Neural Comput & Applic
DOI 10.1007/s00521-016-2412-7

Covering-based rough set classification system

S. Senthil Kumar[1] · H. Hannah Inbarani[1] · Ahmad Taher Azar[2] · Kemal Polat[3]

Received: 6 May 2016 / Accepted: 1 June 2016
© The Natural Computing Applications Forum 2016

Abstract Medical data classification is applied in intelligent medical decision support system to classify diseases into different categories. Several classification methods are commonly used in various healthcare settings. These techniques are fit for enhancing the nature of prediction, initial identification of sicknesses and disease classification. The categorization complexities in healthcare area are focused around the consequence of healthcare data investigation or depiction of medicine by the healthcare professions. This study concentrates on applying uncertainty (i.e. rough set)-based pattern classification techniques for UCI healthcare data for the diagnosis of diseases from different patients. In this study, covering-based rough set classification (i.e. proposed pattern classification approach) is applied for UCI healthcare data. Proposed CRS gives effective results than delicate pattern classifier model. The results of applying the CRS classification method to UCI healthcare data analysis are based upon a variety of disease diagnoses. The execution of the proposed covering-based rough set classification is contrasted with other approaches, such as rough set (RS)-based classification methods, Kth nearest neighbour, improved bijective soft set, support vector machine, modified soft rough set and back propagation neural network methodologies using different evaluating measures.

Keywords Rough set · Covering-based rough set (CRS) · UCI healthcare data · Classification · Experimental analysis

1 Introduction

Healthcare data classification systems are used for a variety of healthcare domains such as medical image classification and biological signal classification [1–6]. The key bases of this common unpleasantness are because of the lacking comprehension of biomedical systems and their connections, and the anonymity of healthcare outcomes and extents. Besides, various sicknesses show up in various levels, in mix with other related sicknesses and with various manifestations of mutable expansion and arrangement. Subsequently, it would be fundamental and extremely gainful to guarantee quick, precise and important analysis for various across-the-board and deadly diseases. That would furthermore enhance the adequacy of healthcare treatment and in addition the quickness and accuracy of the cure response, distressing the recuperation and future of the ill human being and the functioning proficiency of the medicinal elements. If we additionally consider the increasingly growing amount of various collected medical data, we can easily appreciate the necessity of its categorization and the expediency of such a classification framework. The regular development of different

✉ Ahmad Taher Azar
ahmad_t_azar@ieee.org; ahmad.azar@fci.bu.edu.eg

S. Senthil Kumar
pkssenthilmca@gmail.com

H. Hannah Inbarani
hhinba@gmail.com

Kemal Polat
kemal_polat2003@yahoo.com

[1] Department of Computer Science, Periyar University, Salem, Tamil Nadu 636 011, India

[2] Faculty of Computers and Information, Benha University, Banha, Egypt

[3] Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Abant Izzet Baysal University, 14280 Bolu, Turkey

Published online: 14 June 2016

sicknesses, the incomprehensible natural surroundings of UCI healthcare data and the fundamental obscurity of healthcare problems necessitate a reliable background that can handle uncertainty by permitting variable and facilitate approximate reasoning. This definitely makes the covering rough set a valued tool for representing medical diagnosis. Our prior proposed two classification methods (IBS and MSR) for medical data sets provide better classification accuracies [7–9].

Rough set hypothesis is an important instrument in information digging for managing instability issues [10–16]. This hypothesis exhibited by Pawlak has been mathematically praiseworthy and discovered applications in these fields of choice examination, information technology, machine learning framework, data revelation in databases and pattern recognition. Generalized rough sets are inspected by utilizing comparability relations. Some extension of rough set is called covering rough set [17]. Covering rough sets are logically analysed by summing up these characterizations. One can expand the all objects that surely belong to the set (i.e. lower approximation) as the union of covering components contained in the set and amplify the all objects that possibly belong to the set (i.e. upper approximation) as the union of covering components intersecting the set. It is multi-mapping items for inadequate information about the genuine characteristic [18–28]. Hybrid rough set systems have been used in different applications for feature selections [29–34], classifications [35–38, 46–48] and image segmentation [39]. Jaganathan et al. [40] suggested amount of feature significance based on fuzzy entropy, experienced with a radial basis function network classifier for classification using five UCI healthcare benchmark data sets. Experimental classification accuracy indicates that the recommended technique is useful for generating a respectable outcome.

Clinical decisions are regularly made focused by medical practitioner. This practice prompts undesirable predispositions, slips and an extreme medical expense which influences the nature of administration given to patients. Pattern recognition can possibly produce an information-rich environment which can help to essentially enhance the nature of clinical choices. In this paper, medical finding is treated as a patient categorization issue and our attempt to match manifestations in contradiction to sicknesses by gaining from clinical information. In this way, we can classify potential patients as per numerical estimations of their indications and their degrees of enrolment in different classes, apply for covering based rough sets. The preparation stage then makes the learning procedure complete by generating all possible rules for classification subsequent to performing attribute importance took after by categorization. Chang et al. [41] proposed hybrid approach of case-based reasoning method and particle swarm optimization (PSO) and applied it for UCI machine learning data sets. This hybrid proposed classification method provides accuracy of breast cancer as 97.4 % and liver disorders as 76.8 %. The test stage decides the accuracy of the classifier when given test information and by surveying the returned class label. It might be comprehended that while average categorization precision gives learning about the average execution of the method over numerous utilized data sets under this examination, the factual test, as talked about underneath, gives a thought regarding the unwavering quality of an approach over another on every test data set. A data set consists of n numbers of objects and m number of features. In the medical data set, each of genuine vectors normally relates to clinical characterization of a patient. The data set is a union of two disjoint sets. One speaks to the "positive" gathering connected with patients having a particular healthcare condition or malady, and the other speaks to the "negative" gathering connected with patients who do not have that condition or infection. Seera and Lim [42] projected hybrid classification method applied for medical data. In this paper, outcome provides effectiveness for medical decision support. Subsequently, despite the fact that the accuracy contrast is little amongst the methods, one might be considered altogether superior to the next. As of late, medical analysis has been appeared to be enhanced by applying information classification in clinical settings.

