Rough Sets-Based Rules Generation Approach: A Hepatitis C Virus Data Sets

Ahmed Zaki[1,⋆], Mostafa A. Salama[2,*], Hesham Hefny[1], and Aboul Ella Hassanien[3,*]

[1] Department of Computer Sciences and Information, ISSR, Cairo University, Egypt
[2] Department of Computer Science, British University in Egypt, Cairo, Egypt
[3] Faculty of Computers and Information, Cairo University, Cairo, Egypt
http://www.egyptscience.net

Abstract. The risk of hepatitis-C virus is considered as a challenge in the field of medicine. Applying feature reduction technique and generating rules based on the selected features were considered as an important step in data mining. It is needed by medical experts to analyze the generated rules to find out if these rules are important in real life cases. This paper presents an application of a rough set analysis to discover the dependency between the attributes, and to generate a set of reducts consisting of a minimal number of attributes. The experimental results obtained, show that the overall accuracy offered by the rough sets is high.

Keywords: rough sets, rough reduct, rule generation, HCV.

1 Introduction

Medical informatics or Bioinformatics deals with the resources, devices, and methods required optimizing the acquisition, storage, retrieval, and use of information in health and biomedicine. Medical informatics tools include not only computers but also clinical guidelines, formal medical terminologies, and information and communication systems. It is applied to the areas of nursing, clinical care, dentistry, pharmacy, public health and (bio) medical research. As more data is gathered, medical doctors cannot use the tools for their own data analysis individually. User-centered universal tools should be applied for medical researchers to analyze their own data. An example of these researchers is the study in [1] has been started to discover the differences in the temporal patterns of hepatitis B (HBV) and C (HCV) which has not been clearly defined, and more importantly, to examine whether the methods applied can work well and be applied to other fields. Hepatitis is swelling and inflammation of the liver. It is not a condition, but is often used to refer to a viral infection of the liver. Hepatitis can be caused by: Immune cells in the body attacking the liver and causing autoimmune hepatitis Infections from viruses (such as hepatitis A, B, or C), bacteria, or parasites Liver damage from alcohol, poisonous mushrooms,

⋆ Scientific Research Group in Egypt (SRGE).

A. Ell Hassanien et al. (Eds.): AMLTA 2012, CCIS 322, pp. 52–59, 2012.

or other poisons, medications, such as an overdose of acetaminophen, which can be deadly. Liver disease can also be caused by inherited disorders such as cystic fibrosis or hemochromatosis, a condition that involves having too much iron in your body (the excess iron deposits in the liver). The World Health organization (WHO) estimates that 170 million people, i.e. 3% of the world's population, are currently infected with the hepatitis-C virus (HCV) [2]. In majority of such cases, Symptoms are not appeared in the infected people for many years thus leaving them totally unaware of their condition. Liver damage is not caused by the virus itself but by the immune reaction of the body to the attack. This damage can be extremely serious, therefore resulting in liver failure and death of the patient. The indications of this type of hepatitis are normally less critical than hepatitis B. Hepatitis C spreads through contaminated blood or blood products, Sexual contact, contaminated intravenous needles. With some cases of Hepatitis C, no approach of transmission can be recognized. The current treatment for HCV, according to the United Kingdom's clinical guidelines, is with a combination therapy of two drugs: Interferon-alpha and Ribavirin [3]. A chief factor in prescribing combination therapy is that both drugs generate side effects in most inhabitants. The cost of combination therapy is between Ł3000 and Ł12,000 per patient per year. It is a general thinking that treating patients from pricey drugs with potentially severe side effects may be unsuitable unless there is a clear evident that patient has been infected from the virus. A liver biopsy is presently the only technique available to assess HCV activity. The biopsy involves removing a small core of tissue, which is approximately 15 mm in length and 2-3 mm in Diameter. This core is then goes on in paraffin wax, cut into pieces along its length and stained. At this level, a trained histopathology's will investigate the Samples under a light microscope and use his/her practice, Combined with a comprehensive definition, to evaluate the level of damage. The damage can usually be classified into two types and it is general to assign a numerical score relative to the level of damage for each type. One of the most widely used scoring methods is the Ishak system [4], which can be summarized as: Inflammation: assigned a necroinflammatory2 (activity) score from 0 to 18.

The rest of this paper is organized as follows: Section 2 gives an overview of the rough set theory and its techniques. Section 3 describe the proposed rule generation approach. Section 4 describes the experimental results and analysis, while the conclusion is presented in Section 5.

2 Rough Sets: A Brief Overview

Due to space limitations we provide only a brief explanation of the basic framework of rough set theory, along with some of the key definitions. A more comprehensive review can be found in sources such as [11,12,13].

