Abstract

Clinical diagnosis is an unavoidable part of the treatment process for a healthy society. Computer Diagnosis has brought the scenario to great extents from where the whole clinical practice used to be. Now for it to evolve further we need advancements at the very premium standards like the improvements in accuracy and precision of a diagnosis results. Inflow treatment procedures require computationally sound results in the best possible time complexity.

In this paper, we propose a rough set based network method that can help a neural network to reduce the decision computational time, as the rough set attribute reductions will help the network to gain better efficiency by removing the unwanted attribute relations from the dataset. To demonstrate this method we use a modified version of the Wisconsin Breast Cancer Dataset obtained from the UCI Repositories.

Keywords: Rough Set, Breast Cancer, Diagnosis, Relative Reduction, Neural Networks.

1. Introduction

The continuous evolution of medical diagnostic systems have helped us in treatment of many deadly diseases present today. This has been possible due to grow of technology in image processing mechanisms which can collect large amounts of data from a diagnostic image set. An Image can contain millions of minute data, which can have the detailed description of different group of image clusters. Prime challenges faced in this field are determination of the most useful features that we extract from an image and then their classification based on certain criteria. The accuracy of system can be increased by maximising the specificity of the classification. We can overcome this by reducing the unwanted vagueness in image property relations [1]. As for the same reason, defined relations must be exact, or the precise reasoning will be impossible, thus we propose a system based on a recently developed concept of rough sets attribute relations [9].
To achieve the best computation time for dataset analysis one can use the proposed method, which finds the most relevant properties associated with a particular disease and then helps in classification of the final problem using the mainframe classifications techniques used in data classification [10]. An UCI Dataset has been used which contains the diagnostic datasets formed from real medical image diagnosis. This dataset is basically a result of applying knowledge mining techniques on various real world medical diagnosed images. Rough set theory helps us in finding the most relevant relation attributes amongst many attribute sets, thus ultimately reducing the efforts and computation times of running a particular classification technique.

After using the relative reduce algorithm, the neural network classification concept can be applied to get the required diagnosis about the disease treatment. The Dataset is purposefully divided into 2 different sets namely - training and testing sets, so that the Neural Network or the Artificial neural network can be trained using the training set before it gets tested using the testing set. Improvised techniques will be implanted which will help in further diagnosis measures based on requirement schemas.

2. Rough Sets

i. Basic Idea behind this approach

Rough set theory forms the very basis of critical decision making in terms of mathematical approach [20]. Decision tree analysis and other classical approaches can be linked back to this concept. It is based on bound value tables. Lower bound values means that the object’s certainty of belonging to a particular target class is sure [19]. Upper bound values denote those points in the whole set which cannot be classified as belonging to the target set with a sure certainty [2]. Boundary- line elements decides whether an object can be classified as a part of the set or its complement or none. Basically thus we see that rough sets measures the uncertainty in relations not precisely by the membership values of properties but with the boundary value limits of a function [3]. If the boundary region is a null region then the set is said to be crisp set otherwise any not null values in boundary defines that the set is a rough set.

Formal definitions of approximations and the boundary region are as follows:

- **R-lower approximation** of $X$
  $$R_\ast(x) = \bigcup_{x \in U} \{ R(x) : R(x) \subseteq X \}$$

- **R-upper approximation** of $X$
  $$R^\ast(x) = \bigcup_{x \in U} \{ R(x) : R(x) \cap X \neq \emptyset \}$$
R-boundary region of $X$

$$RN^*_r(X) = R^*(X) - R_r(X)$$

Rough sets memberships functions can be defined as:

$$\mu^b_X(x) = \frac{|X \cap B(x)|}{|B(x)|}, \text{ where}$$

$$\mu^b_X(x) \in [0, 1].$$

The values related to the member function $\mu_X(x)$ can be interpreted as a conditional assumption with some certainty for which $x$ belongs to $X$

Fig. 1 – Rough set representation

ii. Applications of the Rough Set theory

Rough set have wide applications in data mining and knowledge discovery methods. Rule induction and feature selection problems have been mainly deal using the concepts of rough set attribute reductions (semantics-preserving reduction patterns). The kind of perfection in results and computation reduction provided by using the rough set approach is really feasible, as mining data from large repositories can be very much time and resource consuming. One of the most useful features of rough sets is the ability to reduce or remove the unwanted attributes [4]. So by using this method we can apply the probabilistic neural network to classify the given image segments or image feature datasets. This framework of task is efficient by computational measures and can be trusted for accuracy and reliability. Neural Networks can be trained and tested for the same purpose.

