World Applied Sciences Journal 29 (Data Mining and Soft Computing Techniques): 70-74, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.29.dmsct.13

# Hepatitis Investigation Using Rough Set Discritized Rule Based Soft Computing Technique

*S. Srimathi, B. Shalini and B. Santhi*

School of Computing,
Sastra University, Thanjavur - 613402, Tamil Nadu, India

**Abstract:** Hepatitis causes severe infection to the liver and prevents liver from its proper functioning. Hence prior judgment of hepatitis is much essential. Soft Computing Techniques provide excellent methodologies to process the medical data and help medical experts in finding out the nature of illness. True data set collection, feature squeezing and classification are the basic steps followed in designing an expert system. The designed expert system acts with intelligence, prevents erroneous decisions and produces sharp results in time. This paper discusses on hepatitis investigation using rough set discritized rule based soft computing technique. The extracted rough set discritized table (RS-DT) based on cuts is used in both training and testing. Finally the rough set discritized rule set (RS-DRS) formed from the discritized table is used in classification. Thus hepatitis investigation using rough set discritized rule based soft computing technique promotes the performance.

**Key words:** Hepatitis · Discritized table · Rule set · Accuracy · Coverage

## INTRODUCTION

Liver performs major functions which are essential for healthy living. It aids in digestion, nutrient extraction, toxin excretion and balancing glucose content. Liver infection existing for less than 6 months is acute hepatitis and infection exiting more than 6 months is chronic hepatitis. Medical experts have divided hepatitis into viral or infectious hepatitis and non-infectious hepatitis.

Viral hepatitis can be spread easily. Liver infection caused due to drug consumption, immune deficiency, genetic disorder, metabolic disorder and obesity leads to non-infectious hepatitis because they cannot be spread. Medical experts take blood test to confirm hepatitis. In medical pronouncement various soft computing techniques are used to design expert systems. Such expert systems assist medical experts in investigation of hepatitis. The first step involves true data set collection in which reports of hepatitis and non-hepatitis cases are collected from medical centers. The second step involves the mechanism of choosing a subgroup of significant attributes and the mechanism is called as feature selection, attribute selection, variable subset selection or dimensionality reduction. The irrelevant attributes add noise to the system which reduces the quality and accuracy of the system and maximizes the size, time complexity and computational resources needed for the system. Hence the mechanism of attribute selection minimizes the negative factors and increases the system performance.

The third step involves classification in which members are categorized into their respective decision classes. The proposed system involves rough set discritized table (RS-DT) and rough set discritized rule set (RS-DRS). The main aim of this paper is to process the data set even in the presence of absent values, to increase accuracy and coverage, to use the cut set, discritized table and rule set mechanisms in hepatitis investigation.

**Literature Survey:** Using Linear Discriminant Analysis and Fisher Discriminant Analysis obtained accuracy were 86.4%, 85.3% and 83.2% [1]. 94.14% of accuracy was obtained using Artificial Immune System and Principal Component Analysis [2]. Using discriminant study on linear systems and Adaptive Neuro Deduction System based on fuzzy, the accuracy rate achieved was 94.16% [3]. Further, 95.0% was achieved with Principal Component Analysis and Least Square Support Vector Machine [4]. Simulated Annealing and Support Vector Machine were used to achieve 96.25% [5]. Using Local

**Corresponding Author:** S. Srimathi, School of Computing, Sastra University, Thanjavur – 613402, Tamil Nadu, India. Tel: +9843443344.

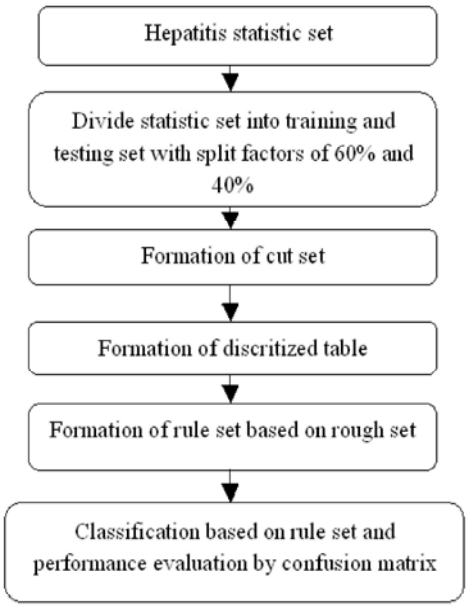


Fig. 1: Work Flow

Fisher Discriminant Study and Support Vector Machine the accuracy obtained was 96.77% [6]. On an average accuracy of 96.49% was obtained using Rough Set and Extreme Learning Machine with data split percentage of 50-50, 70-30 and 80-20 [7]. The limitations are identified as follows, the accuracy level, the methods used in training and testing, the number of reduced sets formed in feature extraction, the working time on each reduced set to find the best reduct.

