International Journal of Multidisciplinary Research and Development

International Journal of Multidisciplinary Research and Development

Volume :2, Issue :4, 664-668
April 2015
www.allsubjectjournal.com
e-ISSN: 2349-4182
p-ISSN: 2349-5979
Impact Factor: 3.762

Femina B
CSE, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

Anto S
CSE, Sri Krishna College Of Technology, Coimbatore, Tamil Nadu, India

Correspondence:
Anto S
CSE, Sri Krishna College Of Technology, Coimbatore, Tamil Nadu, India

Disease diagnosis using rough set based feature selection and K-nearest neighbor classifier

Femina B, Anto S

Abstract
The use of machine learning tools in medical diagnosis is growing progressively. This is mainly because the use of recognition and classification systems has improved in a great deal to help medical experts in diagnosing diseases. Such disease is called hepatitis. Hepatitis disease diagnosis was conducted using K-Nearest Neighbor (KNN) classifier. The proposed system includes two modules: the feature extraction module and the predictor module. In the feature extraction module, rough set theory is used to preprocess the attributes on condition that the important information is not lost, delete redundant attributes. K-Nearest Neighbor (KNN) classifier is used to classify the given data's. Experiments have been conducted on a widely used Hepatitis dataset taken from University of California Irvine (UCI) machine learning repository dataset. The experimental results show that the proposed system can improve the rate of correct diagnosis. The proposed classifier with rough set-based feature selection achieves 84.52 % of accuracy. Different performance metrics are used to show the effectiveness of the proposed system.

Keywords: Hepatitis disease diagnosis, K-Nearest neighbor Ensemble classifier, Rough set, Feature selection

1. Introduction
Viral hepatitis is one of the most important health problems of the world. It is one of the most frequent infectious diseases .It causing an estimated 1.5 million deaths worldwide each year. Viral hepatitis is an inflammation and damage to hepatocytes in the liver caused by at least six different viruses [1]. These viruses called A, B, C, D, E, and G [2]. Most of the time the hepatitis diagnoses is made by a routine blood testing or during a blood donation. So far, many studies have been performed in the diagnosis of hepatitis diseases. Medical diagnostics is quite difficult and visual task which is mostly done by expert doctors. An expert doctor commonly takes decisions by evaluating the current test results of a patient or the expert doctor compares the patient with other patients with the same condition by referring to the previous decisions. Therefore, it is very difficult for a physician to diagnose hepatitis [3]. For this cause, in modern times, many machine learning and data mining techniques have been considered to design automatic diagnosis system for hepatitis. The automatic diagnosis problem can be approached by using both single and hybrid machine learning algorithm [4, 5].
Medical diagnostic decision support systems have become an established component of medical technology. The main idea of the medical technology is an inductive engine that learns the decision characteristics of the diseases and can then be used to diagnose future patients with uncertain disease states. Typically, ensemble learning involves either statistical parametric classifiers or neural networks trained on the same data, and a method that combines their outputs into a single one. If one could select the best classifier to use for every sample, the misclassified samples in the output would be the ones that were wrongly classified by all methods [6].
The main contribution of the paper is to build intelligent systems which combine two methodologies: rough set theory as a preprocessing step for selecting the most discriminatory features and a combined classifier using K-Nearest Neighbor (KNN) as base classifier so as to automatically produce a diagnostic system. We find that the proposed hybrid approach produces a system exhibiting two prime characteristics: first, it attains high classification performance which is the best shown to date; second, the resulting systems involve a few set of discriminatory features, only four features, and are therefore interpretable.

2 Related works
As for other clinical diagnosis problem, classification systems have been using for breast cancer problem, too. When learning in the literature related with the classification application

were examined, it can be seen that the great variety of methods were used which reached high classification accuracies using the dataset taken from UCI machine learning repository. Amongst these, Quinlan reach 94.74% classification correctness using 10-fold cross confirmation with C4.5 decision tree method [7]. Goodman et al. applied three different methods to the problem which were resulted with the following accuracies: optimized-LVQ method's performance was 96.7%, big-LVQ method reached 96.8% and the last method, AIRS, which proposed to depending on the artificial immune systems, obtained 97.2% classification accuracy [8].