Dennis and Muthukrishnan [43] offered adaptive genetic fuzzy system (AGFS) by using membership functions and reducing number of rules for UCI healthcare information in the classification process. It generates rules from medical data, and they are also used for optimized rule selection and good classification accuracy. Tomczak and Zieba [49] proposed a new probabilistic combination of soft classification rules for medical data sets and other benchmark data sets, proposed a particle swarm optimization (PSO) technique and extreme learning machine (ELM) for biomedical data classification (breast cancer, hepatitis and Pima Indian) and obtained 95, 98 and 91 % classification accuracy using PSO-ELM and proposed method, respectively. Neshat et al. [50] used particle swarm optimization (PSO) and case-based reasoning (CBR) to classify the hepatitis disorder data set and obtained 93.25 % classification accuracy. Dash et al. [53] proposed a hybrid training algorithm for fuzzy MLP, called fuzzy MLP-GSPSO, which has been proposed by combining two meta-heuristics: gravitational search (GS) and particle swarm optimization (PSO). These suggested models have been applied for classification of medical data. Lin and Hsieh [51] proposed a hybrid system based on endocrine-based particle swarm optimization (EPSO) and artificial bee

colony (ABC) methods in combination with a support vector machine (SVM) for the selection of optimal feature subsets for the categorization of medical data sets. Muhaideb and Menai [52] proposed a novel mixture metaheuristic that is obtainable for the categorization task of medical data sets. The hybrid techniques are ant colony optimization (ACO) and an artificial bee colony (ABC). Kemal Polat and Günes [44] proposed a new fuzzy artificial immune recognition system algorithm for medical data classification and showed an accuracy of 99 % for classification of breast cancer and 84 % for classification of Pima Indians diabetes data set. The traditional rough set hypothesis depends on equivalence relations; however, in a few conditions, similarity relations are not reasonable to adapt granularity, and in this way, numerous viable information sets cannot be taken care of glowing. In light of this, similarity relation has been summed up to equivalence relation and even random dual relation in a few augmentations of the established traditional rough sets. Covering-based rough set (CRS) is an expansion of segment or equivalence relation; be that as it may, this necessity is not fulfilled in a few circumstances. It builds up the equivalence of the unary covering and the covering with the belongings that the intersection of any two components is the union of limited components in this covering. Really, distinctive models might be relevant to various circumstances. A methodology is the relaxation of the partition arising from equivalence relation to a covering. Covering of a universe is utilized to develop the upper and lower approximations of any subset of the universe. We can discover the distinction and association amongst them and provide the categorization of covering rough set, which can support a decision-maker to pick a reasonable rough set method for data examination. The proposed classification method at first perceives approximation spaces of the classes by setting up an arrangement of named data exploring the traits of the recognized upper and lower approximation space for every one of the classes. Right when that happens, the fitting principle is to isolate the most legitimate class executed to perceive the most appropriate class [17–28]. Covering rough set methodology is proposed for the productive classification of medical information. Finally, our proposed method compares the experimental results with the other methods for four UCI machine learning healthcare data. Table 5 shows the diagnostic accuracies achieved by all categorization methods. From the table, it is perceived that the covering-based rough set (CRS) classification system accomplished a superior classification performance than BPN, RS, KNN, IBS, MSR and SVM. There have been numerous relationships of distinctive classification strategies; however, no single classification technique has been discovered to be better over all data sets.

1.1 Problem statement

All around the world, there are numerous individuals who experience the ill effects of numerous diseases. A large portion of those sicknesses are straightforward and simple to be distinguished and analysed by specialists, nonetheless, there are some that are risky, and in the meantime, it is extremely hard to analyse them, bringing about the trouble to recommend a legitimate treatment. Therefore, early discovery of medical problems, for example, breast cancer, Pima Indian diabetes, liver disorders and hepatitis is vital to enhance the chance of an effective treatment. In the commonplace setting, a data set of notable information, which depicts some medicinal characteristics, types of malady or a medical issue, is thought to be accessible. Healthcare data sets comprise of records of patients portraying physical and research centre examinations identified with that types of malady or medical issue. Along these lines, the computational test is the means by which to build up a demonstrative framework that could help the symptomatic this sort of sickness in view of the information removed from the noteworthy medicinal data sets. Medicinal information sets are taken from UCI repository [45].

