Asian Journal of Information Technology 1 (1): CCCC, 2008
ISSN: 1682-3915
© Medwell Journals, 2008

Correlated Rough Set Based Classificatory Decomposition for Breast Cancer Diagnosis Using Fuzzy Art Neural Network

A. Punitha and T. Santhanam
P.G and Research Department of Computer Science, D.G. Vaishnav College,
Arumbakkam, Chennai, India

Abstract: Feature selection is an essential step in preprocessing and it refers to the process of selecting input variables that are most predictive for a given outcome. Reducing the input space is of major concern in areas like pattern recognition, signal processing, medical research and machine learning. Use of rough set theory for preprocessing of dataset has been very recent since other methods are inadequate at finding minimal reductions that too with uncertain data. The essence of this study is to introduce an innovative approach by fusing rough set theory with feature correlation for reducing the input space and then applying fuzzy ART neural network for breast cancer diagnosis. The intended approach has produced satisfactory results as opposed to the conventional methods.

Key words: Rough set theory, correlation, breast cancer diagnosis and fuzzy ART

INTRODUCTION

Breast cancer is the most common cancer in women worldwide. Trends and Statistics indicate that one out of nine women will develop breast cancer in their lifetime, and one out of 27 women die due to breast cancer (http://www.cancer.org/).To aid clinicians in the diagnosis of breast cancer, recent research has looked into the development of computer aided diagnostic tools.

Numerous samples and high dimensionality of the feature space are the major obstacles in processing large databases. One of the major challenges in breast cancer domain is the extraction of comprehensible knowledge from lab test results. It might be expected that the inclusion of an increasing number of features would further increase the likelihood of including enough information to distinguish between classes. Unfortunately, this is not true if the size of the training dataset is not directly proportional to the rapid increase of additional features. This is the so-called curse of dimensionality (Rich and Dayne, 1994).

Rough set theory provides a new mathematical tool to deal with uncertainty and vagueness of an information system in Data mining. The information system may contain a certain amount of redundancy that will not aid knowledge discovery and may in fact mislead the process. It is necessary to eliminate the redundant attributes by retaining the essential ones or to construct the core of the attribute set (Pawlak, 1991).

The study elaborates the experimental results using the well-known Wisconsin Breast Cancer Dataset (WBCD) with fuzzy ART neural network for classification and the final section deals with conclusion.

ROUGH SET THEORY

In many applications, data is automatically generated and therefore the number of objects to be mined can be large. The time needed to extract knowledge from such large data sets is an issue, as it may easily run in to days, weeks, and beyond. One way to reduce computational complexity of knowledge discovery with data mining algorithms and decision making based on the acquired knowledge is to reduce the volume of data to be processed at a time, which can be accomplished by decomposition (Zdzis, 1991).

Rough Set Theory (RST) (Komorowski *et al*., 1998) has indeed become a topic of great interest to researchers and plays a predominant role in many domains as reported in Chouchoulas and Shen (2001), Drwal (2000) and Ho *et al*. (2003) for classification, clustering (Jensen and Shen, 2001) systems monitoring (Sebban and Mock, 2002) and expert systems (Swiniarski, 1996). Given a dataset with discretized attribute values, it is possible to

Corresponding Author: T. Santhanam, P.G and Research Department of Computer Science, D.G. Vaishnav College, Arumbakkam, Chennai, India

find a subset (termed a reduct) of the original attributes using RST that are the most informative; all other attributes can be removed from the dataset with minimal information loss. From the dimensionality reduction perspective, informative features are those that are most predictive of the class attribute. This success is due in part to the following aspects of the theory (Pawlak, 1982).

- Only the facts hidden in data are analyzed.
- No additional information about the data is required such as thresholds or expert knowledge on a particular domain.
- It finds a minimal knowledge representation.