Rough sets theory provides a novel approach to knowledge description and to approximation of sets. Rough theory was introduced by Pawlak during the early eighties [11] and is based on an approximation space-based approach to classifying sets of objects. In rough sets theory, feature values of sample objects

are collected in what are known as information tables. Rows of a such a table correspond to objects and columns correspond to object features.

Let $\mathcal{O}, \mathcal{F}$ denote a set of sample objects and a set of functions representing object features, respectively. Assume that $B \subseteq \mathcal{F}, x \in \mathcal{O}$. Further, let $x_{\sim_B}$ denote

$$x_{/\sim_B} = \{y \in \mathcal{O} \mid \ \forall \phi \in B, \phi(x) = \phi(y)\},$$

i.e., $x_{/\sim_B}$ (description of x matches the description of y). Rough sets theory defines three regions based on the equivalent classes induced by the feature values: lower approximation $\underline{B}X$, upper approximation $\overline{B}X$ and boundary $BND_B(X)$. A lower approximation of a set X contains all equivalence classes $x_{/\sim_B}$ that are proper subsets of X, and upper approximation $\overline{B}X$ contains all equivalence classes $x_{/\sim_B}$ that have objects in common with X, while the boundary $BND_B(X)$ is the set $\overline{B}X \setminus \underline{B}X$, *i.e.*, the set of all objects in $\overline{B}X$ that are not contained in $\underline{B}X$. Any set X with a non-empty boundary is *roughly* known relative, *i.e.*, X is an example of a rough set.

The indiscernibility relation $\sim_B$ (also written as Ind_B) is a mainstay of rough set theory. Informally, $\sim_B$ is a set of all classes of objects that have matching descriptions. Based on the selection of B (*i.e.*, set of functions representing object features), $\sim_B$ is an equivalence relation that partitions a set of objects $\mathcal{O}$ into classes (also called elementary sets [11]). The set of all classes in a partition is denoted by $\mathcal{O}_{/\sim_B}$ (also by $\mathcal{O}/Ind_B$). The set $\mathcal{O}/Ind_B$ is called the quotient set. Affinities between objects of interest in the set $X \subseteq \mathcal{O}$ and classes in a partition can be discovered by identifying those classes that have objects in common with X. Approximation of the set X begins by determining which elementary sets $x_{/\sim_B} \in \mathcal{O}_{/\sim_B}$ are subsets of X.

3 Rule Generation Approach on Hepatitis C Virus Data Sets

With increasing sizes of the amount of data stored in medical databases, efficient and effective techniques for medical data mining are highly sought after. Applications of Rough sets [5, 6, 7]in this domain include inducing propositional rules from databases using Rough sets prior to using these rules in an expert system. Tsumoto [8] presented a knowledge discovery system based on rough sets and feature-oriented generalization and its application to medicine. Diagnostic rules and information on features are extracted from clinical databases on diseases of congenital anomaly. Experimental Results showed that the proposed method extracts expert knowledge correctly and also discovers that symptoms observed play important roles in differential diagnosis. Hassanien el al. [9] presented a rough set approach to feature reduction and generation of classification rules from a set of medical datasets. They introduced a rough set reduction technique to find all redacts of the data that contain the minimal subset of features associated with a class label for classification. To evaluate the validity of the rules based on the approximation quality of the features, a statistical test

to evaluate the significance of the rules was introduced. Rough sets rule-based classifier performed with a significantly better level of accuracy than the other classifiers. Therefore, the use of reducts concept in combination with rule-based classification offers an improved solution for data set recognition. The medical diagnosis process can be interpreted as a decision-making process, during which the physician induces the diagnosis of a new and unknown case from an available set of clinical data and from clinical experience. This process can be computerized in order to present medical diagnostic procedures in a rational, objective, Accurate and fast way. In fact, during the last two or three decades, diagnostic decision support systems have become a well-established component of medical technology. Podraza et. al [10] presented an idea of complex data analysis and decision support System for medical staff based on rough set theory. The main aim of their system is to provide an easy to use, commonly available tool for efficiently diagnosing Diseases, suggesting possible further treatment and deriving unknown dependencies between different data coming from various patients' examinations. A blueprint of a possible architecture of such a system is presented including some example algorithms and suggested solutions, which may be applied during implementation. The unique feature of the system relies on removing some data through rough set decisions to enhance the quality of the generated rules. Usually such data is discarded, because it does not contribute to the knowledge acquisition task or even hinder it. In their approach, improper data (excluded from the data used for drawing Conclusions) is carefully taken into considerations. This methodology can be very important in medical applications as a case not fitting to the general classification cannot be neglected, but should be examined with special care.

One way to construct a simple model computed from data, easier to understand and with more predictive power, is to create a set of minimal number of rules. Some condition values may be unnecessary in a decision rule produced directly from the database. Such values can then be eliminated to create a more comprehensible minimal rule preserving essential information. The presented rough set approach for rule generation contains two phases (a) Discretization and (b) rule Generation. More details within the following subsections.