Medicinal data frameworks in cutting edge healthcare organizations have enlarged bringing on incredible challenges in extricating helpful data for decision support. Instability and imprecision are the most basic issues in medicinal finding according to the expert systems, and in the midst of past decades, various specialists have been composed towards this zone. Various issues in medical indicative zones ought to be addressed at various degrees of determination to be understood. In this way, classification is vital in computerized helped medical analysis. Thus, exactness is extremely vital in classifiers utilized for medical applications. Intelligent tools that are usually being utilized for medicinal determination incorporate rough set-based strategies. As rough set-based classification strategies have some points of interest, for example, simple adjustment to various sorts of information and information structures, and great speculation abilities, it has been effectively utilized as a part of numerous medical applications including pattern classification [46–48] and decision-making.

The motivation of this study is to extract valuable information from medical data sets and therefore determine decision support perceptions for the analysis and treatment of sicknesses. The medical data mining involves pre-treatment of medical data. The main issue in medical data sets is choosing the useful information amongst the large number of data to be classified. Categorization is a simple yet extensively used method, which is the task of assigning entities to one of numerous well-defined classes and is a general problematic that involves many different

applications. To decrease deaths due to sicknesses, an exact and a reliable prediction method is required. There are several pattern classification systems established for medical analysis assignment. For various medical data categorizations, many researchers have been used various machine learning algorithms such as neural network [1, 2], decision tree [3] rough sets theory [14], neuro-fuzzy sets [4, 6], support vector machine [5], extension of rough set methods [36, 37] and extension of soft set [7, 8]. In this paper, a proposed covering-based rough set (CRS) method for classification for medical data sets is applied.

This paper is structured as follows: Sect. 2 discusses preliminaries, Sect. 3 presents the proposed system, Sect. 4 depicts results and discussion, and Sect. 5 discusses conclusion and future work of the proposed work.

2 Preliminaries

2.1 Rough sets

Rough set hypothesis was launched by Pawlak for managing dubiousness and granularity in data frameworks. This hypothesis handles the approximation of a discretionary subset of a universe by two quantifiable or recognizable subsets called lower and upper approximations. It has been successfully applied to various fields such as biomedical field, E-learning systems, scientific mathematical reasoning problems, pattern classification field, image processing domains, signal processing applications, big data analysis, clinical decision analysis and many other fields [10–16].

Definition 1 Let R be an equivalence relation on U. The pair (U, R) is called a Pawlak approximation space. The equivalence relation R is often called an indiscernibility relation. Using the indiscernibility relation R, one can define the following two rough approximations:

$$R_*(x) = \{x \in U : [x]_R \subseteq \mathrm{X}\} \tag{1}$$

$$R^*(x) = \{x \in U : [x]_R \cap X = \theta\} \tag{2}$$

where $R_*(x)$ and $R^*(x)$ are called the Pawlak lower approximation and the Pawlak upper approximation of X, respectively [10–17, 27].

2.2 Covering rough sets

Covering generalized rough set is regarded as a meaningful extension of the classical rough set model to manage more unpredictable viable issues. Covering rough set is an extension of partition or similarity relation. The covering generated by a reflexive and symmetric relation is characterized. It establishes the equivalence of the unary covering and the covering with the property that the intersection of any two elements is the union of finite elements in this covering [18–28].

Definition 2 In the ordered pair (U, C), U is any non-empty set called a universe and C its finite covering (i.e. C is a finite family of non-empty subsets of U), and covering approximation space $\bigcup C = U$, and the covering C is called the family of approximating sets. Let (U, C) be an approximation space, where $C = \{K(x)/x \in U\}$, for any set X $\subseteq$ U, the family of sets.

$$C_*(X) = \{K(x) \in C : K(x) \subseteq \mathrm{X}\} \tag{3}$$

is called the family of sets approximating the set X. $\mathrm{X}_* = \bigcup \mathrm{C}_*(\mathrm{X})$ is called the lower approximation of the set X.

$$C^*(X) = \{\cup\{\cap\{K \in C/x \in K(x)\}/x \in X\} \tag{4}$$

is called family of sets approximating the set X, and the set $\mathrm{X}^* = \mathrm{U}\ \mathrm{C}^*(\mathrm{X})$ is called the upper approximation of the set X.

Example 1 Table 1 depicts the set of cases U = {1, 2, …, 10} and the set of conditional attributes {temperature, body pain, cold} and the decision attribute {fever}. Consider a complete idea of an indiscernibility relation, the following are defined for complete decision tables with some extensions [27].

[Temperature, Low] = {1, 5, 6}
[Temperature, Medium] = {2, 3, 7, 10}
[Temperature, High] = {4, 8, 9}
[Body Pain, No] = {1, 4, 5, 7, 9}
[Body Pain, Yes] = {2, 3, 6, 8, 10}
[Cold, No] = {2, 3, 6, 7}
[Cold, Yes] = {1, 4, 5, 8, 9, 10}
$C_1 = \{1, 5, 6\} \cap \{1, 4, 5, 7, 9\} \cap \{1, 4, 5, 8, 9,10\} = \{1, 5\}$
$C_2 = \{2,3,7,10\} \cap \{2,3,6,8,10\} \cap \{2,3,6,7\} = \{2,3\}$
$C_3 = \{2,3,7,10\} \cap \{2,3,6,8,10\} \cap \{2,3,6,7\} = \{2,3\}$
$C_4 = \{4, 8, 9\} \cap \{1, 4, 5, 7, 9\} \cap \{1, 4, 5, 8, 9, 10\} = \{4, 9\}$

