Incorporating Genetic Algorithm into Rough Feature Selection for High Dimensional Biomedical Data

Vinh Quoc Dang, Chiou-Peng Lam, Chang Su Lee
School of Computer and Security Science
Edith Cowan University
Western Australia, Australia
vdang@our.ecu.edu.au, {c.lam, chang-su.lee}@ecu.edu.au

***Abstract*—In this paper, a hybrid approach incorporating genetic algorithm and rough set theory into Feature Selection is proposed for searching for the best subset of optimal features. The approach utilizes K-means clustering for partitioning attribute values, the rough set-based approach for reducing redundant data, and the genetic algorithm for searching for the best subset of features. A set of six attributes was obtained as the best subset using the proposed algorithm on the colon cancer dataset. Classification was carried out using this set of six attributes with 23 classifiers from WEKA (Waikato Environment for Knowledge Analysis) software to examine their significance to classify unseen test data. In addition, the set of 6 genes found by the proposed approach was also examined for their relevance to known biomarkers in the colon cancer domain.**

Keywords: feature selection, k-means clustering, genetic algorithm, rough set theory, pattern classification

I. INTRODUCTION

Recently, mass throughput technologies such as microarrays and mass spectrometry (MS) have been developed and used widely in the bioinformatics domain. Microarray data have been used in classification of various types of cancers such as colon, prostate, breast, skin, lymphoma and many others [1]. Microarrays and mass spectrometry data are extremely high in dimension and suffer 'the curse of dimensionality' [2], resulting in over-fitting in learning, too much redundant information for classification, and a high computational cost in the search for optimal solutions through the feature subspaces [3]. Owing to the high complexity of high dimensional data, it is very important that the number of features is reduced in order to perform data analysis and processing with less computational cost. Feature selection (FS) has been widely employed as a technique to select a subset with a minimal number of features that describes the given dataset without losing any inherent information [9]. Rough Set Theory (RST) [5] has also been used to generate reducts, or sets of a minimal number of the most relevant features, in order to reduce the given data efficiently into an optimal compact form. Genetic Algorithm (GA) [6] has also been utilized as another methodology to optimize the search to find the most significant features by the evaluation on the fitness of chromosomes.

In this paper, a hybrid algorithm, incorporating GA and Rough Feature Selection, is proposed for finding the optimal subset of features. Unlike Banerjee, Mitra and Banka's approach [4], K-means clustering, incorporating quartile statistics was applied first to partition each attribute of the training set. RST was then utilized to reduce the given high dimensional data that were re-coded using thresholds from the final clusters generated via K-means clustering. The best subset of features was found by employing GA. The generation of a distinction table in RST and the fitness function used in GA were based on Banerjee *et. al*. [4]. A tuning process for values of GA parameters was carried out for the best performance of the GA. The proposed approach was evaluated using a colon cancer dataset [15]. Once the optimal subset of features was found, further evaluation was performed on an unseen test dataset using 23 independent classifiers. Experimental results showed that the best 6 genes found from this data set had higher classification accuracies than that of the 15 genes found by Banerjee et al. [4]. The significance of these 6 genes to the medical domain, colon cancer is also outlined in the discussion.

The rest of the paper is organized as follows: Section 2 describes previous work related to FS methods, including the applications of rough set–based approaches and genetic algorithms for feature selection. Section 3 describes preliminary basics for K-means clustering, RST, and GA. The proposed GA-based rough feature selection approach is described in Section 4, followed by a description of a parameter tuning process for GA in Section 5. Experiments with results are shown in Section 6, and conclusions are drawn with discussions in Section 7.

