2011 11th International Conference on Control, Automation and Systems
Oct. 26-29, 2011 in KINTEX, Gyeonggi-do, Korea

Feature Reduction using a GA-Rough Hybrid Approach on Bio-medical data

Chang Su Lee

School of Computer and Security Science,
Edith Cowan University, Mt. Lawley, WA Australia
(E-mail: chang-su.lee@ecu.edu.au)

Abstract: In this paper, a new approach is proposed for feature reduction using a GA-Rough hybrid approach on Bio-medical data. The given set of bio-medical data is pre-processed with the min-max normalization method. Then the subsequent evaluation on each feature with respect to the output class is carried out utilizing the information gain-based approach using the entropy-based discretization. Features with zero worth on the evaluated set of features are eliminated. The genetic algorithm is applied for performing a search for most relevant features on the set of features remained. These processes continue until there is no further change on the final reduced set of features. The rough set-based approach is applied on this set of features by applying discernibility matrix-based approach in order to obtain the final reduct. The reduced set of features, or a final reduct, is tested for classification using a TS-type rough-fuzzy classifier to show the viability of the proposed feature reduction approach. The results showed that the proposed feature reduction approach effectively achieved to reduce number of features significantly which reduced to 7 out of 120 features along with compatible classification results on the given bio-medical data compared to other approaches.

Keywords: feature reduction, information gain, rough set theory, genetic algorithm, GA-rough hybridization.

1. INTRODUCTION

Feature reduction refers to the elimination process of features that are not significantly relevant to the IO mapping relation for the given knowledge. This is one of the problems encountered in many areas such as pattern recognition and signal processing. There is often a significant amount of redundant or misleading information among the given data, and this will need to be reduced or removed before the main process for data classification is applied on it. The high dimensionality can be governed by just a few variables which are much more relevant than others to present the given information without losing the same classificatory power. In order to resolve and improve this, many approaches have been proposed so far; the PCA (Principle Component Analysis), the ICA (Independent Component Analysis), GA (Genetic Algorithm)-based methods, etc.

In this paper, a new approach is proposed for an effective feature reduction utilizing a GA-Rough hybrid approach. This approach employs the followings; 1) an evaluation on features using information gain-based method with the entropy-based discretization [1], 2) a simple GA search process [2] for finding the most relevant features with respect to the given output class information, 3) a rough set-based feature selection and reduction approach utilizing a discernibility matrix-based method, and 4) a classification on the testing set of the given data by deploying a adaptive TS-type rough-fuzzy inference system, ARFIS [9], [10].

This paper is organized as follows: Sections 2, 3, 4 provide brief descriptions on an entropy-based feature evaluation, a simple GA search process used, and the rough set theory employed. Section 5 presents the design of the proposed GA-Rough hybrid approach for the feature reduction process. Experimental results on the Alzheimer's disease data are shown with results on feature reduction along with classification results compared to other approaches, and final conclusions are drawn in Section 7.

2. EVALUATION ON FEATURES USING AN ENTROPY-BASED METHOD

In order to evaluate a worth of each feature of the given data set, an information gain-based evaluation using the entropy-based discretization method is employed, which was introduced by Fayyad and Irani [1]. The entropy-based method uses the given class information entropy of candidate partitions to select boundary points for discretization. Class information entropy is a measure of purity and it measures the amount of information that would be needed to specify to which class an instance belongs.

The entropy, *E(S)* defined by (1) is calculated on the given set of data samples *S* with respect to the given *N* number of output class information, where p_i = (number of samples that belong to class *i*)/(total number of samples). If the given *S* is partitioned, for example, into two intervals S_1, S_2 using a boundary point *T*, the entropy after partitioning is defined by (2). The boundary point *T* is selected from the mid-points of the given attributes values. The notation | • | indicates a cardinality of a given set.

$$E(S) = -\sum_{i=1}^{N} p_i * \log_2(p_i) \tag{1}$$

$$E(S,T) = \sum_{j=1}^{2} \frac{|S_j|}{|S|} E(S_j) \tag{2}$$

$$G(S,T) = E(S) - E(S,T) \tag{3}$$

The goal of this method is to find the best *T* which produces the maximum information gain *G* defined by

(3). The best mid-point T is found by examining all possible splits and then selecting the optimal one. The boundary that minimizes the entropy over all possible cases is chosen as an optimal one. Once the best T was found, the given attributes values are converted into the labels corresponding to their discretized category.