Table 1 Sample data set

U	Temperature	Body pain	Cold	Fever
1	Low	No	Yes	No
2	Medium	Yes	No	No
3	Medium	Yes	No	Yes
4	High	No	Yes	Yes
5	Low	No	Yes	No
6	Low	Yes	No	No
7	Medium	No	No	No
8	High	Yes	Yes	Yes
9	High	No	Yes	Yes
10	Medium	Yes	Yes	Yes

$C_5 = \{1, 5, 6\} \cap \{1, 4, 5, 7, 9\} \cap \{1, 4, 5, 8, 9, 10\} = \{1, 5\}$
$C_6 = \{1, 5, 6\} \cap \{2, 3, 6, 8, 10\} \cap \{2, 3, 6, 7\} = \{6\}$
$C_7 = \{2,7,10\} \cap \{1, 4, 5, 7, 9\} \cap \{2, 3, 6, 7\} = \{7\}$
$C_8 = \{4, 8, 9\} \cap \{2, 3, 6, 8, 10\} \cap \{1, 4, 5, 8, 9, 10\} = \{8\}$
$C_9 = \{4, 8, 9\} \cap \{1, 4, 5, 7, 9\} \cap \{1, 4, 5, 8, 9, 10\} = \{4, 9\}$
$C_{10} = \{2,7,10\} \cap \{2, 3, 6, 8, 10\} \cap \{1, 4, 5, 8, 9, 10\} = \{10\}$
Let $C = \{\{1, 5\}, \{2, 3\}, \{4, 9\}, \{1, 5\}, \{6\}, \{7\}, \{8\}, \{4, 9\}, \{10\}\}$
[Fever No] = {1, 2, 5, 6, 7}
[Fever Yes] = {3, 4, 8, 9, 10}
Covering rough set lower approximation

$$C_*(X) = \{K(x) \in C : K(x) \subseteq \mathrm{X}\}$$

[C_*, No] = {1, 5, 6, 7}
[C_*, Yes] = {4, 8, 9, 10}
Covering rough set upper approximation

$$C^*(X) = \{\cup\{\cap\{K \in C/x \in K(x)\}/x \in X\}$$

[C^*, No] = {1, 2, 3, 5, 6, 7}
[C^*, Yes] = {2, 3, 4, 8, 9, 10}

3 Proposed system

Foremost phase of characterization is performed on healthcare data gathered from UCI Repository [45]. Data gathering is the procedure of captivating information which ought to be satisfactory to the assuming gadget for supplementary preparing equipments. Proposed covering-based rough set (CRS) classification methodology is exhibited in Table 2. In this methodology, CRS lower approximation of the given sample information based on decision attribute X is built in phase 1. In the phase 2, CRS approximation space derived from the sample information based on judgement attribute X is built. In the phase 3, finite set of rules is produced in view of CRS lower approximation. In the phase 4, non-finite set of rules is produced in view of CRS upper approximation. In the phase 5, rule duplication performed from the covering approximation space is removed. In final phase, validation measures using non-deterministic rules are computed.

The classification rules produced using proposed framework for the example as given in Table 1 are given in Table 3.

4 Results and discussion

Categorization is a pattern recognition technique used to guide a healthcare data objects to one of a few pre-defined feature labels. Classification has various applications in every field of human life [1–6]. The goal of categorization is to precisely anticipate the target feature label for every one case in the information. In our trials, four information sets are utilized, accessible in the UCI store website (http:/archive.ics.uci.edu/ml/) [45, 48]. For every information set, 80 % of all illustrations were arbitrarily chosen as training cases and the remaining 20 % as testing ones. The objective of this study is to give a far reaching survey of various

Table 2 Proposed technique

Proposed technique: covering based rough set classification
Input: Specified information table with conditional uncertain features 1, 2,… , *n*-1 and the judging feature *n*.
Output: Generated Decision Rules
Step 1: Compute the equivalence relation for covering objects (i.e. covering objects Ex. $C_1, C_2, C_3, \ldots, C_n$)
Step 2: To build the CRS based lower approximation for the specified information table $\boldsymbol{C_*(X) = \{K(x) \in C: K(x) \subseteq X\}}$
Step 3: To build the CRS based upper approximation for the specified information table $\boldsymbol{C^*(X) = \{\cup\{\cap\{K \in C/x \in K(x)\} / x \in X\}}$
Step 4: Generate the finite set of rules using CRS based $\boldsymbol{C_*(X) = \{K(x) \in C: K(x) \subseteq X\}}$.
Step 5: Generate the non-finite rules using CRS based $\boldsymbol{C^*(X) = \{\cup\{\cap\{K \in C/x \in K(x)\} / x \in X\}}$.
Step 6: To remove rule duplication from the covering approximation
Step 7: To compute validation measures utilizing CRS upper approximation

Table 3 Example for proposed work

A sample information in Table 1 is used as sample data in order to mine the rules.

Input: Conditional features: Temperature, Body Pain and Cold.

Decisions feature: Fever.