In [9], the combination of further division of partition space (FDPS) and flexible neural tree (FNT) is proposed to improve the neural network classification performance and the obtained result is 98.25 %. A method based on gravitational potential energy between particles [10] is applied for the whole WBCD dataset, and the finest result obtained is 98.81 %. L. Ariel [11] used Fuzzy Support Vector Clustering to classify heart disease. This algorithm applicable a kernel induced metric for conveying each piece of data and experimental results were obtained using a well known benchmark of heart disease. Polat and Gunes [12] considered an expert system to diagnose the diabetes disease based on principal component analysis. Polat *et al.* as well as developed a cascade learning system for diagnose the diabetes.

3 Dataset

Viral hepatitis is an alternate significant health issue around the globe. Hepatitis is the aggravation and damage to hepatocytes in the liver and can be caused by autoimmunity, viruses, infections with fungi and bacteria, or exposure to toxins such as alcohol. Hepatitis diseases can be infected through blood, shared syringes. The hepatitis disease dataset is obtained from the UCI Machine Learning Repository Databases [13]. The dataset's purpose is to predict the presence or absence of hepatitis disease by using the results of various medical tests carried out on a patient. Hepatitis dataset contains 155 instances belonging to two different classes, die with 32 instances and live with 123 instances. The total number of attributes is 19. A brief description of this hepatitis dataset shown in Table 1

Table 1: Description of the hepatitis dataset

Dataset	Number of attributes	Number of instances	Number of classes
Hepatitis	19	155	2

4 Proposed Scheme

K-Nearest Neighbor (KNN) algorithms are known especially with their simplicity in machine learning literature. They are also advantageous in that the information in training data is never lost. But, there are few problems with them. First of all, for large datasets, these algorithms are very time-consuming because each sample in training set is processed while classifying a new data and this requires longer classification times. This cannot be problem for some application areas but when it comes to a field like medical diagnosis, time is very important as well as classification accuracy. So, an attempt has been made in this study to reduce the size of training data. This data-reducing stage was realized by using rough set theory. The most discriminatory set of features are obtained using rough set technique. Then, these features are used in classification phase as input to a combined classifiers using KNN as base classifier. The final results are united using a majority voting (MV) technique. The block diagram of the whole classification system can be seen in Figure 1 which contains feature rough set based feature selection and classification.