Rough sets remove superfluous information by examining attribute dependencies. It deals with inconsistencies, uncertainty and incompleteness by imposing an upper and a lower approximation to set membership. Rough sets estimates the relevance of an attribute by using attribute dependencies regarding a given decision class (Alexios, 2001) It achieves attribute set covering by imposing discernibility relation. Same as in fuzzy logic, in rough sets every object of interest is associated with a piece of knowledge indicating relative membership. This knowledge is used to drive data classification and is the key issue of any reasoning, learning, and decision making (Angela and Juan, 2006).

The Rough Set based Attribute Reduction (RSAR) technique is herein explained in terms of the following notions: U, the set of all samples in the dataset, along with their corresponding class labeling; A, the set of all variables; and B, the set of class labeling. The value of variable $q \in A$ in sample $x \in U$ is written as $f(x, q)$, which defines an equivalence relationship over U. With respect to a given q, the function partitions the universe into a set of pair wise disjoint subsets of U:

$$Rq = \{x: x \in U \wedge f(x, q) = f(x0, q) \ \forall \ x0 \in U\} \qquad (1)$$

Assume a subset of the set of variables, $P \subseteq A$. Two samples x and y in U are indiscernible with respect to P if and only if $f(x, q) = f(y, q) \ \forall \ q \in P$. The indiscernibility relation for all $P \subseteq A$ is written as IND (P). U/IND(P) is used to denote the partition of U given IND(P) and is calculated as $U/IND(P) = \{q \otimes P : U/IND(\{q\})\}$, where $A \otimes B = \{X \cap Y: \forall X \in A, \forall Y \in B, X \cap Y \neq \{\}\}$.

A rough set approximates traditional sets using a pair of sets named the lower and upper approximation of the set in question. The lower and upper approximations of a set $P \subseteq U$ (given an equivalence relation IND(P)) are defined as $PY = [\{X : X \in U/IND(P), X \subseteq Y\}$ and $PY = [\{X: X \in U/IND(P), X \cap Y \neq \{\}\}$ respectively. Assuming P and Q are equivalence relations in U, the important concept positive region POSP (Q) is defined as:

$$POSP(Q) = \bigcup_{X \in Q} PX(Q)$$

A positive region contains all patterns in U that can be classified in attribute set Q using the information in attribute set P. Thus, the concept of degree of dependency can be defined (Carpenter *et al*., 1991) as the degree of dependency γ (P, Q) of a set P of variables with respect to a set Q of class labeling is defined as

$$\gamma P(Q) = \frac{|POSP(Q)|}{|U|}$$

Where |S| denotes the cardinality of set S. The degree of dependency provides a measure of how important P is in mapping the dataset examples into Q. If $\gamma(P, Q) = 0$, then classification Q is independent of the attributes m in P, hence the decision attributes are of no use to this classification. If $\gamma = 1$, then Q is completely dependent on P, hence the attributes are indispensable. Values $0 < \gamma(P, Q) < 1$ denote partial dependency, which shows that only some of the attributes in P may be useful, or that the dataset was flawed to begin with. In addition, the complement of γ, gives a measure of the contradictions in the selected subset of the dataset.

Given P, Q and a variable $x \in P$, the significance σx (P, Q) of x in the equivalence relation denoted by P and Q is $\sigma x(P, Q) = \gamma(P, Q) - \gamma(P - \{x\}, Q)$. The higher the change in dependency, the more significant x is. RSAR takes advantage of this to remove variables that have little or no significance to the classification task at hand.

Rough Set Attribute Reduction (RSAR): Given a classification task mapping a set of variables C to a set of labeling D, a reduct is defined as any subset $R \subseteq C$, such that $\gamma(C,D) = \gamma(R,D)$. Given a classification task mapping a set of variables C to a set of labeling D, a reduct set is defined with respect to the power set $\wp(C)$ as the set $R \subseteq \wp(C)$ such that

$$R = \{A \in \wp(C): \gamma(A, D) = \gamma(C, D)\}.$$

That is, the reduct set is the set of all possible reducts of the equivalence relation denoted by C and D. Given a classification task mapping a set of variables C to a set of labeling D, and R, the reduct set for this problem space, a minimal reduct is defined as any reduct R such that $|R| \leq |A|, \forall A \in R$. That is, the minimal reduct is the reduct of least cardinality for the equivalence relation denoted by C and D.