3. Neural Networks

i. Introduction to the network bounds

Neural networks are known for their ability to learn and generalise from example data even when the data is noisy and incomplete. This feature of neural network has provided a way for automated knowledge usage for improving the
system workability. The system working is done by acquiring the important relation attributes from a training set (derived from the original dataset) which indirectly helps the network to create a model and learn from the same in processing the whole dataset [5]. Knowledge of a neural network is created through loops of iterative functions such as if-then rules and other conditions specified separately. Because of the automated network formulation the whole system guarantees data perfection and consistency.

ii. **Rough Sets Neurons**

Neural network is basically based on the predictive modelling concept. The artificial neuron generation is evolved using a test case scenario or a test case model [18]. Thereafter the whole scenario learns using the training set given to the system. The integrated concept of rough neural network incorporates the upper and lower bound values. The resultant workaround processed using rough sets in neural networks are significantly better than the conventional neural network model. Upper and lower bound models have been in existence in many applicative ways particularly in fields like machine learning and artificial intelligence [14]. Rough set helps in pattern recognition between attribute values that define the classification of the final decision attribute in a given decision table. Neural Networks play a major role in this step, as it takes the sample of data which contains less attributes and forms a model pattern recognition algorithm. Hence heavy dataset can be handled and reduced to a computationally feasible dataset by using this system. The conventional neural network models is ought to be modified to accommodate rough patterns [15]. Each neurons contains a bound value. The decision classification will be done based on the defined attributes.

4. **Current work based on rough set classifications**

The Features present in an image can go over a million, so similar can happen with the computation time for the proper decision computation [16]. The time complexity increases exponentially. Even if we consider a small set of data the property selection can became really hectic. A remedy to this problem is to reduce the number of features in an informative datasets [11]. Several unsupervised and semi-/supervised, approaches have been proposed to address this issue. Association or classification rule formation is discussed in many papers and conferences. Most of these are formulated upon concepts of probable support and confidence [8]. But the major lacking if these work is the absence of differentiation of strong and weak relation properties. For a more consolidated effort, we need to find something that can mine out the strong relation properties from a given set of relational attributes [12]. However, the case of vagueness and uncertainty cannot be decided just by differentiating the strong and weakly related attributes. Many papers are based on fuzzy logic decision rule inductions, but they lack the ability to add new data in the model set at a
later stage [13]. The proposal for an incremental model may enhance such models. Although many incremental models have been proposed by various authors, they don’t support the inconsistent data sets. And if we are dealing with real world entities then there has to be some inconsistent data which needs to be accommodated while running the pre-processing step.

The currently proposed processes have not gone further into the field of classification of the results obtained or otherwise the computational time for processing the data has been must greater than that expected [17]. This however has restricted the development of this field in further expanding the work.

We will be dealing with such issues in the paper. Furthermore we will be processing the dataset obtained from the image for rough set redundant value reduction. This will further lead to classification of the data through neural rough set design process.

5. Ideology and Integration of the concepts

Clustered study of a bio functions can only be used for theoretical practice and thus remains less effective in normally functioning systems. But practically, doctors are more interested in systems that can help them with early diagnosis, so that they can proceed for the correct treatment strategies beforehand.

Thus in reference to the above context we will propose the required system for Breast Cancer Diagnosis.

We are using the dataset - Wisconsin Breast Cancer Diagnosis. To allow for a supervised regression model with no over-training, the dimensionality of the samples was reduced by using centralized values as input.

Dataset Description:

Wisconsin Breast Cancer Diagnosis Dataset taken from UCI machine learning repository is considered for this study. The original dataset contains 569 samples [23]. There is no missing value. For maintaining the accuracy standards we have customised the dataset as described below:

Attributes or properties found in the original dataset are:

1. Radius: obtained by averaging the length of radial line segments from the centroid to the individual snake points
2. Perimeter: Sum over the total distance of the snake points
3. Area: Number of pixels in the interior of the snake and adding 1/2 of the perimeter pixels
4. Compactness: perimeter2/area
5. Smoothness: Difference between length of a radial line and the mean length of the lines surrounding it
6. Concavity: Draw chords between non-adjacent snake points and measure distance to object boundary
7. Concave Points: Counts the number of contour concavities
8. Symmetry: Similar to relation between major and minor axis

9. Fractal Dimension

10. Texture Variance of the intensity levels in the interior of the snake.

We are using the dataset after reducing the sample size to 198[23].