**Proposed System:** The proposed system for hepatitis investigation uses rough set discritized table (RS-DT) and rough set discritized rule set (RS-DRS) through methods like cut set, discritized table and rule set. The steps included in the new system are given below and Figure 1: Represents the steps followed

**Step1:** Hepatitis statistic set

- Collect the hepatitis statistic set which includes relevant objects and attributes.

**Step2:** Divide data set

- Divide the hepatitis statistic set into training and testing sets using split factor of 60% and 40%.

**Step 3:** Cut set

- Form cut set for the training set which represents the non overlapping subsets thus divides the attributes into set of intervals. Cuts act as a boundary values representing those intervals.

Form discritized table for training based on cuts of hepatitis statistic set. The main theme of discritized table is feature reduction thus it has reduced the size of actual attribute table.

Discritized table for testing set is formed based on the cuts formed from training set.

**Step 6:** Rule set

- Rules are formed from the discritized table by combining the discritized values through logical AND. It also states the result class to which a particular rule points to.

**Step 7:** Classification

- Rule based classifier which categorizes the testing objects.

**Step 8:** Performance Evaluation

- Performance is evaluated through confusion matrix which shows performance in terms of accuracy and coverage.

$$Accuracy = \frac{Number\ of\ objects\ correcly\ classicfied \times 100}{Total\ Number\ of\ objects}$$

$$Coverage = \frac{Number\ of\ Objects\ \mathrm{cov}ered \times 100}{Total\ Number\ of\ Objects}$$

**Attributes Statistics:** The feature set is acquired from UCI machine learning medical database [8]. The feature set has 155 objects and 19 attributes along with 1 class attribute. 123 members belong to live class attribute and 32 members belong to die class attribute. Number of absent values is found as 167 (5.39%). To increase the availability of data the system retains the absent values. It can also be replaced with mean or standard deviation. Except the attributes stated below all others take inputs 1 and 2 for Yes and No respectively. Bilirubin values are 0.39, 0.80,

Table 1: Data Table for Hepatitis\

| 155/20 | Age | Sex | Steroid | Antivirals |
|---|---|---|---|---|
| O:1 | 44 | 1 | 1 | 2 |
| O:2 | 30 | 1 | 2 | 2 |
| O:3 | 38 | 1 | 1 | 2 |
| O:4 | 38 | 1 | 1 | 2 |
| O:5 | 50 | 2 | 1 | 2 |

Table 2: Attributes Statistics

| Attribute | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|
| Age | 41.2 | 12.566 | 7 | 78 |
| Sex | 1.103 | 0.305 | 1 | 2 |
| Steroid | 1.506 | 0.502 | 1 | 2 |
| Antivirals | 1.845 | 0.363 | 1 | 2 |
| Fatigue | 1.351 | 0.479 | 1 | 2 |
| Malaise | 1.604 | 0.491 | 1 | 2 |
| Anorexia | 1.792 | 0.407 | 1 | 2 |
| Liver big | 1.828 | 0.379 | 1 | 2 |
| Liver firm | 1.583 | 0.495 | 1 | 2 |
| Spleen palpable | 1.8 | 0.401 | 1 | 2 |
| Spiders | 1.66 | 0.475 | 1 | 2 |
| Ascites | 1.867 | 0.341 | 1 | 2 |
| Varices | 1.88 | 0.326 | 1 | 2 |
| Bilirubin | 1.427 | 1.211 | 0.3 | 2 |
| Alk phosphate | 105.325 | 51.508 | 26 | 295 |
| Sgot | 85.894 | 89.651 | 14 | 648 |
| Albumin | 3.817 | 0.651 | 21 | 6.4 |
| Protime | 61.852 | 22.875 | 0 | 100 |
| Histology | 1.452 | 0.499 | 1 | 2 |

Table 3: Logic Table

| Logics | Form |
|---|---|
| Associative | P.(Q.R) = (P.Q).R |
| | P+(Q+R) = (P+Q)+R |
| Absorption | P.(P+Q) = P |
| | P+P.Q = P |
| Distributive | P.(Q+R) = P.Q+P.R |
| | P+(Q.R) = (P+Q).(P+R) |
| Commutative | P.Q = Q.P |
| | P+Q = Q+P |
| Idempotent | P.P = P |
| | P+P = P |

1.20, 2.00, 3.00 and 4.00. Alk phosphate values are 33, 80, 120, 160, 200 and 250. Sgot values are 13, 100, 200, 300, 400 and 500. Albumin values are 2.1, 3.0, 3.8, 4.5, 5.0 and 6.0. Protime values are 10, 20, 30, 40, 50, 60, 70, 80 and 90. Table 1 represents the data table for hepatitis.

