Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 47 (2015) 351 – 359

Procedia Computer Science

A Novel Neighborhood Rough set Based Classification Approach for Medical Diagnosis

S. Udhaya kumar[a], H. Hannah Inbarani[b]

[a]University Research Fellow, Department of computer science, Periyar University, TN, India. Email: uk2804@gmail.com
[b]Asst. Professor, Department of computer science, Periyar University, TN, India. Email: hhinba@gmail.com

Abstract

Medical datasets consume enormous amount of information about the patients, diseases and the physicians. Diseases diagnosis required many expensive tests to predict the diseases. Cost of disease prediction and diagnosis can be reduced by applying machine learning and data mining methods. Disease prediction and decision making plays asignificant role in medical diagnosis. In this study, a novel neighborhood rough set classification approach is presented to deal with medical datasets. Five benchmarked medical datasets have been used in this research work for studying the impact of proposed work in decision making.Experimental resultof the proposed classification algorithm is compared with other existing approaches such as rough set, K^{th} −nearest neighbor, support vector machine, Back propagation algorithm and multilayer perceptron to conclude that the proposed approach is cheaper way for disease prediction and decision making. The performance of classification algorithms measured based on various classification accuracy measures.

Peer-review under responsibility of organizing committee of the Graph Algorithms, High Performance Implementations and Applications (ICGHIA2014)

*Keywords:*Rough set; Neighborhood rough set; Neighborhoodrough set classification algorithm; Medical diagnosis.

1. Introduction

In the opening of 1980's, Pawlak initiated new mathematical tool called as rough set theory for handling vagueness and uncertainty in certain dataset. Pawlak's rough set based classification algorithms are recognized based on the equivalence relation and only appropriate for discrete data sets [1]. When handling continuous data with Pawlak model, the cost of computation becomes very high. Hence to overcome this problem, several extensions of rough set theory wereintroduced for replacing equivalent relation, dimensionality reduction and classification system, such as fuzzy rough set model [2], probabilistic rough sets [3], similarity rough set [4], tolerance relation rough set[5], decision-theoretic rough sets [6], covering rough set [7], dominance approximation

*S.Udhaya Kumar Email:uk2804@gmail.com

From the extensions of classical rough set, neighborhood rough set (NRS) is anintelligent system to

Peer-review under responsibility of organizing committee of the Graph Algorithms, High Performance Implementations and Applications (ICGHIA2014)
doi:10.1016/j.procs.2015.03.216

observe as a specified implementation of the neighborhood granular system. The neighborhood rough set can deal with both discrete and continuous data sets by using θ-neighborhood relation. NRS model has been applied in feature selection [9, 13], but it quiet undergoes the low computation performance of the given datasets. In this paper, we present neighborhood rough set based classification (NRSC) algorithm for handling medical diagnosis system.The NRS model describes the neighborhood connection among each two instances based on Euclidean metric function between those two records less than θ (θ>0) with θ- neighborhood relation.

1.1. Problem statement

Computer-based intelligent system is an important and sensational domain for many medical applications and also provides significant support for disease diagnosis. The identification of the exact diseases prediction approach depends upon doctor's experience but the increasing of the new diseases and corresponding medicines, it's quite difficult task to keepup-to-date medical information (diseases and medicines). To reduce deaths due to diseases, we need an initial diagnosis and prognosis, which necessitates an exact and a consistent diagnostic technique [15].There are various data mining techniques developed for medical diagnosis task. For example, neural network [16, 17], hybridized rough set and PSO [18, 19, 20], fuzzy learning vector quantization networks [21], Association rules [22], Principal Component Analysis and Radial Basis Function Neural Network [23], Rough set [24], fuzzy soft set [25], modified soft-rough [26], Bijective soft set theory [27, 28, 29], and hybrid rough set – Bijective soft set [30] have been used as classifier system for several medical applications. The utilization of computerized medical decision support systems turn out to be a practicable approach to assist doctors to quickly and perfectly diagnose patients [31]. So that, the main contribution of this paper is to exploit the neighborhood rough set classification system for diagnosis of five medical data sets.

The remaining chapters of this paper are structured as Sections 2–6. In section 2, related notations and definitions of the NRS presented. Section 3 explains various steps of the proposed methodology. The proposed novel neighborhood rough set based classification algorithm (NRSC) is presented in Section 4. Section 5 reports the results of experimental evaluations and comparisons of the proposed algorithms to other classifiers and finally, in Section 6, we conclude the paper and discuss directions for future research.