Output: Generate decision rules

1. Compute the equivalence relation for covering objects (i.e. covering objects Ex. $C_1, C_2, C_3, \ldots, C_n$)
2. Construct the CRS lower approximation from there covering objects.

$$\boldsymbol{C}_*(\boldsymbol{X}) = \{\boldsymbol{K}(\boldsymbol{x}) \in \boldsymbol{C}: \boldsymbol{K}(\boldsymbol{x}) \subseteq \boldsymbol{X}\} \quad (7)$$

$[C_*, \text{No}] = \{1, 5, 6, 7\}$

$[C_*, \text{Yes}] = \{4, 8, 9, 10\}$

3. Construct the CRS upper approximation from there covering objects

$$\boldsymbol{C}^*(\boldsymbol{X}) = \{\cup \{\cap \{\boldsymbol{K} \in \boldsymbol{C} / \boldsymbol{x} \in \boldsymbol{K}(\boldsymbol{x})\} \ / \ \boldsymbol{x} \in \boldsymbol{X}\} \quad (8)$$

$[C^*, \text{No}] = \{1, 2, 3, 5, 6, 7\}$

$[C^*, \text{Yes}] = \{2, 3, 4, 8, 9, 10\}$

4. Generate certain rules using CRS lower approximation

If temperature=Low, Body Pain=No, Cold=Yes , =>No

If temperature=Low, Body Pain=No, Cold=Yes , =>No

If temperature=Low, Body Pain=Yes, Cold= No, =>No

If temperature=Medium, Body Pain=No, Cold=No , =>No

If temperature=High, Body Pain=No, Cold=Yes , =>Yes

If temperature=High, Body Pain=Yes, Cold=Yes , =>Yes

If temperature=High, Body Pain=No, Cold=Yes , =>Yes

If temperature=Medium, Body Pain=Yes, Cold=Yes , =>Yes

5. Generate possible rules using CRS upper approximation

if temperature=Low, Body Pain=No, Cold=Yes , =>No

If temperature=Medium, Body Pain=Yes, Cold=Yes , =>No

If temperature=Medium, Body Pain=Yes, Cold=Yes , =>Yes

If temperature=Low, Body Pain=No, Cold=Yes , =>No

If temperature=Low, Body Pain=Yes, Cold= No, =>No

If temperature=Medium, Body Pain=No, Cold=No , =>No

If temperature=High, Body Pain=No, Cold=Yes , =>Yes

If temperature=High, Body Pain=Yes, Cold=Yes , =>Yes

If temperature=High, Body Pain=No, Cold=Yes , =>Yes

If temperature=Medium, Body Pain=Yes, Cold=Yes , =>Yes

Table 3 continued

6. Remove rule duplication from the covering approximation space

$[C_*, No] = \{1, 5, 6, 7\} \Rightarrow \{\{1 \Leftrightarrow 5\}, 6, 7\}$

$[C_*, Yes] = \{4, 8, 9, 10\} \Rightarrow \{\{4 \Leftrightarrow 9\}, 8, 10\}$

If temperature=Low, Body Pain=No, Cold=Yes , =>No
If temperature=Low, Body Pain=Yes, Cold= No, =>No
If temperature=Medium, Body Pain=No, Cold=No , =>No
If temperature=High, Body Pain=Yes, Cold=Yes , =>Yes
If temperature=High, Body Pain=No, Cold=Yes , =>Yes
If temperature=Medium, Body Pain=Yes, Cold=Yes , =>Yes

$[C^*, No] = \{1, 2, 3, 5, 6, 7\} \Rightarrow \{\{1 \Leftrightarrow 5\}, 2, 3, 6, 7\}$

$[C^*, Yes] = \{2, 3, 4, 8, 9, 10\} \Rightarrow \{\{4 \Leftrightarrow 9\}, 2, 3, 8, 10\}$

if temperature=Low, Body Pain=No, Cold=Yes , =>No
If temperature=Medium, Body Pain=Yes, Cold=Yes , =>No
If temperature=Medium, Body Pain=Yes, Cold=Yes , =>Yes
If temperature=Low, Body Pain=Yes, Cold= No, =>No
If temperature=Medium, Body Pain=No, Cold=No , =>No
If temperature=High, Body Pain=No, Cold=Yes , =>Yes
If temperature=High, Body Pain=Yes, Cold=Yes , =>Yes
If temperature=Medium, Body Pain=Yes, Cold=Yes , =>Yes

7. Compute validation measures utilizing CRS upper approximation

Table 4 Different validation measures for pattern classification

$$\textbf{Precision} = \frac{\textbf{True positive}}{\textbf{True positive}+\textbf{False positive}}$$

$$\textbf{Sensitivity} = \frac{\textbf{True positive}}{\textbf{True positive}+\textbf{False negative}}$$

$$\textbf{F-Measure (Czekanowski-Dice index)} = \mathbf{2} \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision}+\textbf{Recall}}$$

$$\textbf{Folkes-Mallows index} = \sqrt{\textbf{Precision} \times \textbf{Recall}}$$

$$\textbf{Kulczynski index} = \frac{1}{2}(\textbf{Precision} + \textbf{Recall})$$

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

$$\textbf{Russel-Rao index} = \frac{\textbf{True positive}}{(\textbf{True positive}+\textbf{True negative}+\textbf{False positive}+\textbf{False negative})}$$

five classification approaches that are rough set, SVM, KNN, BPN, improved bijective soft set (IBS) and modified soft rough set (MSR) in pattern classification methods. Keeping this in mind, the end goal is to look at these six approaches based on the accuracy and evaluation parameters (precision, sensitivity, F-measure (Czekanowski-Dice index), Folke-Mallows index, Kulczynski index, Rand index and Russel–Rao index) to table out which one approach is more effective for diseases diagnosis. The performance of the each method is applied on four distinct UCI healthcare information data sets such as liver disorder, breast cancer, Pima Indian diabetes and hepatitis.

