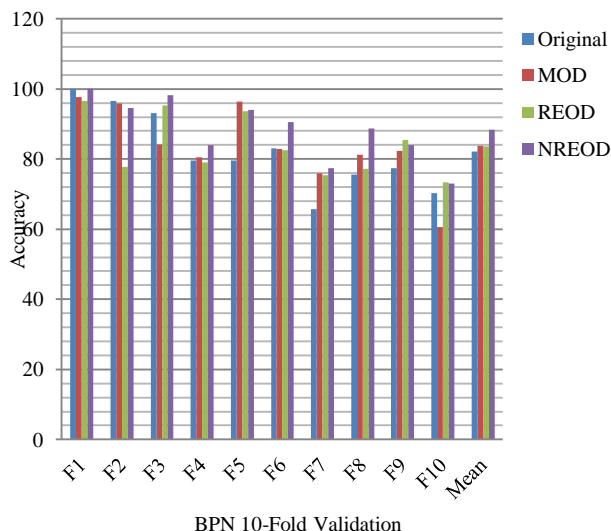


Fig. 3 Classification Accuracy of Car dataset

The classification accuracy of BPN is represented in the figure 3 for car evaluation dataset. The highest classification accuracy is achieved as 88.44%.

Table 11. BPN 10-Fold Validation for entire Yeast Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	149 to 1484	1 to 148	62/148	41.89
2	1 to 148, 297 to 1484	149 to 296	54/148	36.49
3	1 to 296, 445 to 1484	297 to 444	61/148	41.22
4	1 to 444, 593 to 1484	445 to 592	66/148	44.59
5	1 to 592, 741 to 1484	593 to 740	54/148	36.49
6	1 to 740, 889 to 1484	741 to 888	61/148	41.22
7	1 to 888, 1037 to 1484	889 to 1036	56/148	37.84
8	1 to 1036, 1185 to 1484	1037 to 1184	56/148	37.84
9	1 to 1184, 1333 to 1484	1185 to 1332	49/148	33.11
10	1 to 1332	1333 to 1484	53/152	34.86
Mean Accuracy				38.49

The computational results of yeast data set by applying BPN with 10 fold cross validations are tabulated in table 11. The mean accuracy of classification result is 38.49% before removing outliers.

Table 12. BPN 10-Fold Validation for MOD Outliers Removed Yeast Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	146 to 1459	1 to 145	63/145	43.45
2	1 to 145, 291 to 1459	146 to 290	44/145	30.34
3	1 to 290, 436 to 1459	291 to 435	81/145	55.86
4	1 to 435, 581 to 1459	436 to 580	60/145	41.37
5	1 to 580, 726 to 1459	581 to 725	43/145	29.66
6	1 to 725, 871 to 1459	726 to 870	52/145	35.86
7	1 to 870, 1016 to 1459	871 to 1015	72/145	49.65
8	1 to 1015, 1161 to 1459	1016 to 1160	56/145	38.62
9	1 to 1160, 1306 to 1459	1161 to 1305	57/145	39.31
10	1 to 1305	1306 to 1459	35/154	22.72
Mean Accuracy				38.68

Table 13. BPN 10-Fold Validation for REOD Outliers Removed Yeast Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	143 to 1425	1 to 142	54/142	38.03
2	1 to 142, 285 to 1425	143 to 284	65/142	45.77
3	1 to 284, 427 to 1425	285 to 426	59/142	41.55
4	1 to 426, 569 to 1425	427 to 568	59/142	41.55
5	1 to 568, 711 to 1425	569 to 710	46/142	32.39
6	1 to 710, 853 to 1425	711 to 852	59/142	41.55
7	1 to 852, 995 to 1425	853 to 994	57/142	40.14
8	1 to 994, 1137 to 1425	995 to 1136	63/142	44.37
9	1 to 1136, 1279 to 1425	1137 to 1278	47/142	33.10
10	1 to 1278	1279 to 1425	59/147	40.14
Mean Accuracy				39.86

The computational results of yeast data set by applying BPN with 10 fold cross validations are tabulated in table 12 and 13. The mean accuracy of classification result is 38.68% and 39.86% obtained by applying existing MOD and REOD methods.