The statistics for the hepatitis data table is formed with parameters like attribute, data type, status, precision, mean, standard deviation, minimum and maximum values. For class attribute the data type is symbolic and for other attributes the data type is numeric. The status for class attribute is decision and for other attributes it is condition and the precision is set as 2 for all attributes. Table 2 depicts the statistics of attributes.

**Rough Set:** Rough set states a general system represented as S(O, T) consists of set of objects O and set of attributes T. Let V be the value set. Every $v_i \in V$ is assigned to every attribute $t \in T$ mapped with the objects O. Rough set defines an indiscernibility equivalence relation IND(Q) with respect to $Q \subseteq T$ and Q be the subset of attribute set T. The objects in the equivalence relation are indistinguishable from each other with respect to Q. The objects b and a are indistinguishable if (b, a) $\in$ IND(Q).

The lower approximation defines the objects which are surely classified as positive. The upper approximation defines the objects which maybe a member of objective set. The difference between two approximations is called the boundary area. Thus rough set stands as a mix of lower and upper approximations.

**Cut Set:** Cut set defines the mechanism of decomposing attribute value set. The numerical attributes are discritized to generate a set of intervals. These are edge points to highlight the intervals. In case of representative attributes cuts stand for the non overlapping subsets of actual values. Table 4 represents cut set consists of attributes for which the cuts are created. Size represents the quantity of cuts created. In the report decimal represents cuts and * value represents absence of cuts for attribute. Cut set states the result attribute in terms of other attributes. It helps in feature reduction based on Boolean logics of algebra.

**Step1:** For each and every attribute through Boolean squeezing methods the values are continuously combined and reduced to form a minimal edge value.

**Step2:** The edge value is checked to ensure that it can represent cut value with which the values for a single attribute can be easily represented. Table 3 explains the logics used. Here P, Q and R are the attribute values that are combined through Boolean logics to form a cut value for a particular attribute.

**Discritized Table:** The process involves combination of attributes based on cut values. It divides the range of continuous attributes into intervals thus reduces the size of attribute table. The reduced discritized table replaces the actual table values. It prepares the data for further analysis called classification. Supervised, unsupervised, top-down, bottom-up and recursive operation on attributes are the methods used to perform discretization. Table 5 represents discritized table created from cut set.

Table 4: Cut Set

| (1-19) | Attribute | Size | Description |
|---|---|---|---|
| 1 | Age | 1 | 45.5 |
| 2 | Sex | 0 | * |
| 3 | Steroid | 0 | * |
| 4 | Antivirals | 0 | * |
| 5 | Fatigue | 0 | * |
| 6 | Malaise | 0 | * |
| 7 | Anorexia | 1 | 1.5 |
| 8 | Liver big | 1 | 1.5 |
| 9 | Liver firm | 0 | * |
| 10 | Spleen palpable | 0 | * |
| 11 | Spiders | 1 | 1.5 |
| 12 | Ascites | 0 | * |
| 13 | Varices | 0 | * |
| 14 | Bilirubin | 1 | 1.35 |
| 15 | Alk phosphate | 1 | 159.0 |
| 16 | Sgot | 0 | * |
| 17 | Albumin | 0 | * |
| 18 | Protime | 1 | 51.0 |
| 19 | Histology | 0 | * |

Table 5: Discritized Table

| 93/20 | Age | Sex | Steroid |
|---|---|---|---|
| O:1 | "(-Inf,45.5)" | * | * |
| O:2 | "(-Inf,45.5)" | * | * |
| O:3 | "(-Inf,45.5)" | * | * |
| O:4 | "(45.5,Inf)" | * | * |
| O:5 | "(-Inf,45.5)" | * | * |
| O:6 | "(45.5,Inf)" | * | * |

**Step 1:** with values from cut table discritized table produces minimum and maximum interval values for attributes.

**Step 2:** Include the attribute intervals as stated below

- If (attr_value <= cut_value)
- attr_value = -infinity to cut_value
- Else
- attr_value = cut_value to infinity

Thus discritized table reduced the number of objects from 155 to 93. Also instead of checking individual values for an attribute an interval is formed for every attribute with which the values can be easily represented. This method thus helps in rule formation without confusion and promotes feature squeezing with common representation.