4.1 Experimental analysis

A few classification measures are accessible in the example of validation systems. In this paper, seven pattern categorization validation performance indicators were evaluated. There are precision, sensitivity, F-measure (Czekanowski-Dice index), Folke-Mallows index, Kulczynski index, Rand index and Russel–Rao indexes which were applied for assessing the precision of categorization [9, 46, 47]. Precision, sensitivity and Czekanowski-Dice index are universal validation dealings in categorization examination, and remaining four are local validation performance indicators in categorization learning. The greater part of the researchers has connected to only global validation performance indicators. In this study, three global validation performance indicators and four local validation performance indicators are connected to test the proposed classification technique. The different approval validation performance indicators are related to assess the accuracy of projected classification methods for analytic procedure. Table 4 delineates different acceptance measures utilized as a part of this work.

Table 5 Comparative analysis of proposed method and other classification methods

Data sets	Algorithms	Precision	Sensitivity	F-measure	Folke-Mallows	Kulczynski	Rand	Russel–Rao	Over all measure
Liver disorder	RS	0.8278	0.669	0.6558	0.7016	0.7534	0.7217	0.3358	0.6664
	BPN	0.697	0.6934	0.6947	0.6949	0.6952	0.7027	0.3513	0.6704
	KNN	0.6148	0.6092	0.6089	0.6104	0.612	0.6204	0.3102	0.5694
	SVM	0.6444	0.6529	0.6385	0.6436	0.6487	0.6431	0.3216	0.5989
	IBS	0.8365	0.7223	0.7062	0.7110	0.7176	0.7181	0.3412	0.6789
	MSR	0.9708	0.9775	0.9734	0.9738	0.9741	0.9739	0.4870	0.9043
	CRS	0.9876	0.9892	0.9837	0.9841	0.9849	0.9845	0.4883	0.9146
Breast cancer	RS	0.9412	0.8734	0.8964	0.9018	0.9073	0.9127	0.4563	0.8413
	BPN	0.9194	0.9197	0.9193	0.9194	0.9195	0.9192	0.4595	0.8537
	KNN	0.7189	0.7268	0.7228	0.7288	0.7310	0.6144	0.5028	0.6778
	SVM	0.8054	0.8502	0.7965	0.8120	0.8278	0.8025	0.4012	0.7565
	IBS	0.9643	0.9231	0.9502	0.9364	0.9378	0.9311	0.4467	0.8699
	MSR	0.9707	0.9836	0.9766	0.9769	0.9772	0.9785	0.4893	0.9075
	CRS	0.9899	0.9902	0.9874	0.9883	0.9888	0.9863	0.4911	0.9174
Pima Indian diabetes	RS	0.8597	0.6362	0.6825	0.6850	0.7479	0.7461	0.3730	0.6757
	BPN	0.8423	0.7788	0.7942	0.8023	0.8106	0.8218	0.4109	0.7515
	KNN	0.6245	0.6348	0.6264	0.628	0.6297	0.6543	0.3271	0.5891
	SVM	0.7024	0.7039	0.7031	0.7031	0.7032	0.7213	0.3660	0.6575
	IBS	0.8763	0.8124	0.8516	0.8317	0.8388	0.8467	0.4089	0.7809
	MSR	0.8362	0.7928	0.8022	0.7901	0.8114	0.8063	0.4235	0.7517
	CRS	0.8785	0.8432	0.8562	0.8431	0.8448	0.8567	0.4317	0.7934
Hepatitis	RS	0.7381	0.5470	0.6283	0.4984	0.6426	0.5032	0.2516	0.5441
	BPN	0.6537	0.6505	0.6377	0.6461	0.6546	0.6400	0.3200	0.6003
	KNN	0.4745	0.4874	0.4170	0.4469	0.481	0.4960	0.248	0.4358
	SVM	0.4872	0.5183	0.5014	0.5000	0.5264	0.5012	0.2673	0.4716
	IBS	0.7158	0.6656	0.6981	0.6745	0.6792	0.6875	0.3467	0.6382
	MSR	0.8484	0.7357	0.7310	0.7607	0.7920	0.7613	0.3806	0.7156
	CRS	0.8241	0.7765	0.7493	0.7743	0.8003	0.7548	0.3774	0.7223

Table 5 presents results of various approaches for the recognition of various sicknesses and viability of those approaches utilizing different approval measures.