Discretization Based on RSBR. A real world data set, like medical data sets, contains mixed types of data including continuous and discrete valued data sets. The discretization process divides the attribute's value into intervals [11]. The discretization based on RS and Boolean Reasoning (RSBR) shows the best results in the case of heart disease data set. In the discretization of a decision table $S = (U, A \bigcap \{d\})$, where U is a non-empty finite set of objects and A is a non-empty finite set of attributes. And $V_a = [x_a, x_a)$ is an interval of real values x_a, w_a in attribute a. The required is to a partition P_a of V_a for any $a \in A$. Any partition of V_a is defined by a sequence of the so-called cuts $x_1 < x_2 < .. < x_k$ from V_a. The main steps of the RSBR discretization algorithm are provided in algorithm 1.

Algorithm 1. RSBR discretization algorithm

Input: Information system table (S) with real valued attributes A_{ij} and n is the number of intervals for each attribute.
Output: Information table (ST) with discretized real valued attribute.

1: **for** $A_{ij} \in S$ **do**
2: Define a set of boolean variables as follows:

$$B = \{\sum_{i=1}^{n} C_{ai}, \sum_{i=1}^{n} C_{bi} \sum_{i=1}^{n} C_{ci}, ..., \sum_{i=1}^{n} C_{ni}\} \tag{1}$$

3: **end for**
Where $\sum_{i=1}^{n} C_{ai}$ correspond to a set of intervals defined on the variables of attributes a
4: Create a new information table S_{new} by using the set of intervals C_{ai}
5: Find the minimal subset of C_{ai} that discerns all the objects in the decision class D using the following formula:

$$\Upsilon^{u} = \wedge\{\Phi(i,j) : d(x_i \neq d(x_j)\} \tag{2}$$

Where $\Phi(i,j)$ is the number of minimal cuts that must be used to discern two different instances x_i and x_j in the information table.

3.1 Rule Generating and Analysis Phase

Unseen instances are considered in the discovery process, and the uncertainty of a rule, including its ability to predict possible instances, can be explicitly represented in the strength of the rule [11]. The quality of rules is related to the corresponding reduct(s). We are especially interested in generating rules which cover largest parts of the universe U. Covering U with more general rules implies smaller size of a rule set. The main steps of the rule generation algorithm are provided in algorithm 2.

4 Experimental Works and Discussions

The data available at UCI machine [13] learning data repository Contains 19 fields with one output field. The output shows whether patients with hepatitis are alive or dead. The intention of the dataset is to forecast the presence or absence of hepatitis virus. Given the results of various medical tests carried out on a patient. The Hepatitis dataset contains 155 samples belonging to two different target classes. There are 19 features, 13 binary and 6 features with 6-8 discrete values. Out of total 155 cases, the class variable contains 32 cases that died due to hepatitis. In this section, an dataset table has been chosen and hybridization scheme that combines the advantages of PCA,and rough sets

Algorithm 2. Rule generation and classification

Input: reduct sets $R_{final} = \{r_1 \cup r_2 \cup .... \cup r_n\}$
Output: Set of rules

for each reduct r **do**
  for each correspondence object x **do**
    Contract the decision rule $(c_1 = v_1 \wedge c_2 = v_2 \wedge .... \wedge c_n = v_n) \longrightarrow d = u$
    Scan the reduct r over an object x
    Construct $(c_i, 1 \leq i \leq n)$
    for every $c \in C$ **do**
      Assign the value v to the correspondence attribute a
    end for
    Construct a decision attribute d
    Assign the value u to the correspondence decision attribute d
  end for
end for

in conjunction with statistical feature extraction techniques, have been applied to see their ability and accuracy to detect and classify the dataset into two outcomes: die or live.

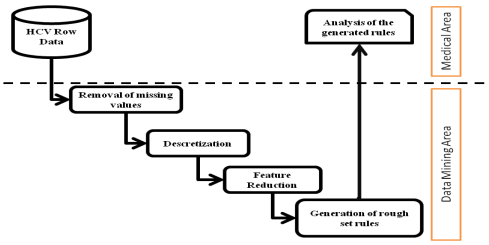

Fig. 1. Data mining services to the medical area

The architecture of the proposed rough sets approache is illustrated in figure 1. It is comprised of four fundamental building phases: In the first phase of the investigation, a preprocessing algorithm based on Normalizing data processing is presented. It is adopted to improve the quality of the data and to make the feature reducts phase more reliable. The set of features relevant to region of interest is extracted, and represented in a database as vector values. The third phase is rough set analysis. It is done by computing the minimal number of necessary attributes, together with their significance, and generating the sets of rules. Finally, a rough is designed to discriminate different regions of interest in order to separate them into die and live cases. These four phases are described in detail in the following section along with the steps involved and the characteristics