2. Neighborhood rough set

The Neighborhood Rough Set (NRS) is used to replace the equivalent approximation of traditional rough set model with neighborhood relation, which supports both continuous and discrete datasets. In this section, we present various essential concepts in NRS used in this work [9, 10, 11, 12, 13].

2.1. θ − Neighborhood relation:

The complete NRS model works based onθ − neighborhood relation which uses distance metric functions (Euclidean distance) based neighborhood relation to replace the equivalence relation in traditional rough set. The Euclidean distance metric function is defined as:

$$f(x_i, x_j) = \sum_{k=1}^{n} \sqrt{(x_i + x_j)}$$

Definition 1. To the nonempty set, universe $U = \{x_1, x_2, x_3, \dots, x_n\}$, metric function satisfies:

(1) Non-negative: $f(x_i, x_j) \geq 0$, if $x_i = x_j$, then $f(x_i, x_j) = 0$;
(2) Symmetry: $f(x_i, x_j) = f(x_j, x_i)$;
(3) Triangle inequality: $f(x_i, x_j) \leq f(x_i, x_k) + f(x_j, x_k)$

Form the above notation f *is* defined as metric function of universe$U$, and $\langle U, f \rangle$ is the neighborhood relation.

Definition 2. Assuming $\langle U, f \rangle$ as the neighbourhood relation,$\forall x_i \in U, \theta \geq 0$, having

$$\theta(x_i) = \{x | f(x_i, x) \leq \theta, x \in U\}$$

Then,$\theta(x_i)$is the θ- neighbourhood relation set of x_i.

2.2. Neighborhood decision system (NDS):

In general, a decision system can be denoted as $\langle U, A \cup D \rangle$, U is universe (set of records), A and D is respectively the conditional and decision attribute set, based on $\theta(\theta > 0)$ condition A will generate a $\theta -$ neighborhood relation$N$. Then basic decision system called as$\theta -$ neighborhood decision system is represented as $NDS = \langle U, A \cup D, \theta \rangle$.

Definition 3. In$NDS = \langle U, A \cup D, \theta \rangle$, B is a subset of $A(B \subseteq A)$, for arbitrary $X \subseteq U$, two sets of records, called lower and upper approximations of X in terms of relation N with respect to B, are defined as:

$$\underline{N_B}X = \{x_i | \theta_B(x_i) \subseteq X, x_i \in U\}$$
$$\overline{N_B}X = \{x_i | \theta_B(x_i) \cap X \neq \emptyset, x_i \in U\}$$

Here $\theta_B(x_i)$ is calculated as follows:

$$\theta_B(x_i) = \{x | f(B(x_i), B(x)) \leq \theta, x \in U\}$$

and $B(x)$is a function to extract the sub vector from x, only the dimensions whose attributes are contained in the attribute set B will be chosen. That is$B(x) = \{ a(x) | \forall a \in B\}$. The boundary region of the decision D with respect to attributes B is defined as

$$BNR(D) = \overline{N_B}X - \underline{N_B}X$$

The union of the lower approximation of each Dclass is called as neighbourhood lower approximation. The neighbourhood lower approximation of decision is also called as positive region. The boundary region is used for reducing uncertainty in decision making process.

Example 1. A sample dataset, containing of both continuous and discrete attributes is presented in Table 1. $A\{a_1, a_2\}$ is conditional attributes and $D\{D_1, D_2\}$is decision attribute.

Table 1
Sample dataset of both continuous and discrete data

Records $x_i \in U$	a_1	a_2	D_i
x_1	0.20	1	1
x_2	0.24	2	1
x_3	0.85	1	2
x_4	0.69	2	2
x_5	0.74	2	1
x_6	0.72	2	2
x_7	0.48	1	1
x_8	0.52	1	1

Here we compute the distances metric values with each records of attributes$A\{a_1, a_2\}$ respectively and neighborhood system of records with $\theta = 0.1$.