The computational results of yeast data set by applying BPN with 10 fold cross validations are tabulated in table 14. The mean accuracy of classification result is 40.47% obtained by applying proposed NREOD method.

Table 14. BPN 10-Fold Validation for NREOD Outliers Removed Yeast Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	142 to 1411	1 to 141	55/141	39.01
2	1 to 141, 283 to 1411	142 to 282	65/141	46.10
3	1 to 282, 424 to 1411	283 to 423	63/141	44.68
4	1 to 423, 565 to 1411	424 to 564	50/141	35.46
5	1 to 564, 706 to 1411	565 to 705	55/141	39.01
6	1 to 705, 847 to 1411	706 to 846	63/141	44.68
7	1 to 846, 988 to 1411	847 to 987	59/141	41.84
8	1 to 987, 1129 to 1411	988 to 1128	62/141	43.97
9	1 to 1128, 1270 to 1411	1129 to 1269	46/141	32.62
10	1 to 1269	1270 to 1411	53/142	37.32
Mean Accuracy				40.47

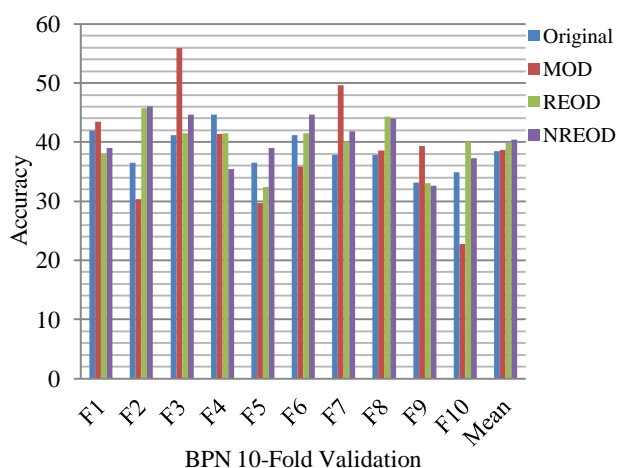


Fig. 4 Classification Accuracy of Yeast dataset

The classification accuracy of BPN is represented in the figure 4 for yeast dataset. The highest classification accuracy is achieved as 40.47%.

Table 15. BPN 10-Fold Validation for entire Breast Tissue Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	11 to 106	1 to 10	6/10	60.00
2	1 to 10, 21 to 106	11 to 20	3/10	30.00
3	1 to 20, 31 to 106	21 to 30	4/10	40.00
4	1 to 30, 41 to 106	31 to 40	4/10	40.00
5	1 to 40, 51 to 106	41 to 50	8/10	80.00
6	1 to 50, 61 to 106	51 to 60	2/10	20.00
7	1 to 60, 71 to 106	61 to 70	3/10	30.00
8	1 to 70, 81 to 106	71 to 80	4/10	40.00
9	1 to 80, 91 to 106	81 to 90	5/10	50.00
10	1 to 90	91 to 106	6/16	37.50
Mean Accuracy				42.75

The computational results of breast tissue data set by applying BPN with 10 fold cross validations are tabulated in table 15. The mean accuracy of classification result is 42.75% before removing outliers.

Table 16. BPN 10-Fold Validation for MOD Outliers Removed Breast Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	10 to 90	1 to 9	5/9	55.55
2	1 to 9, 19 to 90	10 to 18	3/9	33.33
3	1 to 18, 28 to 90	19 to 27	5/9	55.55
4	1 to 27, 37 to 90	28 to 36	6/9	66.66
5	1 to 36, 46 to 90	37 to 45	4/9	44.44
6	1 to 45, 55 to 90	46 to 54	3/9	33.33
7	1 to 54, 64 to 90	55 to 63	6/9	66.66
8	1 to 63, 73 to 90	64 to 72	4/9	44.44
9	1 to 72, 82 to 90	73 to 81	5/9	55.55
10	1 to 81	82 to 90	4/9	44.44
Mean Accuracy				49.99