II. RELATED WORK

Xu, Liu and Huang [23] developed a method that used statistical characteristics of the data to select features. These statistics include the correlation coefficient between variables, the large variance between the variables in different classes, and the probability density of the variables related to small round blue-cell tumor (SRBCT) cancer data. Lopes, Martins and Cesar [7] successfully implemented two wrapper FS algorithms: Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS) for classifying microarray breast cancer data with a high classification accuracy. The RS approaches have also been used effectively in feature selection whereby the number of attributes in a dataset is reduced by removing irrelevant attributes [12]. In the study of Wang, Chen, Li and Zhang

978-1- 61284-704-7/11/$26.00 ©2011IEEE

[8], RST has been applied to 2 microarray cancer data sets, colon and leukemia; and the results showed that the RST approach was able to select the relevant feature gene subsets, and achieving high classification accuracy. Another approach in feature selection that has been used to optimize the search for feature subsets is GA. GA was also used in conjunction with SVM in Peng et al. [9], for selecting features from NCI60 and GCM cancer data sets. The algorithm finds a smaller number of relevant genes, with higher accuracy than previous methods such as rank-based gene selection and all paired binary support vector machine (AP-SVM) [9]. Banerjee *et al*. [4] proposed an evolutionary rough FS based on the RST with a variant of the discernibility matrix-based approach, namely a distinction table, to reduce the large number of redundant features in three cancer datasets, colon, leukemia and lymphoma, and followed by the use of Multi-objective genetic algorithm (MOGA) to search for optimal feature subsets.

III. PRELIMINERIES

This section describes K-Means clustering, RST and GA. The K-means clustering method was proposed by MacQueen [10]. It is an unsupervised learning algorithm that can be employed to separate data into different clusters. It starts with a random centroid for each of the k groups, after that placing each data point to the group that has a closest centroid, and then the positions of the K centroids are recalculated for new data points. This process repeats until the centroids for each group has no change i.e. convergence takes place.

RST was proposed by Pawlak in 1982 [5] and has a mathematical basis. It has been used in data mining to classify vague, uncertain or incomplete data. The rough set is based on equivalence classes containing samples that are identical in terms of attributes describing the data. The rough set describes a set by using lower and upper approximations of the given set. The lower approximation of the set consists of all the samples that can be described as definitely belonging to the set, “positive cases”, whilst the upper approximation of the set consists of all the samples that are described as possibly belonging to the set, “possible cases” [11].

GA was proposed and developed by Holland in 1975 [13] and is based on Darwin’s theory of survival of the fittest. GA consists of components such as population representation, fitness function, selection, crossover, and mutation operators. GA starts with an initial population, which is generated randomly from a population pool. The fitness for each individual is evaluated. The algorithm applies a selection method such as tournament selection to select individuals. Only the fittest individuals are selected and considered as potential parents to produce offspring. Crossover operator is used to combine parents to produce offspring. After the crossover process is done, offspring are mutated to produce new individuals with different features, which are not present in their parents. At this stage optimization is assessed again based on the new generation to determine if the optimal solution has been achieved. The algorithm stops only when a solution is found or some other criteria are met such as a pre-determined number of iterations. Otherwise the process cycles back to the selection step.

IV. THE PROPOSED APPROACH – GENETIC ALGORITHM (GA)-BASED ROUGH FEATURE SELECTION

The proposed approach, GA-based Rough Feature Selection, utilizes K-means clustering for partitioning each attribute, the RST with a variant of the discernibility matrix-based method [4] to reduce the number of high dimensional features, and GA to search for optimal subsets from the outputs of RST. In determining threshold values for each attribute, K-means clustering is selected to obtain three different cluster centroids for low, mid and high levels of each feature (bio-marker). Cluster centroids for these levels for each feature would represent a center point in different level of each bio-marker measure based on the similarity distance measure, for instance Euclidean distance, between data samples. This approach is utilized in order to find more effective intervals in terms of distribution of collected data, rather than employing a simple statistics-based manner using quartiles [4], which gives a single fixed constant threshold for each level of each attribute. Appropriate threshold values that capture the differences between diseased versus non-diseased will lead to better representation of data for finding an optimal feature subset.

The proposed approach consists of a number of steps outlined below.

Step 1: Normalization

Attribute values are converted into the range of [0 1] by applying the Min-Max normalization on each subspace for each attribute domain.