4.2 Discussion

CRS is generally and effectively utilized approach for categorization and decision-making. CRS is proposed for the diagnosing sicknesses. The essential thought about this healthcare information set is to build the projected idea, which will make the conclusion of medicinal information for all intents and purposes. To assess the capability of the CRS system, four standard UCI healthcare data sets, viz. hepatitis, breast cancer, liver disorders and Pima Indian diabetes, are utilized. Various helpful execution measurements in medical applications incorporate precision, sensitivity, F-measure (Czekanowski-Dice index), Folke-Mallows index, Kulczynski index, Rand index and Russel–Rao indexes. The exploratory results emphatically exhibited that the CRS classification approach is compelling undertaking medical information classification assignments. In general, the results show that CRS strategy is healthier than all the other different routines. This CRS can be connected to an assortment of medical information. Trial investigation shows that the proposed system achieves superior than RS, BPN, IBS, KNN, MSR and SVM. Determination of a pertinent methodology to a classification issue can in this way be a troublesome problem. In this manner, there is an incredible potential for the utilization of information digging systems for UCI healthcare data categorization, which has been completely inspected and is a fascinating technique for upcoming exploration.

5 Conclusion and future work

This study looks at current practices, issues and prospects of medical data classification. The attention is put on the rundown of major progressed proposed classification

approaches and the methods utilized for enhancing order exactness. Since researchers have picked up superior and contributed assets to explore apparently interesting data mining applications, it creates the impression that covering rough set strategy has been ended up being fruitful for federation assignments. In this paper, a comparative analysis of the bench mark algorithms in medical data order has been carried out. From these proofs on medical data classification, it can be seen that there is still much space for further change over present medical data order undertakings. More research, notwithstanding, is required to distinguish and diminish instabilities in medical data classification to enhance classification accuracy. The future research will be applied to evaluate CRS classification method with various medical diagnosis applications, other applications and biosignal applications.

References

1. Azar AT (2013) Fast neural network learning algorithms for medical applications. Neural Comput Appl 23(3–4):1019–1034. doi:10.1007/s00521-012-1026-y
2. Azar AT, El-Said SA (2013) Probabilistic neural network for breast cancer classification. Neural Comput Appl 23(6): 1737–1751. doi:10.1007/s00521-012-1134-8
3. Azar AT, El-Metwally SM (2013) Decision tree classifiers for automated medical diagnosis. Neural Comput Appl 23(7–8): 2387–2403. doi:10.1007/s00521-012-1196-7
4. Azar AT, El-Said SA (2013) Superior neuro-fuzzy classification systems. Neural Comput Appl 23(1):55–72. doi:10.1007/s00521-012-1231-8
5. Azar AT, El-Said SA (2014) Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput Appl 24(5):1163–1177. doi:10.1007/s00521-012-1324-4
6. Azar AT, Hassanien AE (2014) Dimensionality reduction of medical big data using neural-fuzzy classifier. Soft Comput 19(4):1115–1127
7. Kumar SS, Inbarani HH, Udhayakumar S (2014) Modified soft rough set for multiclass classification. Adv Intell Syst Comput 246:379–384
8. Udhaya Kumar S, Inbarani HH, Kumar SS (2013). Bijective soft set based classification of medical data. In: International conference on pattern recognition, informatics and medical engineering (PRIME), 1:517–521
9. Udhaya Kumar S, Hannah Inbarani H, Senthil Kumar S (2014) Improved bijective-soft-set-based classification for gene expression data. Adv Intell Syst Comput 246:127–132
10. Pawlak Z (1982) Rough sets. Int J Parallel Prog 11(5):341–356
11. Pawlak Z, Slowinski R (1994) Decision analysis using rough sets. Int Trans Oper Res 1(1):107–114
12. Pawlak Z (1995) Vagueness and uncertainty: a rough set perspective. Comput Intell 11(2):227–232
13. Pawlak Z (1996) Rough sets: present state and the future. Found Comput Decis Sci 18(3–4):157–166
14. Pawlak Z (1999) Rough classification. Int J Hum Comput Stud 51(2):369–383
15. Pawlak Z (2002) Rough sets and intelligent data analysis. J Inf Sci 147(1–4):1–12
16. Pawlak Z, Skowron A (2007) Rough sets and Boolean reasoning. Inf Sci 177(1):41–73
17. Tsang ECC, Degang C, Yeung DS (2008) Approximations and reducts with covering generalized rough sets. Comput Math Appl 56(1):279–289
18. Yang T, Li Q (2010) Reduction about approximation spaces of covering generalized rough sets. Int J Approx Reason 51(3):335–345
19. Ma L (2012) On some types of neighborhood-related covering rough sets. Int J Approx Reason 53(6):901–911
20. Zhu W (2007) Topological approaches to covering rough sets. Inf Sci 177(6):1499–1508
21. Zhu W (2007) Generalized rough sets based on relations. Inf Sci 177(22):4997–5011
22. Zhu W (2009) Relationship between generalized rough sets based on binary relation and covering. Inf Sci 179(3):210–225
23. Zhu W (2009) Relationship among basic concepts in covering-based rough sets. Inf Sci 179(14):2478–2486
24. Yao Y, Yao B (2012) Covering based rough set approximations. Inf Sci 200(2012):91–107
25. Ge X, Bai X, Yun Z (2012) Topological characterizations of covering for special covering-based upper approximation operators. Inf Sci 204(2012):70–81
26. Wang C, Chen D, Sun B, Hu Q (2012) Communication between information systems with covering based rough sets. Inf Sci 216(2012):17–33
27. Medhat T (2012) Missing values via covering rough sets. Int J Data Min Intell Inf Technol Appl (IJMIA) 2(1):10–17
28. Sandeep YS, Reddy PVS, Manoj C, Lakkshmanan KA (2013) Identifying the vague regions by using covering based rough sets. Int J Adv Res Comput Sci Softw Eng 3(7):743–746
29. Inbarani HH, Azar AT, Jothi G (2014) Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Comput Methods Programs Biomed 113(1):175–185
30. Inbarani HH, Banu PKN, Azar AT (2014) Feature selection using swarm-based relative reduct technique for fetal heart rate. Neural Comput Appl 25(3–4):793–806. doi:10.1007/s00521-014-1552-x
31. Jothi G, Inbarani HH, Azar AT (2013) Hybrid tolerance-PSO based supervised feature selection for digital mammogram images. Int J Fuzzy Syst Appl (IJFSA) 3(4):15–30
32. Inbarani HH, Bagyamathi M, Azar AT (2015) A novel hybrid feature selection method based on rough set and improved harmony search. Neural Comput Appl 26(8):1859–1880. doi:10.1007/s00521-015-1840-0
33. Azar AT, Inbarani HH, Devi KR (2016) Improved dominance rough set-based classification system. Neural Comput Appl 2016:1–16. doi:10.1007/s00521-016-2177-z
34. Azar AT, Banu PKN, Inbarani HH (2013). PSORR—an unsupervised feature selection technique for fetal heart rate. In: 5th international conference on modelling, identification and control (ICMIC 2013), 31 August, 1–2 September 2013, Egypt
35. Elshazly HI, Azar AT, Elkorany AM, Hassanien AE (2013) Hybrid system based on rough sets and genetic algorithms for medical data classifications. Int J Fuzzy Syst Appl (IJFSA) 3(4):31–46
36. Kumar S, Inbarani HH, Azar AT, Own HS, Balas VE (2014) Optimistic multi-granulation rough set based classification for neonatal jaundice diagnosis. Adv Intell Syst Comput (Soft Computing Applications) 356:307–317. doi:10.1007/978-3-319-18296-4_26
37. Inbarani HH, Kumar SS, Azar AT, Hassanien AE (2014) Soft rough sets for heart valve disease diagnosis. In: AE Hassanien, M Tolba, AT Azar (eds.) Advanced machine learning technologies and applications: Second International Conference, AMLTA 2014, Cairo, Egypt, November 28–30, 2014. Proceedings, communications in computer and information science, vol 488, Springer GmbH Berlin/Heidelberg. ISBN: 978-3-319-13460-4