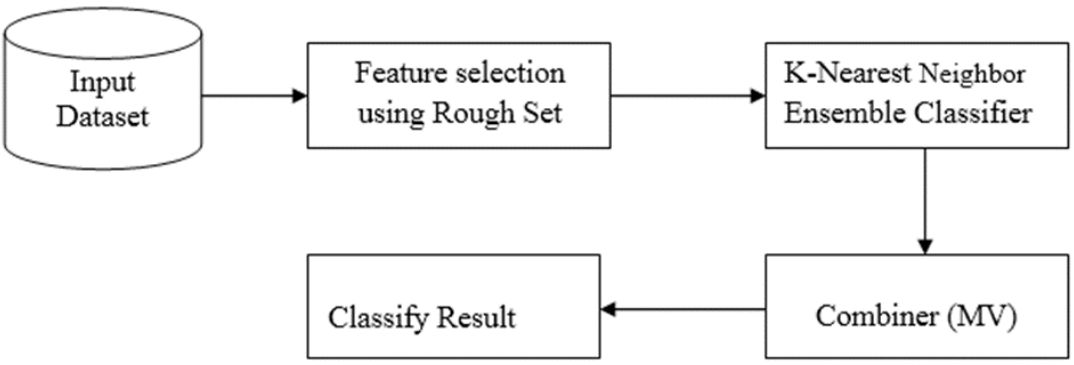

Fig 1: Block diagram of the proposed system

4.1 Feature Selection

Rough Set (RS) theory is a new intelligent mathematical tool proposed by Prof. Pawlak in 1982 to deal with uncertainty and incompleteness [14]. Over the past few years, RST has become topic of great interest to researchers and has been applied to many domains. It is based on the concept of an upper and a lower approximation of a set in the approximation space and models of sets. The main advantage of RS theory is that it does not need any preliminary or additional information about data: like probability in statistics or basic probability assignment in Dempster – Shafer theory and membership grade in fuzzy set theory [15]. One of the major applications of RS theory is the attribute reduction that is possible to find a minimal subset. The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Using the dependency degree to evaluate, attributes are removed and reduced set provides the same dependency degree as the original. This section recalls some essential definitions from RST that are used for feature selection. Detailed description and formal definitions of the theory can be found in [17, 18].

4.1.1 Information system

Knowledge representation in rough sets is done via information system, which is denoted as 4-tuple S = <U, A, V, f>, where U is the closed universe, a finite set of N objects $\{x_1, x_2, \ldots, x_n\}$, A is a finite set of attributes $\{a_1,$

$a_2, \ldots, a_n\}$, which can be further divided into two disjoint subsets of C and D, A = {C ∪ D} where C is condition attributes and D is a set of decision attributes. $V = \cup_{a \in A} V_a$ and V_a is a domain of the attribute a, and f: U × A → V is the total decision function called the information function such that f(x, a) ∈ V_a for every a ∈ A, x ∈ U.

4.1.2 Indiscernibility relation

One of the most significant aspects of RS theory is its indiscernibility relation. The R-indiscernibility relation is denote by IND(R), is defined as:

$$IND(X) = [(x, y) \in U \times U | \forall_a \in R, a(x) = a(y)\} \quad (1)$$

where a(x) denotes the value of attribute a of object x. If (x, y) ∈ IND(R), x and y are said to be indiscernible with respect to R. The equivalence classes of the R-indiscernibility relation are denoted by $[x]_R$. The indiscernibility relation is the mathematical basis of RS theory.

4.1.3 Lower and upper approximation

In RS theory, the lower and upper approximations are two basic operations, for any concept X ⊆ U and attribute set R ⊆ A, X could be approximated by the lower approximation and upper approximation. The lower approximation of X is the set of objects of U that are surely in X, defined as:

$$\underline{R}(X) = \{x \in U | [x]_R \subseteq X\} \quad (2)$$

The upper approximation of X is the set of objects of U that are possibly in X, defined as:

$$\overline{R}(X) = \{x \in U | [x]_R \subseteq X \neq \emptyset\} \quad (3)$$

And the R-boundary region of X is defined as:

$$BND(X) = \underline{R}(x) \quad (4)$$

A set is said to be rough if its boundary region is non-empty, otherwise the set is crisp.

4.1.4 Attribute reduction and core

There often exist some condition attributes that do not provide any additional information about the objects in U in the information system. So, these redundant attributes can be eliminated without losing essential classificatory information. Reduct and core attribute sets are two fundamental concepts of rough set theory. A reduct attribute set is a minimal set of attributes from A (the whole attributes set) that provided that the object classification is the same as with the full set of attributes. Given C and D ⊆ A, a reduct is a minimal set of attributes such that IND(C) = IND(D). Let RED(A) denote all reducts of A. The intersection of all reducts of A is referred to as a core of A, i.e., CORE(A) = ∩ RED(A), the core is common to all reducts.

4.1.5 Dependency degree

Various measures can be defined to represent how much C, a set of decision attributes, depends on D, a set of condition attributes. One of the most common measure is the dependency denoted as cC(D),is defined as: γ_c(D) = |POS_C(D)|/|U| where |U| is the cardinality of set U, POS_C(D) called positive region, is defined by $POS_C(D) = \cup_{x \in \frac{U}{D}} (X)$.