Step 2: Calculation of Thresholds using K-means clustering

The procedure is as follows for the K-means clustering incorporating the quartile statistics to calculate threshold values for each attribute:

i. Values for each attribute are sorted in ascending order.

ii. Divide the measurements into small class intervals of equal width (δ).

iii. Calculate quartile statistics. The lower and upper thresholds are calculated using the formula defined in [4] and as shown below

$$Th_k = L_c + (R_k - cfr_{c-1})/fr_c * \delta \quad (1)$$

, where L_c is lower limit of the c_{th} class interval, R_k is the rank of the k_{th} interval value; R_k = (N*k)/(number of partitions), N is a number of objects, k is k^{th} partition value, k=1,2,3 for 4 partitions, cfr_{c-1} is the cumulative frequency of the immediately preceding class interval such that $cfr_{c-1} <= R_k <= cfr_c$, fr_c is the class frequency

iv. Unlike Banerjee et al.’s approach [4] that uses the three threshold values from step (iii) for subsequent processing, this approach uses these thresholds as

initial centroid values for K-means clustering of each attribute with k =3. Output: 3 centroid values.

Step3: Generation of Reduced Attribute Value Table

Using the final centroids values obtained step 2(iv), assign thresholds Th_l and Th_u as the lower and upper thresholds which are used subsequently for converting attribute values of each attribute using the following conditions:

- If an attribute value is less than or equal Th_l then insert 0 into the table. If the value is greater or equal to Th_u then insert 1, otherwise insert '*' to represent a 'don't care' condition [4].

The average frequency of '*'s in the table is calculated. This is then used as a threshold value (th_a) to eliminate attributes with a total number of '*' greater than or equal to th_a and subsequently produces a reduced attribute value table (A_r).

Step 4: Generation of Distinction Table

The A_r table is used to create a distinction table [4]. The distinction table is a variant of discernibility matrix-based approach with special conditions in order to reduce its dimension from m * (m-1)) / 2 to m1 * m2, where m = m1*m2, m1 = total number of samples in Class1, m2 = number of samples in Class2.

- The conditions to insert object pairs to the d-distinction table are:

$$b\big((k,j),i\big) = 1 \text{ if } a_i(x_k) \neq a_i(x_j)$$

$$b\big((k,j),i\big) = 0 \text{ if } a_i(x_k) = a_i(x_j)$$

such that $\forall(k,j)\exists i \epsilon R: b((k,j),i)=1$ where $d(x_k) \neq d(x_i)$.

- Either object in the pair has '*', then insert 0 for its entry.
- Object pairs of the same class are ignored.

Step 5: Application of Single Objective GA

- Single objective GA is employed in this study aggregating the two objective functions, f1 and f2 [4] as a single objective, f, defined in (2).

$$f = \alpha_1 * f_1 + \alpha_2 * f_2 \quad (2)$$

, where $\alpha_1 = 0.9$, $\alpha_2 = 0.1$ (adapted from Banerjee *et al*. [4]),

$f_1(\underset{v}{\rightarrow}) = \frac{N - L\underset{v}{\rightarrow}}{N}$, $f_2\left(\underset{v}{\rightarrow}\right) = \frac{C\underset{v}{\rightarrow}}{(m1 * m2)}$, $f_1(\underset{v}{\rightarrow})$, $f_2\left(\underset{v}{\rightarrow}\right)$: fitness function 1 and 2 of the candidate reduct,

N: length of reduct, $L\underset{v}{\rightarrow}$: number of 1s in the candidate reduct, m1: number of objects in class 1, m2: number of objects in class 2, $C\underset{v}{\rightarrow}$: number of objects can be discerned by the candidate reduct.

The objective function, *f*, guides the algorithm to find a relevant feature subset, which has less number of features but gives more accuracy in discerning between objects.

In order to carry out experiments, parameter settings are needed for all the parameters associated with GA. The following section describes the parameter tuning process to obtain the best parameter set for GA.