The QuickReduct algorithm attempts to calculate a reduct without exhaustively generating all possible subsets. It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset. The existing algorithm is outlined below (Jensen and Shen, 2004a, b, 2005).

Quick Reduct (C, D)
Input: C, the set of all features attributes; D, the set of class attributes.
Output: R, the attribute reduct, $R \subseteq C$

```
$R \leftarrow \{\}$
do
    $T \leftarrow R$
    for each $x \in (C - R)$
        if $\gamma(R \cup \{x\}, D) > \gamma(T, D)$
            $T \leftarrow R \cup \{x\}$
    $R \leftarrow T$
until $\gamma(R, D) = \gamma(C, D)$
return R
```

Algorithm 1: The QuickReduct algorithm

According to the Quickreduct algorithm, the dependency of each attribute is calculated, and the best candidate is chosen. This, however, is not assured to find a minimal subset which has been proposed by Thangavel *et al.* (2006) with a potential solution to the problem by altering it into an n-look ahead approach. However, even this cannot promise a reduct unless n is equal to the original number of attributes, but this reverts back to generate-and-test. It still suffers from the same problem as the original Quickreduct, i.e., it is impossible to tell at any stage whether the current path will be the shortest to a reduct. In order to obtain the minimal reduct, apply the Improved Quickreduct Algorithm by using the best degree of dependency value by normalizing the information system. As in the normalization process in the data base system, the size of the information system can also be reduced horizontally by eliminating the objects which are involved in the construction of lower approximations.

CORRELATION

The usefulness of a feature or feature subset is determined by both its relevancy and redundancy. A feature is said to be relevant if it is predictive of the decision feature(s), otherwise it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. Hence, the search for a good feature subset involves finding those features that are highly correlated with the decision feature(s), but are uncorrelated with each other (John *et al*., 1994).

Like the majority of feature selection programs, Correlation based Feature Selection (CFS) uses a search algorithm along with a function to evaluate the merit of feature subsets. The heuristic by which CFS measures the goodness of feature subsets, takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The hypothesis on which the heuristic is based can be stated by the following equation.

$$Gs = \frac{kr_{ci}}{\sqrt{k + k\,(k\text{-}1)\, r_{ff}}}$$

Where, G_s is the heuristic merit of a feature subset s containing k features r_{ci} is the mean feature class correlation ($f \in s$) and r_{ff} is the average feature _feature intercorrelation.

The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator of how much redundancy there is among them. The heuristic goodness measure should filter out irrelevant features as they will be poor predictors of the class. Redundant features should be ignored as they will be highly correlated with one or more of the other features.

FUZZY ART

The best and most precise description of disease entities uses linguistic terms that are also imprecise and vague. Uncertainty is now considered essential to science and fuzzy logic is a way to model and deal with it using natural language. It can be said that fuzzy logic is a qualitative computational approach. Fuzzy logic is a method to render precise what is imprecise in the world of medicine.

Fuzzy art is a clustering algorithm that operates on vectors with analog-valued elements (Carpenter *et al*., 1991). Adding a further layer of processing to Fuzzy art yields a supervised clustering algorithm. Such applications often require the formation of thousands of clusters in a high dimensional feature space and could benefit from parallel implementation of the algorithm for high-speed or real-time classification.