The Attribute $a_1$with $\theta = 0.1$, the neigborhood relations are,

$$\theta(x_1) = \{x_1, x_2\}; \theta(x_2) = \{x_2\}; \theta(x_3) = \{x_3\}; \theta(x_4) = \{x_4\};$$
$$\theta(x_5) = \{x_5, x_6\}; \theta(x_6) = \{x_5, x_6\}; \ \theta(x_7) = \{x_7, x_8\};$$

For attribute a_2, Equivalence relation are computed as follows:

$$U/a_2 = \{\{x_1, x_3, x_7, x_8\}, \{x_2, x_4, x_5, x_6\}\};$$

At the same time, we can also divide the decision attributes based on decision classes into two subsets using equivalence relations:

$$D_1 = \{x_1, x_2, x_5, x_7, x_8\};$$
$$D_2 = \{x_3, x_4, x_6\};$$

Apply " ∧ " operator applied for neighborhood granules of records x induced by a_1 *and* $a_2$are listed as follows:

$$\theta(x_1) = \{x_1\}; \theta(x_2) = \{x_2\}; \theta(x_3) = \{x_3\}; \theta(x_4) = \{x_4\};$$
$$\theta(x_5) = \{x_5, x_6\}; \theta(x_6) = \{x_5, x_6\}; \ \theta(x_7) = \{x_7, x_8\};$$

Then we approximate (lower and upper) D $\{D_1, D_2\}$ with the neighborhood granules brought by attributes $\{a_1 \wedge a_2\}$, we get

$$\underline{N}D_1 = \{ x_1, x_2, x_7, x_8\} \qquad \overline{N}D_1 = \{ x_1, x_2, x_5, x_6, x_7, x_8\}$$
$$\underline{N}D_2 = \{x_3, x_4\} \qquad \overline{N}D_2 = \{ x_3, x_4, x_5, x_6\}$$

3. Proposed Methodology

The complete methodology adopted in this work for the diagnosis of medical data sets is exposed in Fig. 1. First stage of the proposed system is the process of medical data sets gathered using various computerized devices.Real time medical data acquisition is quite difficult and extraordinary domain in data mining and machine learning. In this work, we acquired medical data sets form UCI machine learning repository. The basic information of the medical data sets is explained in section 5. In second stage, data sets are separated into training and testing based on *K*-fold cross validation. After the separation of data sets, neighborhood rough set is applied to training data set based medical data sets classification, NRSC algorithm is explained in section 4. In the next step, generated decision rules are matched with test data set for validating the proposed classifier algorithm. The various classification measures are applied to evaluate the performance of the proposed classification algorithm and the proposed approach is also compared with various classification algorithms for medical diagnosis.

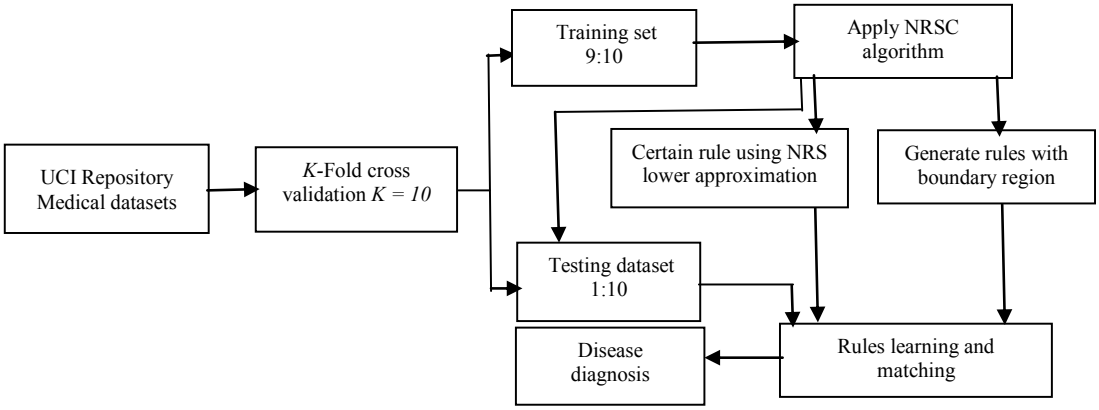

Fig. 1. Block diagram of proposed method

4. Proposed Neighborhood rough set based classification algorithm

The NRS classification algorithm is presented in Fig. 2. There are two importance causes given impression of the proposed algorithm. One is the distance metric function and other one is value of θ – neighborhood relation. The Euclidean distance function defines the shape of neighborhoods and value of θ – neighborhood relation is control the overall system. Neighborhood rough sets are called as generalization of traditional rough set, whenever the size of the neighborhood relation is equal to zero. When the relation is equal to zero, the neighborhood granules act as equivalence relation [11, 12, and 13].