Number of attributes: 34 (ID, outcome, 32 real-valued input features).

Missing values of lymph node status in 4 cases.

The Class distribution is formed as 151 non-recurrent, 47 recurrent.

Attribute information

1) ID order
2) Lymph node status
3) Radius mean, standard error, worst
4) Texture mean, standard error, worst
5) Perimeter mean, standard error, worst
6) Area mean, standard error, worst
7) Smoothness mean, standard error, worst
8) Compactness mean, standard error worst
9) Concavity mean, standard error, worst
10) Symmetry mean, standard error, worst
11) Fractal dimension mean, standard error, worst
12) Tumor size
13) Time

6. Implementations

i. Reduction

Fig. 2 – Graphical representation of the mean attribute values from the given set.
As we know, by default, every attribute is represented in real value measurement but the rough set theorem implementation needs a precise crisp values or discretized values. The Sample value is considered as 1 or 0 based on the mean of the values of attributes given.

Relative reduce feature is used. Using this rough set approach we can relatively reduce the values of the features that depend on the decision class attribute. Thus in large datasets this can result in conservation of immense computational power. Proper resource use is guaranteed after one set goes thought the relative reduct algorithm.

We define the dependency as:

$$K_r(D) = \frac{|U/\text{IND}(R)|}{|U/\text{IND}(R \cup D)|}$$

The relative reduct algorithm is defined as follows:

1. Set of all conditional Features
2. Set of all decision features
3. if \(K_{r-[a]}(D) = 1\)
4. \(R \leftarrow R \cup \{a\}\)
5. return \(R\)

Backward elimination process is used to remove the useless features from the total attribute set if the dependency of a particular attribute is equal to 1 i.e. meaning that the attribute is totally dependent on other attributes and thus won’t affect the decision class attribute even if is it removed from the whole set. The removal process is considered one at a time, starting from the first element [22].

For maintaining a better accuracy standard we consider first 5 attributes and check their dependencies based on relative reduct algorithm.

After taking the initial set as \{R,T,P,A,S\}, we start pruning by taking the attribute R. After using the formula for relative reduct, we get the dependency of R as 1. Thus it is removed. Similarly we prune for T, P, A and S attributes.

As we can observe due to the higher dependencies of 1 for R and A, both attributes get removed and at last the initial set is reduced to \{T,P,S\}. 

A. Manimaran* et al. International Journal Of Pharmacy & Technology

IJPT | July-2015 | Vol. 7 | Issue No.1 | 8049-8060
ii. Classification

For classification techniques we run the M5P Computational classification algorithm.

Firstly we divide the dataset into training and testing sets, so that the training set can help in creating the train model for artificial neural network and the test model can be used later to get better results [6].

```
Test mode: evaluate on training data

--- Classifier model (full training set) ---

M5 pruned model tree:
(using smoothed linear models)
LMI (116/59.412%)

LM num: 1
fractal_dimension_mean =
  0.0068 * radius_mean
  + 0.1569 * perimeter_mean
  + 2.081 * area_mean
  + 0.6309 * symmetry_mean
  - 0.1866

Number of Rules : 1
Time taken to build model: 0.55 seconds

--- Evaluation on training set ---

--- Summary ---

Correlation coefficient   0.804
Mean absolute error   0.0196
Root mean squared error 0.0245
Relative absolute error % 61.054
Root relative squared error % 59.412
Total Number of instances 116
```

Fig. 3 – Plot matrix representing the indiscriminate dependencies of the first 5 sample attributes.

Fig. 4: Training Data Classification.
As we can see the classification algorithm gives a relative absolute error of 61.0848 % in training model creation. The machine learning phase begins once we give the test dataset to the system. The Relative absolute error is reduced to 49.6592 %.

7. Results

Our proposed method is a result of the combination of rough sets approach for relative reduct functions, hybrid artificial networks pattern linking for the dataset and classification algorithms for decision selection [7]. We have mainly dealt with 5 attributes from the Wisconsin breast cancer dataset and have realised that the computation timing and resource usage can be exponentially reduced if one uses the relative reduce algorithm to remove the useless attributes from the dataset before processing it for further classifications.
We plan to expand our work and bring the reduction algorithm and neural networks to real time usage so that large amounts of data can be extracted for the very best related features and can be used according to the requirements of the system [21]. This approach will help in bringing the current medical diagnosis systems to a whole new level of more practical approach and treatment.

8. References


**Corresponding Author:**

A. Manimaran*,

**Email:** marans2011@gmail.com