Several properties of Fuzzy art facilitate implementation in hardware. Notably, because it uses the

L1 distance metric, multiplication is not required at each synapse. Furthermore, the algorithm can perform well with as few as four bits of weight precision (Rubin, 1995). Consequently, very little circuit area is required at each synapse in an electronic implementation. However, as originally specified, Fuzzy art requires bi-directional synapses, weight transport, or weight duplication, any of which is troublesome for a parallel implementation. This mapping relies on a different sequencing of the operations of the algorithm, but the classification of the input vectors remains unchanged.

Carpenter *et al.* (1991a) propose searching for the category J which maximizes Tj(I) and then checking whether $S_J(I)$ r. If not, category J is marked as ineligible (reset) and the search is repeated until a satisfactory category is found (resonance). The neural network realization described by Carpenter *et al.* (1991a) is three-layer architecture, illustrated in Fig. 1. Calculation of the choice function takes place in layer F2, whereas calculation of the match function and comparison to the vigilance parameter takes place in layer F1 and at the reset node. Since the weight values, wji, are required for both calculations, copies are stored and independently updated in each of the two layers. The length of time between input presentation and selection of the corresponding cluster is variable, depending on how many search cycles are required.

Fuzzy art clusters vectors based on two separate distance criteria, match and choice. For input vector I and category j, the match function is defined by

$$\underline{S_j(I) = |I \wedge W_j|}$$
$$|I|$$

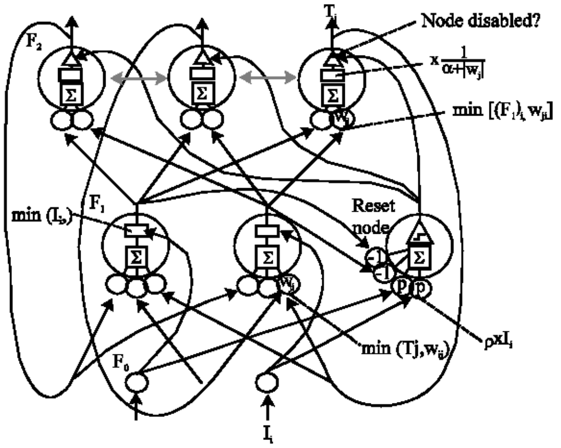

Fig. 1: The Fuzzy art architecture proposed by Carpenter *et al.* (1991a) consists of three layers of processing elements and utilizes two identical sets of weights

Where wj is an analog-valued weight vector associated with cluster j. Ù denotes the fuzzy and operator, $(p \wedge q)_i = \min(p_i, q_i)$ and the norm |. . .| is defined by $|p| \equiv \sum |p_i|$. The choice function is defined by

$$T_j(I) = \frac{|I \wedge W_j|}{\alpha + |W_j|}$$

Where α is a small constant. When the bias is increased, the search is more towards clusters with large wj. Each input vector is assigned to the category that maximizes Tj(I) while satisfying S_j (I) r, where the vigilance, r, is a constant, $0 \le r \le 1$.

It is therefore, desirable to develop techniques to provide the means of data reduction for crisp and real-value attributed datasets which utilize the extent to which values are similar. This can be achieved through the use of fuzzy-rough sets. Fuzzy-rough sets encapsulate the related but distinct concepts of vagueness and indiscernibility, both of which occur as a result of uncertainty in knowledge. Vagueness arises due to a lack of sharp distinctions or boundaries in the data itself. This is typical of human communication and reasoning. Rough sets can be said to model ambiguity resulting from a lack of information through set approximations.

RESULTS AND DISCUSSION

Making a blend of the three technologies of soft computing namely rough set theory, neural network and fuzzy logic, designing a fuzzy art classifier for the breast cancer diagnosis by reducing the feature set is the central focus of this study (Dubois and Prade, 1992). The RSAR acts as the preprocessor to handle all vital issue of feature selection and even after the application of the QuickReduct algorithm there is no guarantee regarding the removal of redundancy. Hence, the redundant features are eliminated using feature correlation. In the end, a fuzzy ART classifier has taken over the task of diagnosing with a reduced feature set and produced very reasonable results (Fig. 2).