Note $0 \leq \gamma_C(D) \leq 1, If\ \gamma_c(D) = 1$

we say that D depends totally on C, if $0 \leq \gamma_c(D) < 1$, we say that D depends partially on C, and if $\gamma_c(D) = 0$ means that C and D are totally independent of each other.

4.2 K-Nearest neighbor (KNN) Classifier

An instance based learning method called the K-Nearest Neighbor or K-NN algorithm has been used in many applications in areas such as data mining, statistical pattern recognition, image processing. The K-nearest-neighbor (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set.

Suppose each sample in our data set has n attributes which we combine to form an n-dimensional vector: x = (x1, x2, . . . , xn). These n attributes are considered to be the independent variables. Each sample also has another attribute, denoted by y (the dependent variable), whose value depends on the other n attributes x. We assume that y is a categoric variable, and there is a scalar function, f, which assigns a class, y = f(x) to every such vectors. We do not know anything about f (otherwise there is no need for data mining) except that we assume that it is smooth in some sense. We suppose that a set of T such vectors are given together with their corresponding classes: x(i), y(i) for i = 1, 2, . . . , T. This set is referred to as the training set. The problem we want to solve is the following. Supposed we are given a new sample where x = u. We want to find the class that this sample belongs. If we knew the function f , we would simply compute v = f(u) to know how to classify this new sample, but of course we do not know anything about f except that it is sufficiently smooth. The idea in k-Nearest Neighbor methods is to identify k samples in the training set whose independent variables x are similar to u, and to use these k samples to classify this new sample into a class, v. If all we are prepared to assume is that f is a smooth function, a reasonable idea is to look for samples in our training data that are near it (in terms of the independent variables) and then to compute v from the values of y for these samples. When we talk about neighbors we are implying that there is a distance or dissimilarity measure that we can compute between samples based on the independent variables. For the moment we will concern ourselves to the most popular measure of distance: Euclidean distance.

The Euclidean distance between the points x and u is

$$d(x, u) = \sqrt{\sum_{i=1}^{n} (x_i - u_i)^2} \quad (5)$$

We will examine other ways to measure distance between points in the space of independent predictor variables when we discuss clustering methods. The simplest case is k = 1 where we find the sample in the training set that is closest (the nearest neighbor) to u and set v = y where y is the class of the nearest neighboring sample.

Find the nearest k neighbors of u and then use a majority decision rule to classify the new sample. The advantage is that higher values of k provide smoothing that reduces the risk of over-fitting due to noise in the training data. In typical applications k is in units or tens rather than in hundreds or thousands. Notice that if k = n, the number of samples in the training data set, we are merely predicting the class that has the majority in the training data for all samples irrespective of u. This is clearly a case of over-smoothing unless there is no information at all in the independent variables about the dependent variable.

$$h_j(x) = \sum_i I\left(arg_i \max\left(p_{ij}(x)\right) = i\right) \quad (6)$$

In which I(.) is the indicator function defined as follows:

$$I(y) = \begin{cases} 1, & if\ y\ is\ true \\ 0, & otherwise \end{cases} \quad (7)$$

5 Experimental result and discussion

Confusion Matrix: Confusion matrix shows classifications and predicted. A confusion matrix for a classification problem with two classes is of size 2×2, and it is given in Table 2.

Table 2: Confusion Matrix

Predicted	Actual	
	Positive	**Negative**
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

- TP represents an instance, which is actually positive and predicted by the model as positive.
- FN represents an instance, which is actually positive but predicted by the model as negative.
- TN represents an instance, which is actually negative and predicted by the model as negative.
- FP represents an instance, which is actually negative but predicted by the model as positive.