As to classification, we are given a set of instance described with a $m \times n$ matrix $[x_{ij}]$, where $x_{ij}$is the j^{th} feature values of instance x_i. Generally, supervised data sets can be written as $\langle U, A \cup D \rangle$, where $U = \{x_1, \dots, x_n\}, A = \{a_1, \dots, a_n\}$ is the set of conditional attributes, $D = \{d_1, \dots, d_n\}$ is the decision attributes.

Fig. 2: Proposed Algorithm: Neighborhood Rough set based classification

Input: $\langle \boldsymbol{U}, \boldsymbol{A} \cup \boldsymbol{D} \rangle$, $\boldsymbol{\theta}$ // The size of Neighborhood relation

Output: Set of Decision Rules

Step 1: Construct the neighborhood relation for the conditional attributes using Definition 1 and 2 (See. Example 1).

Step 2: Construct the equivalence relation for the decision attribute.

Step 3: Apply " ∧ " (“and”) operator of the neighborhood granules of records of $\boldsymbol{U}$ brought by conditional attributes.

Step 4: Construct the neighborhood rough set lower approximation space for decision attributes and the result of neighborhood granules after applying" ∧ " to conditional attributes. (See. Example 1)

$$\underline{N_B}X = \{x_i | \theta_B(x_i) \subseteq X, x_i \in U\}$$

Step 5: Construct the neighborhood rough set upper approximation space for decision attributes and the result of neighborhood granules after applied " ∧ " to conditional attributes. (See. Example 1)

$$\overline{N_B}X = \{x_i | \theta_B(x_i) \cap X \neq \emptyset, x_i \in U\}$$

Step 6: Find the boundary region of data set by using neigborhood rough set boundary region.

$$BNR(D) = \overline{N_B}X - \underline{N_B}X$$

Step 7: Generate the certain rules using Neighborhood rough set based lower approximation space.

Step 8: Generate the possible rules using Neighborhood rough set based upper approximation space.

Step 9: Generate the boundary rules by using neighborhood rough set boundary region.

The initial steps of the proposed classification algorithm is compute the neighborhood relation of the conditional attributes by using Definition 1, 2 and compute equivalence classes of decision attributes. Distance metrics values are calculated based on Euclidean distance and the size of the neighborhood is $\theta = 0.1$.Apply "and" operator to neighborhood granules of conditional attributes. Furthermore, construct NRS lower and upper approximations to the result of after applied "and" operator and decision classes. The NRS lower approximation space of the decision attributes is definite as the union of the lower approximation space of each decision class. Find the boundary region, boundary region of the subset is come from more than one decision class. With the help of neighborhood approximation we generated two types of decision rules: certain rules (deterministic rules) and possible rules (non-deterministic rules). The certain rules are generated by applying lower approximation and possible rules are generated by applying upper approximation of neighborhood rough set. The proposed algorithm is explained with example in Fig. 3. Fig.3 depicts the example for the proposed algorithm provided in Fig.2.

Fig. 3 : Example for the Proposed algorithm
The following rules are extracted for example 1.

Input: $\langle U, A \cup D \rangle$, θ // The size of Neighborhood relation

Output: Set of Decision Rules

Step 1: Construct the neighborhood relation for the conditional attributes (See. Example 1).

Step 2: Construct the equivalence relation for the decision attribute.

Step 3: Apply " ∧ " ("and") operator of the neighborhood granules. (See. Example 1).

Step 4: Construct the neighborhood rough set lower approximation space for decision attributes D_1 andD_2 $respectively$.

$$\underline{N}D_1 = \{x_1, x_2, x_7, x_8\}$$

$$\underline{N}D_2 = \{x_3, x_4\}$$

Step 5: Construct the neighborhood rough set upper approximation space for decision attributes D_1 andD_2 respectively.

$$\overline{N}D_1 = \{x_1, x_2, x_5, x_6, x_7, x_8\}$$
$$\overline{N}D_2 = \{x_3, x_4, x_5, x_6\}$$

Step 6: Find the boundary region of data set by using neigborhood rough set boundary region.

$$BNR(D) = \{x_5, x_6\}$$

Step 7: Generate certain rules using Neighborhood rough set based lower approximation.

If a_1 = 0.20 and a_2 = 1 then D = 1If a_1 = 0.24 and a_2 = 2 then D = 1

If a_1 = 0.48 and a_2 = 1 then D = 1If a_1 = 0.50 and a_2 = 1 then D = 1

If a_1 = 0.85 and a_2 = 1 then D = 2If a_1 = 0.69 and a_2 = 2 then D = 2

Step 8: Generate the possible rules using Neighborhood rough set based Boundary region.