3. GENETIC ALGORITHM FOR FEATURE SELECTION

The genetic algorithm used in this paper is for performing a search for most relevant features using the simple genetic algorithm described in research by Goldberg [2]. This GA method is carried out along with an evaluation method that scores the worth of a subset of attributes (a population) by considering the individual predictive ability of each feature along with the degree of redundancy between them.

With randomly selected initial populations which are subsets of features, the GA process is performed to measure their merits (fitness) in terms of their generalization ability with respect to their output class information under the supervised manner. After the GA process is done by the provided stopping criteria (e.g., number of generations), all populations are sorted in ascending order by merits evaluated for each individual. The population that has the height merit score is used to select only features which are in this population.

4. ROUGH SET THEORY

Rough set theory was developed by Pawlak [3] as a mathematical tool to deal with the classificatory analysis on imprecision, vagueness and uncertainty of the given data. The main objectives of the rough set analysis are to estimate approximations of concepts, to find the most significant attributes based on the gathered data, and to generate decision rules for classification process.

In regards to feature selection and reduction, an information system, or a given set of data samples S, may be represented by a discernibility matrix and a discernibility function. They help to construct efficient algorithms related to a generation of minimal subsets of attributes which are sufficient to describe concepts in a given information system. The aim of the knowledge reduction is to obtain irreducible, but essential parts of the knowledge. A discernibility matrix for an information system S with the set of attributes Q is a $n \times n$ matrix such that

$$m_{ij} = \{a \in Q : f(x_i, a) \neq f(x_j, a)\} \quad (4)$$

, where $U = \{x_1, x_2, \ldots, x_n\}$ is a universe - a finite and non-empty set of n objects, $Q = \{a_1, a_2, \ldots, a_m\}$ is a finite and non-empty set of m attributes. Each entry m_{ij} in the given i-th row and j-th column of the discernibility matrix is a subset of attributes that discerns all objects in U.

An information system can be represented as a decision table, $DT = < U, C \cup D >$ if the set of attributes Q can be represented by a set of condition attributes C, and a set of decision attributes D, in this form $Q = C \cup D$. In a decision table, every pair of condition and decision attributes determines the implication of $C \rightarrow D$ and a set of decision rules constitutes a decision algorithm. An elimination of redundant attributes is required in order to obtain a minimal set of decision rules. If $C^* = \{X_1, X_2, \ldots, X_p\}$ and $D^* = \{Y_1, Y_2, \ldots, Y_q\}$ are defined as a C-definable set and a D-definable set of U, a set of decision rules r_{ij} for all D-definable sets Y_j is defined by (5).

$$\{Des_C(X_i) \Rightarrow Des_D(Y_j) : X_i \bigcap Y_j \neq \phi, \forall X_i \in C^*, \forall Y_j \in D^*\} \quad (5)$$

, where $Des_C(X_i)$ and $Des_D(Y_j)$ are unique descriptions of the sets (classes) X_i and Y_j, respectively.

5. GA- ROUGH HYBRID APPROACH FOR FEATURE REDUCTION

Once the pre-processing is done using the max-min minimization on the given raw data, the entropy-based feature evaluation and the GA procedures are applied in an iterative process until there is no change on the final reduced set of features as stated earlier.

Using the final reduced set of features the rough set-based feature reduction approach will be performed on the discretized given data. The discretized (clustered) data table, however, may contain much redundant and conflicting data. One of the main advantages in using the rough set methodology is that it reduces the given knowledge using the degree of dependency of attributes. This process requires finding minimal sets of condition attributes, or *reducts* with respect to the decision (class) attribute in order to obtain the smallest possible number of decision rules for higher compactness.

In order to reduce the number of features, an algorithm based on the decision-relative discernibility matrix [4] with Boolean calculation is employed for the reduction of attributes. The algorithm is described as follows. First, obtain the discernibility matrix, m_{ij} defined by (6) with respect to the decision attribute, d.

$$m_{ij} = \begin{cases} \{a \in C : f(x_i, a) \neq f(x_j, a)\} & d(x_i) \neq d(x_j) \\ 0 & d(x_i) = d(x_j) \end{cases} \quad (6)$$

Calculate Tij, the disjunctive Boolean expressions with the entries of the discernibility matrix as defined by (7).

$$T_{ij} = \{ \bigvee_{a_i \in m_{ij}} a_i : m_{ij} \neq 0, m_{ij} \neq \phi \} \quad (7)$$

Compute the Boolean expression in conjunctive normal form as defined by (8).

$$T = \bigwedge_{m_{ij} \neq 0, m_{ij} \neq \phi} T_{ij} \quad (8)$$

Calculate the Boolean expression in disjunctive normal form as defined by (9).

$$T^* = \bigvee_i T_i \quad (9)$$

Finally, find a minimal set of attributes, or a *reduct* which has the least number of attributes from the normal form of T^*.