The vertical reduct is possible using the Quick Reduct algorithm, only when the superfluous attributes are eliminated from the normalized data. But even after that, redundancy exists. Hence, it is necessary to fine tune the original algorithm to remove

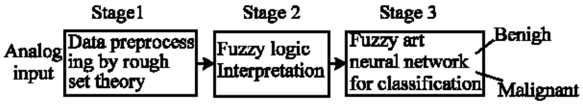
Stage1
Stage 2
Stage 3
Analog input
Data preprocessing by rough set theory
Fuzzy logic Interpretation
Fuzzy art neural network for classification
Benigh
Malignant

Fig. 2: Fuzzy art results

Table1: Comparison of feature selection accuracies of customary methods (17) and the proposed CRSAR

Database WBCD	Original no of features	ID3	SOAP	C4.5	RLF	CRSAR
Features and Accuracy (%)	10/95.01	9.1/94.57	5.2/94.64	7.0/95.42	5.7/95.02	2/96.24

the redundancy by incorporating feature correlation with correlation coefficient factor exceeding 90. The modified form of the original algorithm given

Correlated Rough Set Attribute Reduction (CRSAR)
Correlated Quick Reduct algorithm (C, D)
C-Set of highly correlated features.
D-Set of decision features.

```
$R \leftarrow \{\}$
do
    $T \leftarrow R$
    for each $x \in (C - R)$
        if $\gamma(R \cup \{x\}, D) > \gamma(T, D)$
            $T \leftarrow R \cup \{x\}$
    $R \leftarrow T$
until $\gamma(R, D) = \gamma(C, D)$
for each $x \in R$
    if FC $(x_i, x_R - x_i) > 90$
        remove $x_i$ from R
return R
```

Algorithm 2: Correlated quick reduct algorithm

The algorithm 2 has been tested on the popular breast cancer database namely Wisconsin Breast Cancer Database (WBCD). The WBCD consists of 683 samples with 16 missing attributes (Blake and Merz, 2006) has been employed in this research. The database consists of one id number and nine features obtained from fine needle aspirates, each of which is ultimately represented as an integer value between 1 and 10. The measured variables are as follows: (1) Clump Thickness (X1); (2) Uniformity of Cell Size (X2); (3) Uniformity of Cell Shape (X3); (4) Marginal Adhesion (X4); (5) Single Epithelial Cell Size (X5), (6) Bare Nucleoli (X6); (7) Bland Chromatin (X7); (8) Normal Nucleoli (X8) and (9) Mitoses (X9) where 444 of the data set belong to benign, and remaining 239 are malignant. The vigilance parameter r is set as 0.75 by trial and error method. Eighty percent of the reduced data set is used for training the network and the remaining 20% for testing. Table 1 compares the number of features and the accuracy for the traditional methods (Kohavi and John, 1997) with the method (CRSAR) outlined in this study.

CONCLUSION

The correlated QuickReduct algorithm has been described in this research and its efficiency has been compared with the well established methods reported in the literature. From the experimental results, one can very easily conclude that this approach has yielded better results with less features than the regular methods. This research can be further investigated with other neural networks in the ART family, SVM, etc. for breast cancer diagnosis.

REFERENCES

American Cancer Society. Breast Cancer Facts and Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

Alexios Chouchoulas, 2001. Incremental Feature Selection Based on Rough Set Theory- PhD Proposal Centre for Intelligent Systems and their Applications Division of Informatics, The University of Edinburgh.

Angela Torres and J. Juan, 2006. Nieto Fuzzy Logic in Medicine and Bioinformatics J. Biomed Biotechnol. Published online 2006.doi: 10.1155/JBB/2006/91908. pp: 91908.

Blake, C. and C. Merz, 2006. UCI Repository of Machine Learning Databases. Available at: http://www. ics.uci.edu/~mlearn/MLRepository.html.