Table 17. BPN 10-Fold Validation for REOD Outliers Removed Breast Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	9 to 84	1 to 8	4/8	50.00
2	1 to 8, 17 to 84	9 to 16	4/8	50.00
3	1 to 16, 25 to 84	17 to 24	4/8	50.00
4	1 to 24, 33 to 84	25 to 32	3/8	37.50
5	1 to 32, 41 to 84	33 to 40	5/8	62.50
6	1 to 40, 49 to 84	41 to 48	3/8	37.50
7	1 to 48, 57 to 84	49 to 56	4/8	50.00
8	1 to 56, 65 to 84	57 to 64	3/8	37.50
9	1 to 64, 73 to 84	65 to 72	4/8	50.00
10	1 to 72	73 to 84	5/12	41.67
Mean Accuracy				46.67

The computational results of breast tissue data set by applying BPN with 10 fold cross validations are tabulated in table 16 and 17. The mean accuracy of classification result is 49.99 and 46.67% by applying existing MOD and REOD methods.

The computational results of breast tissue data set by applying BPN with 10 fold cross validations are tabulated in table 18. The mean accuracy of classification result is 51.62% by applying proposed NREOD method.

The classification accuracy of BPN is represented in the figure 5 for breast tissue dataset. The highest classification accuracy is achieved as 51.62%.

Table 18. BPN 10-Fold Validation for NREOD Outliers Removed Breast Dataset

Fold	Training	Testing	Correctly Classified	Accuracy (%)
1	9 to 89	1 to 8	5/8	62.50
2	1 to 8, 17 to 89	9 to 16	4/8	50.00
3	1 to 16, 25 to 89	17 to 24	3/8	37.50
4	1 to 24, 33 to 89	25 to 32	3/8	37.50
5	1 to 32, 41 to 89	33 to 40	5/8	62.50
6	1 to 40, 49 to 89	41 to 48	4/8	50.00
7	1 to 48, 57 to 89	49 to 56	4/8	50.00
8	1 to 56, 65 to 89	57 to 64	6/8	75.00
9	1 to 64, 73 to 89	65 to 72	4/8	50.00
10	1 to 72	73 to 89	7/17	41.18
Mean Accuracy				51.62

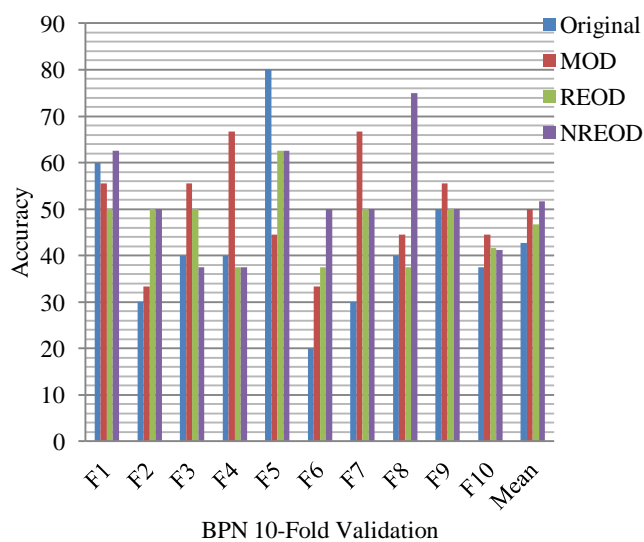


Fig. 5 Classification Accuracy of Breast tissue dataset

It is interesting to note that an increase in classification accuracy is recorded for the proposed and the MOD, REOD methods, with respect to the unreduced data in some cases. This increase in classification accuracy is a little bit high when comparing the MOD, REOD and the NREOD methods to the unreduced data. Also, when comparing classification results, proposed NREOD method outperformed, compared with the existing methods.