This feature reduction process is performed using the discernibility matrix-based algorithm in the rough set

approach. The final reduct obtained on the training set of the given Alzheimer's data, is a set of seven attributes; {EGF, IL1α, IL3, MCP3, MIP1δ, TNFα, IL11}. All these 7 attributes belong to 18 bio-markers found in Ray et al.'s research [6], which are {EGF, GDNF, IL1α, IL3, MCP3, MCSF, MIP1δ, PARC, PDGFBB, RANTES, TNFα, ANG2, GCSF, ICAM1, IGFBP6, IL11, IL8, TRAILR4}. Also the 4 attributes {EGF, IL1α, IL3, TNFα} from our 7 features are in common with 5 bio-markers {EGF, IL1α, IL3, TNFα, GCSF} found by Ravetti and Moscato's research [7]. It can be stated that the proposed feature reduction method in the rough set approach may have a significant contribution to find the most relevant bio-markers on the given Alzheimer's disease data set.

6. EXPERIMENTS AND RESULTS

In order to show the viability of the proposed system, a comparison on results of feature reduction against the other existing approaches is shown in this section with experiments carried out using the proposed GA-rough hybrid approach. To perform the proposed approach, the Weka software package [5] and a recently generated Alzheimer's disease data by Ray et al. [6] are utilized.

6.1 Alzheimer's disease data and other approaches

The Alzheimer data set used in this paper is from a research work performed by Rat et al. [6]. They collected a total of 259 archived plasma samples from individuals with pre-symptomatic to late-stage Alzheimer's disease and from various controls and measured the abundance of 120 known signaling proteins in these samples. Their Alzheimer's disease (AD) and Non-Demented Control (NDC) samples were divided equally into a training set for predictor discovery and supervised classification and a test set for class prediction of blinded samples. The training data set has 83 data samples - 43 AD and 40 NDC with 120 known attributes. The testing data set has 92 data samples - 42 AD, 39 NDC, and 11 OD (Other Dementias) with the same 120 attributes. This is shown in Table I below.

TABLE I
ALZHEIMER'S DISEASE DATA FROM RAY ET AL[6].

Classes / Data Set	*AD*	*NDC*	*OD*	Attributes
Training	43	40		120
Testing	42	39	11	120

In Ray et al.'s research [6], they employed the following two approaches. Firstly, the statistical analysis of the training set is performed by significance analysis of microarrays (SAM) which identified 19 proteins with highly significant differences. Through the subsequent unsupervised clustering analysis on the training set, in the next procedure the predictive analysis of microarrays (PAM) is done with a shrunken centroid algorithm on the clustered training data. The PAM identified 18 predictors out of the total 120 proteins which are {ANG-2, CCL5/RANTES, CCL7/MCP-3, CCL15/MIP-1δ, CCL18/PARC, CXCL8/IL-8, EGF, G-CSF, GDNF, ICAM-1, IGFBP-6, IL-1α, IL-3, IL-11, M-CSF, PDGF-BB, TNF-α, TRAIL-R4}. The PAM classified Alzheimer's and NDC samples on the training set with 95% positive (AD) and 83% negative (NDC) agreements with the clinical diagnosis as shown in Table II. With the 18 bio-markers found, the classification on the testing set is done using the PAM-class prediction algorithm. The PAM classified the testing data including OD samples with 90% positive (AD) and 88% negative (NDC & OD) agreements with the clinical diagnosis as shown in Table III. The 10 out of the 11 other dementia samples were classified as 'non-ADs'.