V. PARAMETER TUNING FOR GA

Convergence of fitness in GA is important as premature convergence will result in a local optimal solution. Parameter tuning for GA is necessary to ensure that the algorithm performed with the best parameter setting, because "the choice of crossover probability, *Pc* and mutation probability, *Pm* is known to be critical and will affect the behavior and performance of the GA" [14]. Parameter tuning for *Pc* and *Pm* in this study is based on the approach of measuring GA convergence in Srinivas and Patnaik's experiment [14]. This involves varying different values of *Pc* and *Pm* and observing the average fitness (favg) and maximum fitness value (fmax) of the population to determine whether the GA converges to a local (e.g., fitness value of 0.6) or global optimum (e.g., fitness value of 1). The difference between fmax – favg is a measurement to determine the convergence. The smaller the difference between fmax and favg, the better convergence for the algorithm and a better optimal solution [14].

As a result of parameter tuning, the best parameter values of *Pc* and *Pm* are 0.7 and 0.03 respectively. These parameters were used to run the GA and the convergence is illustrated in Figure 1.

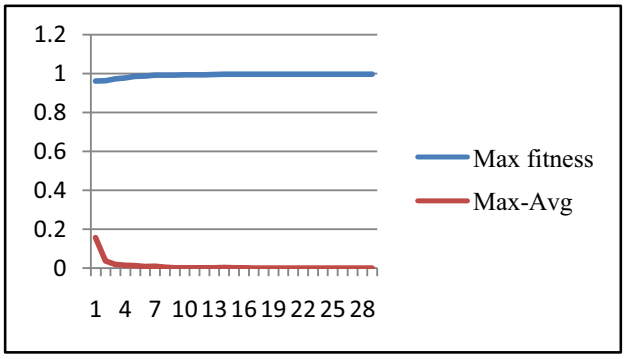

Figure 1. Example of convergence for maximum and (maximum – average) fitness values on colon dataset

As seen from Figure 1, the algorithm converged to a global optimum with the difference of maximum and average fitness values (max-avg) being 0 and the maximum fitness value reached to 0.99 (1 is maximum). Note that the convergence is quick due to the initial population of individuals being selected randomly from the distinction table, which consists of reducts that are already near to optimal solutions. The "reducts correspond to the minimal feature sets that are necessary and sufficient to represent a correct decision about the classification of the domain." [4]. Therefore the algorithm does not need to run through a large number of generations to find the global optimal solution. The best set of parameter values used to run GA in this study is shown in TABLE I.

TABLE I. A SET OF ADJUSTED PARAMETERS USED TO RUN GA

Parameters	Values / Methods
Population Size	100
Pc	0.7
Pm	0.03
Generation	15000
Selection	Tournament
Crossover	SinglePointCross
Mutation	SimpleBitFlip
Elitist	Single

VI. RESULTS

The dataset used for the following experiments to evaluate the proposed approach is the colon cancer dataset [15] consisting of 62 samples, 2 classes - 22 normal and 40 colon cancer samples, 2000 attributes. The dataset is sub-divided equally into 2 subsets as training and testing sets as shown in TABLE II.

TABLE II. COLON DATASET

DataSet	Type	Number of samples
Colon cancer : 62 samples (40 cancer, 22 normal), 2000 attributes	TrainingSet (31)	C1 (cancer) : 20
		C2 (normal): 11
	TestingSet (31)	C1 (cancer) : 20
		C2 (normal): 11

Using the parameter settings in Table I and the colon dataset, 10 independent runs of the algorithm were carried out. The optimal subset of 6 attributes was obtained and evaluated subsequently by 10-fold cross validation (CV) on the training dataset, and re-evaluated on the unseen test dataset using the k-NN classifier with different k values of 1, 3, 5 and 7. The results are shown in TABLE III.