5 Conclusion and Future work

In this paper, we have proposed rough entropy based novel approach (NREOD) to discover outliers for the given dataset. We studied and implemented the MOD and REOD outlier detection algorithm successfully. The proposed NREOD method utilizes the framework of rough set and rough entropy for detecting outliers. BPN classifier has been used for

classification. Experimental results on different data sets

have shown the efficiency of the proposed approach.

The proposed work may be extended for gene expression data set. This is the direction for further research. Future researches should be directed to the following aspect. For the NREOD-based outlier detection algorithm, we can adopt rough set feature selection method to reduce the redundant features while preserving the performance of it. This technique can also be applied to other high dimensional data besides gene expression data.

Acknowledgment

The present work is supported by Special Assistance Programme of University Grants Commission, New Delhi, India (Grant No. F.3-50/2011 (SAP-II)).

The first author immensely acknowledges the partial financial assistance under University Research Fellowship, Periyar University, Salem – 636 011, Tamilnadu, India.

References:

- [1] Nguyen, T.T.: "Outlier Detection: An Approximate Reasoning Approach." Springer, Heidelberg (2007).
- [2] Chen, Y., Miao, D., Wang, R.: "Outlier Detection Based on Granular Computing." Springer, Heidelberg (2008).
- [3] Breunig, M.M., Kriegel, H.P., Ng, R.T., and Sander, J.: "LOF: Identifying density based local outliers", In Proc. ACM SIGMOD Conf., (2000) 93–104
- [4] Jiang, F., Sui, Y., Cunge: "Outlier Detection Based on Rough Membership Function." Springer, Heidelberg (2006).
- [5] Jiang, F., Sui, Y., Cunge: "Outlier Detection Using Rough Set Theory." Springer, Heidelberg (2005).
- [6] Xiangjun Li, Fen Rao: "An Rough Entropy Based Approach to Outlier Detection." Journal of Computational Information Systems 8: 24 (2012) 10501-10508
- [7] Pawlak Z.(1982), "Rough sets", International Journal of Computer and Information Sciences 2 (1982) 341–356.
- [8] Zalinda Othman et.al, "Dynamic Tabu Search for Dimensionality Reduction in Rough Set," WSEAS Transactions on Computers, Issue 4, Volume 11, April 2012.
- [9] Krupka Jiri and Jirava Pavel, "Modelling of Rough-Fuzzy Classifier", WSEAS Transactions on Systems, Issue 3, Volume 7, March 2008.

- [10] Irad Ben-Gal, "Outlier detection", Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.
- [11] Kun Gao, Zhongwei Chen and Meiqun Liu, "Predicting performance of Grid based on Rough Set," WSEAS Transactions on Systems, Issue 3, Volume 7, March 2008.
- [12] Yan WANG and Lizhuang MA, "Feature Selection for Medical Dataset Using Rough Set Theory", Proceedings of the 3rd WSEAS International Conference on Computer Engineering and Applications (CEA'09), ISSN: 1790-5117.
- [13] Bay, S. D., The UCI KDD repository, <http://kdd.ics.uci.edu>, 1999
- [14] E.N.Sathishkumar, K.Thangavel and A.Nishama, "Comparative Analysis of Discretization Methods for Gene Selection of Breast Cancer Gene Expression Data", Proceedings of ICC3, Advances in Intelligent Systems and Computing, Publisher Springer India (2014), Vol. 246, pp 373-378.
- [15] E.N.Sathishkumar, K.Thangavel and T.Chandrasekhar, "A New Hybrid K-Mean-QuickReduct Algorithm for Gene Selection," WASET: International Journal of Computer, Information Science and Engineering, Vol.7, No. 2, 2013, Pages: 47-52. ISSN: 1307-6892
- [16] E.N.Sathishkumar, K.Thangavel and T.Chandrasekhar, "A Novel Approach for Single Gene Selection Using Clustering and Dimensionality Reduction," International Journal of Scientific & Engineering Research, Vol: 4, Issue 5, May-2013. ISSN: 2229-551
- [17] Jiawei Han and Micheline Kamber.: Data Mining: Concepts and Techniques, Second Edition, Elsevier (2006).