TABLE II
CLASSIFICATION RESULTS ON THE TRAINING DATA IN RAY ET AL. [6]

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	41	2	95 (84-99)
NDC	7	33	83 (67-93)
Overall			**89 (p<0.001)**

TABLE III
CLASSIFICATION RESULTS ON THE TESTING DATA IN RAY ET AL. [6]

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	38	4	90 (77-97)
NDC	5	34	88 (76-95)
OD	1	10	
Overall			**89 (p<0.001)**

In another recent study by Ravetti and Moscato [7], the same Ray et al.'s molecular dataset was used which has attracted worldwide attention on the Alzheimer's disease research. This research showed that improved results were obtained with the abundance of only 5 proteins – {EGF, IL1α, IL3, TNFα, GCSF} which achieved 96% classification accuracy on average in predicting the clinical AD results. This research utilized an integrative data analysis which consists of the following for steps: 1) abundance quantization and 2) feature selection using Fayyad and Irani's entropy-based discretization [1], 3) literature analysis, and 4) selection of a classifier algorithm that is independent of the feature selection process. The first two methods created an instance of the (alpha-beta)-k-Feature Set problem [8]. Fayyad and Irani's method filtered only 14 out of the total 120 proteins of the training set. After the removal of the 3 conflict training samples, the numerical solution of the (alpha-beta)-k-Feature Set problem led to the selection of only 10 bio-markers. Using an undirected graph of the known important association of these 10 proteins, the maximum subset of proteins was identified as a core set of the 10 signatures, which is {EGF, IL1α, IL3, TNFα, IL6, GCSF}. Among

these 6, the IL6 was ignored as they claimed that many classifiers in Weka did not make use of its abundance to inform decisions. To test the selected final 5attributes, the Simple Logistic (SL) classifier in the Weka software package was employed for pattern classification task. With the 5 protein signatures, the SL classifier performed 86% of classification accuracy after 10 times of the 10-fold Cross-Validation (CV) over the training set. On the test set the 6 features found with the SL method achieved 97% of accuracy – 100% positive agreement on AD and 92% negative on NDC. These results are shown in Table IV and V.

TABLE IV
CLASSIFICATION RESULTS ON THE TRAINING DATA IN RAVETTI. [7]

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	34.9	6.1	85
NDC	5.6	33.4	86
Overall			**85.4 (p<0.01)**

TABLE V
CLASSIFICATION RESULTS ON THE TESTING DATA IN RAVETTI. [7]

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	42	0	100
NDC	3	36	92
OD	1	10	
Overall			**96 (p<0.01)**

6.2 The proposed GA-Rough Hybrid approach for feature selection

In our experiments on the Alzheimer's disease data, the final 7 features of proteins {EGF, IL1α, IL3, MCP3, MIP1δ, TNFα, IL11} are found by the rough set-based feature reduction process. Comparing with other sets of proteins found by Ray et al. [6] and Ravetti and Moscato [7], the following relations can be obtained between the following three sets of proteins found.

$5_{\text{Ravetti}} \subset 18_{\text{Ray}}$

$7_{\text{GA-ROUGH}} \subset 18_{\text{Ray}}$

$7_{\text{GA-ROUGH}} \cap 5_{\text{Ravetti}}$ = { EGF, IL1α, IL3, TNFα } (10)

Using the 7 bio-markers found, a framework for Adaptive Rough-Fuzzy TS-type Inference Systems (ARFIS) [9], [10] is deployed for classification. The classification results on the training and the testing sets are shown in Table VI and VII below.

TABLE VI
CLASSIFICATION RESULTS ON THE TRAINING DATA USING THE PROPOSED GA-ROUGH-FUZZY HYBRID APPROACH

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	32	11	74.42
NDC	3	37	92.50
Overall			**83.46**

TABLE VII
CLASSIFICATION RESULTS ON THE TESTING DATA USING THE PROPOSED GA-ROUGH-FUZZY HYBRID APPROACH

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	39	3	92.86
NDC	6	33	84.62
OD	0	11	
Overall			**88.74**

This classification accuracy on the testing set is quite compatible with the 89% result from Ray et al.'s work [6]. However, this result on the testing set is much lower than the 96% from Ravetti and Moscato's [7].

Therefore, for more investigation, the following set of experiments were carried out for more comparisons: classification using the GA-rough-fuzzy hybrid approach with 1) the 18 proteins found by Ray et al. [6], 2) the 5 found by Ravetti and Moscato [7], and 3) the 4 found in common between $7_{\text{GA-ROUGH}}$ and 5_{Ravetti} in (10). Classification results using the proposed GA-rough-fuzzy hybrid approach for each configuration are shown in tables from Table VIII to XIII.