If $\boldsymbol{a_1}$ = 0.74 and $\boldsymbol{a_2}$ = 2 then D = 1

If $\boldsymbol{a_1}$ = 0.20 and $\boldsymbol{a_2}$ = 2 then D = 1

5. Experimental result and discussion

The simulation of proposed algorithm and benchmark algorithms used for comparison were performed using an Intel (R) Core (TM) i3 CPU 2330M–2.20 GHz machine with 4 GB RAM and a Microsoft Windows 7 64-bit operating system. The essential of the NRSC algorithm calculations was implemented using the MATLAB software package (MATLAB R2013b).

5.1. Medical Data set descriptions

The applicability of NRSC classification is validated in publicly available real-world medical data sets. In this paper, we used five different medical data sets acquired from the well-known UCI repository. The medical data sets are Pima diabetes database [32, 34], heart disease (echocardiogram data) [33], Breast cancer data set [28, 32], liver disorder [28] and Hepatitis's [28, 32]. Pima diabetes database contains 768 instance, 7 attributes and 2 decision classes (negative and positive). In this diabetes dataset, all patients were females with $age \geq 21$and its disturbed with appearance or nonappearance of diabetes. Echocardiogram dataset contains 132 instance, 13 attributes and 2 decision classes (Alive or Not alive). All samples of echocardiogram affected by heart attack, some patients are died and some are still alive. The major problem statement of this dataset is to exactly classify the NOT Alivepatient's samples. The Wisconsin breast cancer data set contains 699 samples, 11 attributes and 2 decision classes (Benign and Malignant). Liver disorder data set covers 345 instances, 7 attributes and two decision classes (Sick or Normal). The major cause of the liver disorder is alcohol intake, so initially patient's bloods are need for further diagnosis purpose. In this liver disorder data set initially 5 types of blood test are taken and recorded in first 5 attributes. The hepatitis's contains 155 instance, 19 attributes and two classes (die and live). The k-fold cross validation (CV) method is applied for evaluate the classification accuracy of test results. The k-fold CV method is extensively used by many researchers with the purpose of random sampling of the training. Initially, all the data's form database randomly separated to k equally select and almost same size subsets. Furthermore, the classification algorithm is trained and tested k times. In this paper, we defined k as 10, so data is divided into ten subsets.

5.2. Performance analysis

The performance of the proposed Neighborhood rough set based classification algorithm is compared with traditional Pawlak's rough set (RS), K-nearest neighbor algorithm (KNN), Back propagation algorithm (BPN), Multilayer perceptron (MLP) and support vector machine (SVM). The obtained results of above classification algorithms are validated based on classification validation accuracy measures. Validation is important for the classification of medical data sets because accurate classification and decision making system is very important in data mining and medical diagnosis. There are many validation methods available for evaluating the accuracy of classification algorithm. In this paper, we demonstrated the performance of classification algorithm using the most familiar metrics such as Precision, Recall, F-Measure, and some other validation measures such as Folke-mallows Index, Kulcznski Index and rand Index [35]. These various measures are presented in Table 2.

Sensitivity or recall is a measure of the ability of a prediction model to select instances of a certain class from a dataset. Positive predictive value is the amount of positive test results that are true positives (correct diagnoses). It is a critical measure of the performance of an analytical method, as it reproduces the probability that a positive test reflects the underlying conditions. Table 3shows the performance of proposed neighborhood rough set classification algorithm and comparative algorithms for the five medical data sets.

Table 2: Classification algorithms performance validation Measures used in this paper.

$$(1)\ \boldsymbol{Precision\ (or)\ Positive\ predictive\ value} = \frac{\boldsymbol{TP}}{\boldsymbol{(TP+FP)}} \quad (2)\ \boldsymbol{Recall\ (or)\ Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{(TP+FN)}}$$

$$(3)\ \boldsymbol{F-Measure} = \frac{\boldsymbol{(2 * Precision * Recall)}}{\boldsymbol{(Precision + Recall)}}$$

$$(4)\ \textbf{Folkes - Mallows index} = \sqrt{\textbf{Precision} * \textbf{Recall}}$$

$$(5)\ \boldsymbol{Kulczynski} = \frac{\mathbf{1}}{\mathbf{2}}(\textbf{Precision} + \textbf{Recall})$$

$$(6)\ \textbf{Rand index} = \frac{(\textbf{True positive} + \textbf{False negative})}{(\textbf{True positive} + \textbf{True negative} + \textbf{False positive} + \textbf{False negative})}$$