38. Banu PKN, Inbarani HH, Azar AT, Hala S. Own HS, Hassanien AE (2014). Rough set based feature selection for Egyptian Neonatal Jaundice. In: AE Hassanien, M Tolba, AT Azar (eds.) Advanced machine learning technologies and applications: Second International Conference, AMLTA 2014, Cairo, Egypt, November 28–30, 2014. Proceedings, communications in computer and information science, vol 488, Springer GmbH Berlin/Heidelberg. ISBN: 978-3-319-13460-4
39. Roy P, Goswami S, Chakraborty S, Azar AT, Dey N (2014) Image segmentation using rough set theory: a review. Int J Rough Sets Data Anal 1(2):62–74
40. Jaganathan P, Kuppuchamy R (2013) A threshold fuzzy entropy based feature selection for medical database classification. Comput Biol Med 43(12):2222–2229
41. Chang PC, Lin JJ, Liu CH (2012) An attribute weight assignment and particle swarm optimization algorithm for medical database classifications. Comput Methods Programs Biomed 107(3):382–392
42. Seera M, Lim CP (2014) A hybrid intelligent system for medical data classification. Expert Syst Appl 41(5):2239–2249
43. Dennis B, Muthukrishnan S (2014) AGFS: adaptive genetic fuzzy system for medical data classification. Appl Soft Comput 25:242–252
44. Polat K, Günes S (2007) An improved approach to medical data sets classification: artificial immune recognition system with fuzzy resource allocation mechanism. Expert Systems 24(4):252–270
45. Lichman M (2013) UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
46. Kumar SS, Hannah Inbarani H (2015) Optimistic multi-granulation rough set based classification for medical diagnosis. Proc Comput Sci 47:374–382
47. Kumar SS, Inbarani HH, Azar AT, Hala SO, Balas VE, Olariu T (2015) Optimistic multi-granulation rough set based classification for neonatal jaundice diagnosis. Adv Intell Syst Comput 356:307–317
48. Gadaras I, Mikhailov L (2009) An interpretable fuzzy rule-based classification methodology for medical diagnosis. Artif Intell Med 47(1):25–41
49. Tomczak JM, Zieba M (2015) Probabilistic combination of classification rules and its application to medical diagnosis. Mach Learn 101:105–135
50. Neshat M, Sargolzaei M, Nadjaran Toosi A, Masoumi A (2012) Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization. In: ISRN Artificial Intelligence, vol 2012
51. Lin KC, Hsieh YH (2015) Classification of medical datasets using SVMs with hybrid evolutionary algorithms based on endocrine-based particle swarm optimization and artificial bee colony algorithms. J Med Syst 119:1–9
52. AlMuhaideb Sarab, Menai MEB (2014) HColonies: a new hybrid meta-heuristic for medical data classification. J Appl Intell 41:282–298
53. Dash T, Nayak SK, Behera HS (2015) Hybrid gravitational search and particle swarm based fuzzy MLP for medical data classification. Comput Intell Data Min 1:35–43