Accuracy is calculated by using equation 8

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad (8)$$

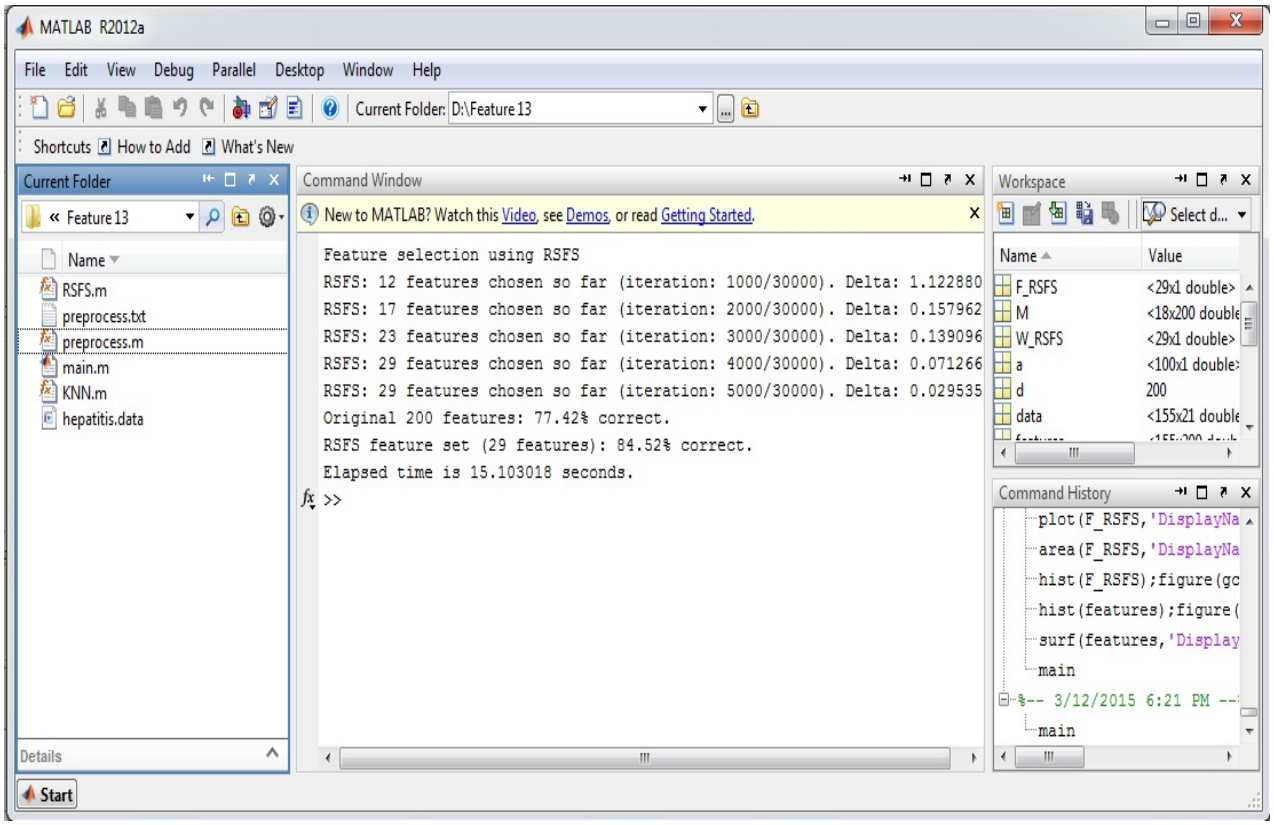

Fig 1: Feature Subsets using Rough Set Algorithm and their accuracy

The experiments are conducted using Intel Pentium IV 2.2 GHz CPU, 512 MB RAM, Windows XP operating system. The development environment is MATLAB 7.14.0.739. In this paper, four diseases are diagnosed by using Matlab software. The UCI Machine Learning Repository Datasets is used, namely, Hepatitis disease dataset.