TABLE III. RESULTS OF K-NN CLASSIFIER WITH K=1, 3, 5, AND 7. ALG.B (PROPOSED APPROACH): GA-RST INCORPORATING K-MEANS CLUSTERING WITH QUARTILE STATISTICS

	K=1			K=3			K=5			K=7		
Reduct Length	C1	C2	Net	C1	C2	Net	C1	C2	Net	C1	C2	Net
Banerj [4] (15)	75	63.6	71	70	36.4	58.1	75	0	48.4	90	9.1	61.3
Alg.B (6)	90	72.7	83.9	95	72.7	87.1	90	63.6	80.1	95	72.7	87.1

As shown in TABLE III, Banerjee *et al*.'s approach [4] found 15 attributes and the results they obtained using these attributes and k-NN classification on the colon dataset are 71%, 58.1%, 48.4% and 61.3 % for k=1, 3, 5 and 7, respectively. The proposed approach (i.e. Alg.B in Table III) found a set of 6 attributes and the classification results are 83.9%, 87.1%, 80.1% and 87.1% for k=1, 3, 5 and 7 respectively. It is also not possible to compare the sets of attributes as the list of 15 attributes is not listed in [4].

Furthermore, to ensure that the classification is not biased by using a particular k-NN classifier, other 22 independent classifiers from WEKA were used to test the 6 attributes obtained from Alg.B on the unseen colon test dataset. The results of the classification are shown in Table IV.

TABLE IV. RESULTS OF CLASSIFICATION FOR THE 6 SELECTED GENES WITH 22 WEKA CLASSIFIERS ON COLON TEST DATASET

	Algorithm B		
Classifier	C1	C2	Net
SMO	70	81.8	74.2
Simple Logistic	65	100	77.4
Logistic	65	100	77.4
Multilayer Perceptron	83.5*	89*	85.5*
Bayes Net	85	18.2	61.3
Naïve Bayes	80	63.6	74.2
Naïve Bayes Simple	80	63.6	74.2
Naïve Bayes Up	80	63.6	74.2
IB1	90	72.7	83.9
KStar	**90**	**90.9**	**90.3**
LWL	70	63.6	67.7
AdaBoost	80	63.6	74.2
ClassVia Regression	80	63.6	74.2
Decorate	85*	58.16*	75.5*
Multiclass Classifier	65	100	77.4
Random Committee	77.5*	55.41*	69.7*
j48	90	54.5	77.4
LMT	65	100	77.4
NBTree	80	54.5	71
Part	90	54.5	77.4
Random Forest	79*	59.97*	72.3*
Ordinal Classifier	90	54.5	77.4
Average	79.09	69.35	75.65

Note that the Multilayer Perceptron, Decorate, Random Committee and Random Forest are non-deterministic classifiers, therefore 10 independent runs with different seeds were carried out and the results with * is an average on these 10 runs.

As seen from Table IV, in the 'Net' column, the results associated with the 6 genes from Algorithm B are as follows: KStar (in bold) achieved above 90% classification accuracy; the remaining classifiers also performed well including 2 classifiers with above 80% and others are above 70%;

however, 3 classifiers Bayes Net, LWL and Random Committee, has 61.3%, 67.7%, 69.7%, respectively. The average of classification accuracy for 22 classifiers is 75.65%.

VII. DISCUSSION AND CONCLUSION

Relevance of the selected attributes to its corresponding domain is crucial. Hence this study examines the relevance of the genes found to the medical diagnosis using known biomarkers associated with colon cancer. The 6 selected genes are reported as significantly associated with cancer and other diseases. The U31248 (Human zinc finger protein (ZNF174) attribute is related to the expression of colon tissues [16]. The L08069 (Human heat shock protein, E. coli DnaJ homologue) is not only "shown to increase tumorigenicity in rat colon cancer, promotes resistance to apoptosis" [17], but also associated with tumour development in human [18]. The H49870 (UTR 2a 178915 Mad Protein (Homo sapiens)) is involved in the detection of over-expression for colon cancer disease [19]. The M18216 (Human nonspecific crossreacting antigen) is considered as a major component of Carcinoembryonic antigen involved in expression of lung cancer, tumour specimens, and tumour cell lines at mRNA levels [20], also increasing level of expression in colon cancer [21]. The M22538 (Nadh-Ubiquione Dehydrogenase 24 KD Subunit Precursor) is involved in schizophrenia, bipolar disorder, and Parkinson disease [24]. The H08393 (Collagen Alpha 2(XI) Chain) is involved in the process of degrading activity of colon cells. It is also one of 66 differently expressed genes for colon cancer data [22].