TABLE VIII
CLASSIFICATION RESULTS ON TRAINING SET FOR THE EXPERIMENT 1)

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	38	5	88.37
NDC	2	38	95
Overall			**91.69**

TABLE IX
CLASSIFICATION RESULTS ON TESTING SET FOR THE EXPERIMENT 1)

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	37	5	88.01
NDC	4	35	89.74
OD	0	11	
Overall			**88.92**

TABLE X
CLASSIFICATION RESULTS ON TRAINING SET FOR THE EXPERIMENT 2)

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	35	8	81.40
NDC	5	35	87.50
Overall			**84.45**

TABLE XI
CLASSIFICATION RESULTS ON TESTING SET FOR THE EXPERIMENT 2)

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	41	1	97.62
NDC	4	35	89.74
OD	1	10	
Overall			**93.68**

TABLE XII
CLASSIFICATION RESULTS ON TRAINING SET FOR THE EXPERIMENT 3)

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	36	7	83.72
NDC	6	34	85.00
Overall			**84.36**

TABLE XIII
CLASSIFICATION RESULTS ON TESTING SET FOR THE EXPERIMENT 3)

Class / Class	*AD*	*NDC*	Classification Accuracy (%)
AD	39	3	92.86
NDC	4	35	89.74
OD	0	11	
Overall			**91.30**

As shown in tables from Table VI to XIII, the following points can be observed.

• The classification results produced on the testing set by the proposed GA-rough-fuzzy hybrid approach are quite compatible with the results achieved by Ray et al.'s, and Ravetti and Moscato, i.e., $88.92\%_{ARFIS}$ compared to $89\%_{Ray}$ using the 18 proteins found, and $93.68\%_{ARFIS}$ compared to $96\%_{Ravetti}$ using the 5 proteins found respectively. The difference on results in between $93.68\%_{ARFIS}$ and $96\%_{Ravetti}$ is one mis-classified point for each class.

• The 7 bio-markers found by the proposed GA-rough hybrid feature reduction approach achieved quite similar classification results on the testing set, i.e., $88.74\%_{ARFIS}$ compared to $89\%_{Ray}$.

• However, it is much lower performance when it is compared to $96\%_{Ravetti}$. Thus, the 4 features in (10) are used to investigate towards higher accuracy. The 4 proteins produced slightly higher classification rate $91.3\%_{ARFIS}$ using the proposed GA-rough-fuzzy hybrid.

7. CONCLUSION

The proposed approach reduced the given number of features (bio-markers) significantly by the GA-rough hybrid feature reduction, and achieved compatible classification results using the rough-fuzzy classifier compared to two recent studies on Alzheimer's disease research. Also the results produced by the proposed approach under different experiment configuration showed that the set of reduced features achieved quite compatible classification accuracies compared to ones from the literatures. It can be concluded that the proposed GA-rough hybrid feature reduction approach is highly efficient in feature reduction and also in classification on the given Alzheimer's data.

REFERENCES

[1] U.M. Fayyad, K.B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," Proc. 13th International Joint Conference on Artificial Intelligence (IJCAI 1993), San Francisco, CA, pp. 1022-1027, 1993.

[2] D. E. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley, 1989.

[3] Z. Pawlak, "Rough Sets," International Journal of Computer and Information Science, vol. 11, pp. 341-356, 1982.

[4] R. Jensen and Q. Shen, "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches," IEEE Trans on Knowledge and Data Engineering, vol. 16, no. 12, pp. 1457-1471, Dec 2004.

[5] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/~ml/weka/, University of Waikato, New Zealand.

[6] Ray et al., "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins," Nature Medicine, pp. 1-4, Nature Publishing Group, October 2007, doi:10.1038/nm1653.

[7] M.G. Ravetti, P. Moscato, "Identification of a 5-Protein Biomarker Molecular Signature for Predicting Alzheimer's Disease," PLoS One, vol. 3, issue 9, e3111, pp. 1-12, Sept 2008, doi:10.1371/journal. pone.0003111.

[8] R. Berretta, W. Costa, P. Moscato, "Combinatorial Optimization Models for Finding Genetic Signatures from Gene Expression Datasets", In: Keith JM, ed. Bioinformatics, vol. II: Structure, Function and Applications Humana Press, 2008.

[9] C. Lee, A. Zaknich, T. Bräunl, "A Framework of Adaptive T-S type Rough-Fuzzy Inference Systems (ARFIS)," IEEE 2008 International Conference on Fuzzy Systems (FUZZ-IEEE 2008), Hong Kong, pp. 567-574, June 2008.

[10] C. Lee, A. Zaknich, T. Bräunl, A Framework of Adaptive T-S type Rough-Fuzzy Inference Systems (ARFIS), Ph.D. Thesis, Department of Electrical, Electronic, and Computer Engineering, The University of Western Australia, Perth, Australia, 2008.