Carpenter, G.A., S. Grossberg and D.B. Rosen, 1991a. Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks, 4: 759-771.

Carpenter, G.A., S. Grossberg and D.B. Rosen, 1991b. A neural network realization of Fuzzy ART (Technical Report CAS/CNS-91-021). Boston, MA: Boston University, Center for Adaptive Systems.

Chouchoulas, A. and Q. She, 2001. Rough set-aided keyword reduction for text categorisation. Applied Artif. Intelligence, 15: 843-873.

Drwal, G., 2000. Rough and fuzzy-rough classification methods implemented in RClass system. In: Proceedings of the 2nd International Conference on Rough Sets and Current Trends in Computing (RSCTC), pp: 152-159.

Dubois, D. and H. Prade,1992. Putting rough sets and fuzzy sets together. In Kluwer Academic Publishers, Dordrecht, pp: 203-232.

Ho, T.B., S. Kawasaki and N.B. Nguyen, 2003. Documents clustering using tolerance rough set model and its application to information retrieval. Studies In Fuzziness and Soft Computing, Intelligent Exploration of the Web, pp: 181-196.

Jensen, R. and Q. She, 2001. A Rough Set-Aided System for Sorting www Bookmarks. In N. Zhong *et al.* (Eds.). Web Intelligence: Research and Development, pp: 95-105.

Jensen, R. and Q. Shen, 2004. Fuzzy rough attribute reduction with application to web categorization. Fuzzy Sets and Sys., 141: 469-485.

Jensen, R. and Q. Shen, 2004. Semantics-preserving Dimensionality reduction: Rough and Fuzzy Rough-based approaches. IEEE Trans. Knowledge and Data Eng., pp : 16.

Jensen, R., 2005. Combining Rough and Fuzzy sets for feature selection. Ph.D Thesis, School of Informatics, University of Edinburgh.

John, G., R. Kohavi and K. Pfleger,1994. Irrelevant feature and the subset selection problem. In: Proceedings of the 11th International Conference on Machine Learning, pp:121-129.

Komorowski, J., L. Polkowski and A. Skowron, 1998. Rough sets: A tutorial. In: Rough-Fuzzy Hybridization: A New Method for Decision Making. Springer-Verlag.

Kohavi, R. and G.H. John, 1997. Wrappers for feature subset selection. Artif. Intelligence, pp: 273-324.

Mark, A.H. and A.S. Lloyd. Feature Subset Selection: A Correlation Based Filter Approach-American Association for Artificial Intelligence.

Pawlak, Z., 1982. Rough Sets. Int. J. Comput. Inform. Sci., 11: 341-356.

Pawlak, Z., 1991. Rough Sets: Theoretical Aspects and reasoning about Data. Kluwer Academic Publishers.

Rich, Caruana and D. Freitag, 1994. Greedy attribute selection. In: Proceedings of the 11th International Conference on Machine Learning, Morgan Kaufmann, pp: 28-36.

Rubin, M.A., 1995. Issues in automatic target recognition from radar range profiles using Fuzzy ART map. In The 1995 World Congress on Neural Networks Mahwah, NJ: Lawrence Erlbaum Associates, pp: 197-202.

Sebban. M. and R. Nock, 2002. A hybrid filter/wrapper approach of feature selection using information theory. Pattern Recog., 4: 835-846.

Swiniarski, R.W., 1996. Rough set expert system for online prediction of volleyball game progress for US olympic team. In: Proceedings of the 3rd Biennial European Joint Conference on Engineering Systems Design Analysis, pp: 15-20.

Thangavel, K., A. Pethalakshmi and P. Jaganathan, 2006. A Comaparative analysis of feature selection algorithms based on Rough set theory. Int. J. Soft Comput., 1: 288-294.

Zdzis, law Pawlak, 1991. Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer, Academic Publishers, Dordrecht, 4: 16-18.