Table 3: Comparison with existing system

Disease	Methodology	Accuracy (%)
Hepatitis Dataset	SVM [19]	74
	GA-SVM [20]	86.12
	KNN [21]	75.0
	Proposed RS-KNN	**84.52**

The proposed system shows a higher performance with feature subsets at an accuracy of 84.52%. The feature subsets and its accuracy are shown in figure 1. Classification accuracies of the studies in the literature and our proposed system are given in Table 3 for comparison. Performances of all methods given in Table 3 were evaluated on the same Hepatitis dataset taken from the UCI machine learning repository.

6 Conclusion

This study aims at diagnosing Hepatitis disease with an intelligent system. A rough set theory with a combined classifier based on k-nearest neighbor algorithm as base classifier, a method was obtained to solve this diagnosis problem via classifying Hepatitis dataset. This dataset is a very commonly used dataset in the literature relating the use of classification systems for hepatitis disease diagnosis, and it was used in this study to evaluate the classification performance of our proposed intelligent system with regard to other studies. A classification accuracy of 84.52 % is obtained. In future, this system can be used for the diagnosis of real life medical data of patients.

References:

1. W. M. Lee, “Hepatitis B virus infection”, N. Engl. J. Med. 337 (1997) 1733.
2. J. Cohen, “The scientific challenge of hepatitis C”, Science 285 (1999) 26.
3. K. Polat, S. Günes, “Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy

resource allocation", Digital Signal Process. 16 (2006) 889–901.
4. K. Polat, S. Günes¸, "A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS", Comput. Methods Programs Biomed. 88 (2007) 164–174.
5. K. Polat, et al., "Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection", Expert Syst. Applicat. 33 (2007) 484–490.
6. Meynet J, Thiran JP (2010), "Information theoretic combination of pattern classifiers". Pattern Recogn 43(10):3412–3421
7. J.R. Quinlan, "Improved use of continuous attributes in C4.5", J. Artific. Intell. Res. 4 (1996) 77–90.
8. D.E. Goodman, L. Boggess, A. Watkins, "Artificial immune system classification of multiple-class problems" , in: Proceedings of the Artificial Neural Networks in Engineering ANNIE 02, 2002, pp. 179–183.
9. Yang B, Wang L, Chen Z, Chen Y, Sun R (2010), "A novel classification method using the combination of FDPS and flexible neural tree", Neurocomputing 73:690–699
10. Shafigh P, Yazdi Hadi S, Sohrab E (2013), "Gravitation based classification", Inf Sci 220:319–330
11. A. L. Gamboa, M.G.M., J. M. Vargas, N. H. Gress, and R. E. Orozco, "Hybrid Fuzzy-SV Clustering for Heart Disease Identification", in Proceedings of CIMCA-IAWTIC'06.2006.
12. K. Polat and S. Gunes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Dig.Signal Process.*, vol. 17, no. 4, pp. 702–710, Jul. 2007.
13. https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data
14. Z. Pawlak, Rough sets, International Journal of Parallel Programming 11 (5) (1982) 341–356.
15. H.L. Chen, B. Yang, J. Liu, D.Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis", Expert Systems with Applications 38(2011) 9014–9022.
16. Z. Pawlak, Why rough sets. In Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, vol, 2, 8–11September 1996, New Orleans, LA, USA, 738–743.
17. N. Rami, N. Khushaba, A. Al-Ani, A. Al-Jumaily, Feature subset selection using differential evolution and a statistical repair mechanism, in: Expert Systems with Applications, Elsevier, 2011, pp. 11515–11526.
18. Javad S.S, Mohammad H.Z, Kourosh M, Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA), *Computer Methods and Programs in Biomedicine*.8, 2012, pp.570-579.
19. Tan K.C, Teoh E.J, Yua Q, Goh, K.C, A hybrid evolutionary algorithm for attribute selection in data mining, Expert Systems with Applications, Elsevier, 2009.
20. A.F. Atiya, A. Al-Ani, A penalized likelihood based pattern classification algorithm, *Pattern Recogn, 42*, 2009, pp.2684–2694.