Table 3: Performance analysis of proposed classification algorithm and other comparative algorithm

Medical data set	Classification algorithms	Precision	Recall	F-Measure	Folkes-Mallows index	Kulczynski index	Rand index
Pima Indian diabetes	NRSC	0.9730	0.9620	0.9674	0.9678	0.9675	0.9824
	RS	0.8597	0.6362	0.6825	0.6850	0.7479	0.7461
	KNN	0.6245	0.6348	0.6264	0.6280	0.6297	0.6543
	BPN	0.8423	0.7788	0.7942	0.8023	0.8106	0.8218
	MLP	0.7480	0.7510	0.7489	0.7494	0.7495	0.7779
	SVM	0.7024	0.7039	0.7031	0.7031	0.7032	0.7321
Heart disease	NRSC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	RS	0.9386	0.8542	0.8819	0.8861	0.8964	0.9054
	KNN	0.8173	0.8458	0.8192	0.8248	0.8163	0.8478
	BPN	0.8561	0.8148	0.8311	0.8456	0.8369	0.8402
	MLP	0.8230	0.8230	0.8230	0.8230	0.8230	0.3470
	SVM	0.7537	0.7536	0.7537	0.7538	0.7537	0.8095
Breast Cancer	NRSC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	RS	0.9412	0.8734	0.8964	0.9018	0.9073	0.9127
	KNN	0.7189	0.7268	0.7228	0.7288	0.7310	0.6144
	BPN	0.9194	0.9197	0.9193	0.9194	0.9195	0.9192
	MLP	0.9531	0.8564	0.9021	0.9034	0.9047	0.9028
	SVM	0.8054	0.8502	0.7965	0.8120	0.8278	0.8025
Liver Disorder	NRSC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	RS	0.8278	0.669	0.6558	0.7016	0.7534	0.7217
	KNN	0.6148	0.6092	0.6089	0.6104	0.6120	0.6204
	BPN	0.6970	0.6934	0.6947	0.6949	0.6952	0.7027
	MLP	0.6782	0.6584	0.6685	0.6681	0.6682	0.6683
	SVM	0.6444	0.6529	0.6385	0.6436	0.6487	0.6431
Hepatitis	NRSC	0.9560	0.9832	0.9694	0.9699`	0.9696	0.9786
	RS	0.7381	0.5470	0.6283	0.4984	0.6426	0.5302
	KNN	0.4745	0.4847	0.4170	0.4469	0.4810	0.4960
	BPN	0.6537	0.6505	0.6377	0.6461	0.6546	0.6400
	MLP	0.7180	0.6504	0.6825	0.6064	0.6842	0.6928
	SVM	0.4872	0.5183	0.5014	0.5000	0.5264	0.5012

Table 3 represents the classification accuracy of rough set is higher than of MLP, SVM, BPN and KNN for Pima Indian diabetes medical dataset. It also shows the effectiveness of proposed NRSC algorithm over the rough set classification algorithm. The proposed NRSC algorithm produces 96.75% accuracy and RS, KNN, BPN, MLP, SVM algorithms accuracies are 73.15%, 62.64%, 69.42%, 71.89%, 70.31% respectively. In heart disease data sets exact prediction of "NOT Alive" is very important. The experiment of heart disease data set the proposed NRSC algorithm provide 100% correctly identified heart failure patients "NOT Alive" and NRSC algorithm is over than other comparative classification algorithms. The table also represent the accuracy of BPN is higher than of MLP, RS, KNN and SVM for breast cancer and liver disorder datasets. The proposed algorithm provides 100% accuracy and it higher than BPN for both breast and liver disorder datasets. Hepatitis dataset's accuracy of MLP is higher thanthe accuracy of BPN, RS, SVM and KNN. It also shows the effectiveness of proposed NRSC algorithm over

MLP classification algorithm. The proposed NRSC algorithm produces 96.94% accuracy and the accuracy of MLP, BPN, RS, SVM and KNN are 68.25%, 63.77%, 62.83%, 50.14%, and 41.70% respectively.