**Rule Set Formation:** A set of rules are used to decide the possibility of attribute in classification. The rule is extracted based on decision matrix. Let the condition and decision attributes be C = $\{C_1, C_,...,C\}_n$ and D respectively where D $\in$ C and the rule is stated as $C_i^a C_i^b,...,C_k^c \to D^d$. It can also be stated as follows.

Table 6: Rule Set

| Match | Rules |
|---|---|
| 27 | Alk phosphate=(Inf,159.0)&Anorexia=(1.5,Inf) |
| 4 | Age=(-Inf,45.5)&Alk phosphate=(Inf,159.0) |
| 3 | Liver big=(1.5,Inf)&Age=(-Inf,45.5) |
| 3 | Anorexia=(1.5,Inf)&Liver big=(1.5,Inf) |

$$(C_i = \alpha) \wedge ... \wedge (C_k = c) \to (D = d)$$

Here a, b, c represents the values in the attribute set. The result matrix for d belongs to result attribute D provides all the D - d pairs that vary within objects that take values may or may not be equal to d. Table 6 represents rule set used for testing. The rule set has 3 columns first is rule count, second is match which represents the number of objects matching the conditional rule and third is decision rules which are framed with logical formula.

## RESULTS AND DISCUSSION

The investigation is reported by generating a result table shown by Table 7. The middle component of the result represents the matrix. Rows represent real result classes presented in the problem. Columns represent result values produced by the classifier. The crossway of matrix represents the number of objects perfectly classified. With split factor of 60% and 40% total number of trained and tested objects are 93 and 62 respectively. Out of 62 objects 61 objects are classified perfectly. In result class die out of 12 objects all 12 are perfectly classified and there is no misclassification. In result class live out of 50 objects 49 objects are perfectly classified and 1 object is misclassified. Thus the true positive result for classes dies and live are 0.92 and 1. The obtained accuracy for result classes die is 1 and for live it is 0.98. The obtained coverage for result class die is 1 and for live is 1. The system performs well with total accuracy of 0.984 and with coverage of 1. Thus system working is improved in terms of accuracy and coverage.

Table 7: Result

| | Predicted | | | | |
|---|---|---|---|---|---|
| | Die | Live | Objectcount | Accuracy | Coverage |
| Die | 12 | 0 | 12 | 1 | 1 |
| Live | 1 | 49 | 50 | 0.98 | 1 |
| TPR | 0.92 | 1 | | | |

- The total number of objects tested: 62
- Total accuracy: 0.984
- Total coverage: 1
- TPR - True Positive Rate

## CONCLUSION

Hepatitis investigation using rough set discritized rule based soft computing technique shows good accuracy and coverage. It has handled 155 objects and 20 attributes. The cut set mechanism forms edge points to form range of intervals through discritized table. Thus discritized table promotes feature squeezing in which it has reduced 155 objects to 93 objects and instead of checking individual values for the attributes range of intervals are formed for attributes. Logically, rule set combines the interval values of different attributes from discritized table. Hence size reduction through discritized table aids in easy classification through rule formation. The classification accuracy and coverage obtained by the above method is 0.984 and 1 respectively. The above method handles attributes even in presence of absent values. In future, the same method can be used to investigate other diseases. Replacing the absent values with mean or standard deviation may be used. The rough set rule based approach can be combined with other classifiers to enhance further. Thus hepatitis investigation using rough set discritized rule based soft computing provides good accuracy and coverage.

## REFERENCES


1. Ster, B. and A. Dobnikar, 1996. Neural networks in medical diagnosis: comparison with Other methods, in: Proceedings of the International Conference on Engineering Applications of Neural Networks, 1(1): 427-430.
2. Polat, K. and S. Gunes, 2007. Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system, Applied Mathematics and Computation, 189(2): 1282-1291.
3. Dogantekin, E., A. Dogantekin and D. Avci, 2009. Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system, Expert Systems with Applications, 36(8): 11282-11286.
4. Calisir, D. and E. Dogantekin, 2011. A new intelligent hepatitis diagnosis system: PCA-LSSVM, Expert Systems with Applications, 38(8): 10705-10708.
5. Javad, S.S., H.Z. Mohammad and M. Kourosh, 2012. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA), Computer Methods and Programs in Biomedicine, 108(2): 570-579.
6. Chen, H., D. Liu, B. Yang, J. Liu and G. Wang, 2011. A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis, Expert Systems with Applications, 38(9): 11796-11803.
7. Yılmaz Kaya and Murat Uyar, 2013. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease, Applied Soft Computing, 13: 3429-3438.
8. Blake, C.L. and C.J. Merz, 1996. UCI repository of machine learning databases, 1996, given in http://www.ics.uci.edu./~mlearn/MLReporsitory.html