Table 1. The architecture of the proposed rough sets approache

Matches	Conditions	Class
86	(B=(-Inf,1.65)), (AP=(-Inf,131.5)) (P=(50.5,Inf))	2
77	(SP=(1.45,Inf)), (AP =(-Inf,131.5)) (P=(50.5,Inf))	2
52	(S=(58.5,Inf)), (P=(50.5,Inf))	2
46	(SP=(1.45,Inf)), (B=(-Inf,1.65)) (AP=(-Inf,131.5)), (S=(-Inf,58.5))	2
33	(AGE=(29.0,37.5)), (P=(50.5,Inf))	2
25	(AGE=(47.5,Inf)), (B=(-Inf,1.65)) (P=(50.5,Inf))	2
25	(AGE=(-Inf,29.0))	2
23	(AGE=(37.5,47.5)), (SP=(1.45,Inf)) (B =(-Inf,1.65)), (AP =(-Inf,131.5))	2
21	(AGE=(47.5,Inf)), (B=(-Inf,1.65)) (S=(-Inf,58.5))	2
19	(AGE=(47.5,Inf)), (SP=(1.45,Inf)) (P=(50.5,Inf))	2

feature for each phase. The features used are {Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices, Bilirubin, Alk Phosphate, Sgot, Albumin, Protime, Histology}. The generated set of reducts are: {Age [A], Spiders [SP], Bilirubin [B], "Alk Phosphate" [AP], Sgot [S], Protime [P]}. The extracted rules from this data set, to successfully classify the input data in 100% accuracy of classification are 51 rules. Table (1) shows only the first 10 rules of the highest matches.

5 Conclusions and Future Works

In the steps of classification, discretization is one of the important steps in medical analysis. It partition the attribute value according to the classification problem. These cuts itself represents an importance in the medical field. It shows at which cut in the values the medical experts shows start to worry about the patient and what is the range of values that is considered as an alarm range. According to results after applying the reducts from the rough set theory and determining the rules, the classification accuracy resulted were 100. This implies that the generated rules are sufficient to classify the patient as If these rules are analyzed by experts in the field of medicine. It provides that the patients is in danger and the percentage of death is high. The future work here is to show this percentage in an accurate methodology.

References

1. Booth, J., OGrady, J., Neuberger, J.: Clinical guidelines on the management of hepatitis C, vol. 1, pp. 11–21 (2001)
2. Kedziora, P., Figlerowicz, M., Formanowicz, P., Alejska, M., Jackowiak, P., Malinowska, N., Fratczak, A., Blazewicz, J., Figlerowicz, M.: Comp utational Methods in Diagnostics of Chronic Hepatitis C. Bulletin of the Polish Academy of Sciences, Technical Sciences 53(3), 273–281 (2005)
3. Hodgson, S., Harrison, R.F., Cross, S.S.: An automated Pattern recognition system for the quantification of Inflammatory cells in hepatitis-C-infected liver biopsies. Image and Vision Computing 24, 1025–1038 (2006)
4. Ishak, K., Baptista, A., Histological, L.B., et al.: Histological grading and staging of chronic hepatitis. Journal of Hepatology 22, 696–699 (1995)
5. Pawlak, Z.: Rough Set. International Journal of Computer and Information Sciences 11, 341–356 (1982)
6. Wojcik, Z.: Rough approximation of shapes in pattern recognition. Computer Vision, Graphics, and Image Processing 40, 228–249 (1987)
7. Pal, S.K., Pal, B.U., Mitra, P.: Granular computing, rough entropy and object extraction. Pattern Recognition Letters 26(16), 2509–2517 (2005)
8. Tsumoto, S.: Automated Extraction of Medical Expert System Rules from Clinical Databases on Rough Set Theory. Journal of Information Sciences 112, 67–84 (1998)
9. Hassanien, A.E., Abraham, A., Peters, J.F., Schaefer, G., Henry, C.: Rough sets and near sets in medical imaging: A review. IEEE Trans. Info. Tech. in Biomedicine (2009), doi:10.1109/TITB.2009.2017017
10. Podraza, R., Dominik, A., Walkiewicz, M.: Decision Support System for Medical Applications. In: Proceedings of the IASTED International Conference on Applied Simulations and Modeling, Marbella, Spain, pp. 329–333. ACTA Press, Anaheim (2003)
11. Chao, S., Li, Y.: Multivariate interdependent discretization for continuous attribute. In: Proceeding of the 3rd International Conference on Information Technology and Applications, vol. 1, pp. 167–172 (2005)
12. Hameed, A.-Q., Ella, H.A., Ajith, A.: A Generic Scheme for Generating Prediction Rules Using Rough Sets Rough Set Theory: A True Landmark in Data Analysis, January 01, vol. 174, pp. 163–186 (2009)
13. UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml/datasets.html`