5.3. Discussion

Medical data classification is a major element of the many decision-making tasks. Decision-making tasks are instances of classification problem that can be easily formulated into a prediction tasks, diagnosis and pattern recognition.To avoid the risks of decision-making, we need computerized decision making techniques. In this work, we proposed the neighborhood rough set based classification for medical diagnosis. In previous study, neighborhood rough set theory was widely applied for feature selection and none of the approaches have been adoptedcompletely for medical diagnosis. The proposed work is applied for five medical data sets and evaluated the efficiency of NRSC compared with five different classification algorithms. The efficiency of the classification algorithm is validated from six performance measures. The performance resultsconfidently demonstrated that the neighborhood rough set based classification method is very effective for medical data classification. Furthermore, the NRSC delivered good results over Pawlak's rough set ($\theta = 0$), BPN, MLP, SVM, KNN. The result is intensely important in decision support tasks not only for exact prognostic or diagnostic prediction, but also would like to be convinced that the prediction is based on reasonable justifications; hence accepting the use of proposed classification system in clinical practice.

6. Conclusion and future work

The neighborhood rough set based classification algorithm has been demonstrated as a valuable method to solve the medical diagnosis problems. The medical big data analysis has become the widespread domain of pattern recognition. This paper presents neighborhood rough set based classification (NRSC) algorithm for classification of five medical data sets. Different experiments are carried out to evaluate the performance of the proposed classification algorithm and the propose approach is compared with different classification models. The acquired result illustrates that the proposed method performs relatively better than other classification algorithms. The experimental results strongly suggest that it can aid in the diagnosis of above five medical datasets and it can be very useful to the doctors for exact decision making task. The future investigations will be applied to evaluate NRS classifier with other medical diagnosis applications and Bio-signal such as ECG, EMG, EEG and EOG.

Acknowledgement

The first author immensely acknowledges the partial financial assistance under University Research Fellowship, Periyar University, Salem. The Second author would like to thank UGC, New Delhi for the financial support received under UGC Major Research Project No. F-41-650/2012 (SR).

Reference

1. Z. Pawlak, "Rough sets", *International Journal of Computer and Information Sciences,* vol. 11, pp. 341–356, 1982.
2. Q. Hu, D. Yu, Z. Xie, J. Liu, "Fuzzy probabilistic approximation spaces and their information measures", IEEE Transaction of Fuzzy System, vol. 14, pp. 191–201, 2006.
3. Y. Yao, "Probabilistic rough set approximations", *International Journal of Approximate Reasoning*, vol. 49, pp. 255–271, 2005.
4. R. Slowinski, D. Vanderpooten, "A generalized definition of rough approximations based on similarity", *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, pp. 331–336, 2000.
5. A. Skowron, J. Stepaniuk, "Tolerance approximation spaces", *Fundamenta Informaticae,* vol. 27, pp. 245–253, 1996.
6. Y. Yao, Y. Zhao, "Attribute reduction in decision-theoretic rough set models", *Information Sciences*, vol. 178, pp. 3356–3373, 2008.
7. Y. Yao, B. Yao, "Covering based rough set approximations", *Information Science*, vol. 200, pp. 91–107, 2012.
8. S. Greco, B. Matarazzo, R. Słowin´ ski, "Rough approximation of a preference relation by dominance relations", *European Journal of Operational Research*, vol. 117, pp. 63–83, 1999.
9. Q. Hu, D. Yu, J. Liu, C. Wu, "Neighborhood rough set based heterogeneous feature subset selection", *Information Science,* vol. 178, pp. 3577–3594, 2008.
10. Q. Hu, D. Yu, Z. Xie, "Neighborhood classifiers", *Expert System with Application,* vol. 34, pp. 866–876, 2008.
11. Y. Du, Q. Hu, P. Zhu, P. Ma, "Rule learning for classification based on neighborhood covering reduction", *Information Science,* vol.

181, pp. 5457–5467, 2011.
12. T.Y. Lin, "Granulation and Nearest Neighborhoods: Rough Set Approach", *Granular Computing: An Emerging Paradigm*, vol. 70, pp. 125–142, 2001.
13. L. Yong, H. Wenliang, J. Yunliang, Z. Zhiyong, "Quick attribute reduct algorithm for neighborhood rough set model", *Information Science,* vol. 271, pp. 65-81, 2014.
14. D. Slezak, W. Ziarko, "The investigation of the Bayesian rough set model", *International Journal of Approximate Reasoning*, vol. 40, pp. 81–91, 2005.
15. P. Meesad, G.G. Yen, "Combined numerical and linguistic knowledge representation and its application to medical diagnosis", *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 33, pp. 206–222, 2003.
16. A.T. Azar, A.E. Hassanien, "Dimensionality reduction of medical big data using neural-fuzzy classifier", *soft computing,* 2014.
17. Adlassing Klaus-Peter, "Fuzzy Set Theory in Medical Diagnosis", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 16, 1986.
18. G. Jothi, H.H. Inbarani, A.T. Azar, "Hybrid Tolerance Rough Set: PSO Based Supervised Feature Selection for Digital Mammogram Images", *International Journal of Fuzzy System Applications*, vol. 3, pp. 15-30, 2013.
19. H.H. Inbarani, P.K.N. Banu, A.T. Azar, "Feature selection using swarm-based relative reduct technique for fetal heart rate", *Neural Computing and Applications*, Springer. DOI: 10.1007/s00521-014-1552-x, 2013.
20. H.H. Inbarani, A.T. Azar, G. Jothi, "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis", *Computer Methods and Programs in Biomedicine,* vol. 113, pp. 175-185, 2014.
21. J. S. Lim, "Finding Features for Real-Time Premature Ventricular Contraction Detection Using a Fuzzy Neural Network System", *IEEE Transactions on Neural Networks*, vol. 20, pp. 522–527, 2009.
22. T.P. Exarchos, A.T. Tzallas, D.I. Fotiadis, S. Konitsiotis, S. Giannopoulos, "EEG Transient Event Detection and Classification Using Association Rules", IEEE Transactions on Information Technology in Biomedicine, vol. 10, pp. 451–457, 2006.
23. S. Ghosh-Dastidar, H. Adeli, N. Dadmehr, "Principal Component Analysis-Enhanced Cosine Radial Basis Function Neural Network for Robust Epilepsy and Seizure Detection", IEEE Transactions on Biomedical Engineering, vol. 55, pp. 512–518, 2006.
24. Aboul Ella Hassanien and Jafar ALI, "Rough Set Approach for Generation of Classification Rules of Breast Cancer Data", *Journal of informatics*, vol. 15, pp. 23-38, 2004.
25. N. Kalaiselvi, and H. Hannah Inbarani, "Fuzzy Soft Set Based Classification for Gene Expression Data", International Journal of Scientific & Engineering Research, vol. 3, 2013.
26. S. Senthil kumar, H. H. Inbarani, and S. Udhaya kumar, "Modified Soft Rough set for Multiclass Classification", In: Proceedings of Advances in Intelligent Systems and Computing, Springer-India, vol. 246, pp. 379-384, 2014.
27. S. Udhaya kumar, H. H. Inbarani, and S. Senthil kumar, "Improved Bijective-Soft-Set-Based Classification for Gene Expression Data", In: Proceedings of Advances in Intelligent Systems and Computing, Springer-India, vol. 246, pp. 127-132, 2014.
28. S. Udhaya kumar, H.H. Inbarani, S.S. kumar, "Bijective soft set based classification of Medical data". *In: Proceedings of Pattern Recognition, Informatics and Medical Engineering (PRIME), International Conference*, pp. 517 – 521, 2013.
29. S. Udhaya kumar, H.H. Inbarani, "Classification of ECG Cardiac Arrhythmias using Bijective Soft Set", *Book chapter: Big Data in Complex Systems Challenges and Opportunities*, Springer-Verlag, Germany, vol. 9, 2014.
30. S. Udhaya kumar, H.H. Inbarani, A.T. Azar, A.E. Hassanien, "Identification of Heart Valve Disease using Bijective Soft Sets Theory", International Journal of Rough Sets and Data Analysis (IJRSDA), vol. 1, pp. 1-14, 2014.
31. S.A. Pavlopoulos, A. Delopoulos, "Designing and implementing the transition to a fully digital hospital", *IEEE Transactions on Information Technology in Biomedicine*, vol. 3, pp. 6–19, 1999.
32. K. Bache, M. Lichman, UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine. http://archive.ics.uci.edu/ml.
33. G. Kan, C.A. Visser, J.J. Koolen, A.J. Dunning, "Short and long term predictive value of admission wall motion score in acute myocardial infarction. A cross sectional echocardiographic study of 345 patients", *British Heart Journal*, vol. 56, pp. 422–427, 1986. http://archive.ics.uci.edu/ml/datasets/Echocardiogram.
34. J.W. Smith, J.E. Everhart, W.C. Dickson, W.C Knowler, R.S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care", *IEEE Computer Society Press*, pp. 261—265, 1988. http://archive.ics.uci.edu/ml/
35. Bernard Desgraupes, "Clustering Indices",*University of Paris Ouest - Lab Modal'X*, pp. 1-34,2013.