This paper proposed a hybrid algorithm which incorporates GA and Rough set theory into Feature Selection for finding the optimal subset of the most significant features. The approach utilizes the K-means clustering for partitioning of attribute values, the rough set-based approach for reducing redundant data, and genetic algorithm for searching for the optimal subset of features. The evaluation process used the same colon cancer dataset as in [4], a set of 6 attributes were obtained using the proposed algorithm, whereas 15 attributes were found in [4]. Classification using the set of 6 attributes and 23 WEKA classifiers were carried out to examine their significance to classify unseen test data. Moreover, the set of 6 genes found by the proposed approach was examined for their relevance to the known biomarkers in the colon cancer domain.

REFERENCES

[1] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," Briefings in bioinformatics - Oxford, p. 12, 2008.

[2] R. Clarke, *et al.,* "The properties of high dimensional data spaces: implications for exploring gene and protein expression data," Nature Review – Cancer. Vol. 8, p. 12, 2008.

[3] G. Kim, Y. Kim, H. Lim, and H. Kim, "An MLP-based feature subset selection for HIV-1 protease cleavage site analysis," Artificial intelligence in medicine, vol. 48, pp. 83-89, 2010.

[4] M. Banerjee, S. Mitra, and H. Banka, "Evolutionary Rough Feature Selection in Gene Expression Data," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 37, pp. 622-632, 2007.

[5] Z. Pawlak, "Rough sets," International Journal of Computer and Information Sciences," vol. 11, pp. 341-356, 1982.

[6] D. E. Goldberg, Genetic Algorithm in Search, Optimization, and Machine Learning. Alabama: Addison-Wesley, 1989.

[7] F.M. Lopes, D.C.J. Martins, and R.M.J, "Cesar, Feature selection environment for genomic applications," BMC Bioinformatics, vol. 9, p. 451, 2008.

[8] S. Wang, H. Chen, R. Li, and D. Zhang, "Gene selection with rough sets for molecular disgnosing of tumor based on support vector machines," in Proc. Int. Comput. Symp, pp. 1368-1373, 2006.

[9] S. Peng, Q. Xu, Z. B. Ling, X. Peng, and W. Du, "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines," Elsevier - Federation of European Biochemical Societies, vol. 555, p. 362, 2003.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press: Berkeley, pp. 281-297, 1967.

[11] J. Han and M. Kamber, Data mining concepts and techniques. Delhi: India: Morgan Kaufmann, 2006.

[12] R. Swiniarski and A. Skowron, "Rough set method in feature selection and recognition," Elsevier Science, vol. 24, pp. 833-849, 2003

[13] J.H. Holland, Adaptation in Natural and Artificial Systems 1975, Michigan: University of Michigan Press, Ann Arbor, 1975.

[14] M. Srinivas and L.M. Patnaik, "Adaptive Probabilities of Crossover Genetic in Mutation and Algorithms," IEEE, vol. 24, pp. 656-667, 1994.

[15] U. Alon, *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc Natl Acad Sci USA, vol. 96, pp. 6745-50, 1999.

[16] GENEBIO.ZNF174 zinc finger protein 174, http://www.nextprot.org/ db/entry/ NX_Q15697/ expression.

[17] GSAEmulator,http://www.clopinet.com/isabelle/Projects/RFE/gene.html.

[18] I. Diesinger, *et al.*, "Toward a more complete recognition of immunoreactive antignes in squamous cell lung carcinoma," Int. J. Cancer, vol. 102, pp. 372-378, 2002.

[19] N.J. Laping, "DNA encoding human MAD proteins 1999," SmithKline Beecham Corporation (Philadelphia, PA), United States (1999)

[20] T. Hasegawa, *et al.*, "Nonspecific crossreacting antigen (NCA) is a major member of the carcinoembryonic antigen (CEA)-related gene family expressed in lung cancer." Br J Cancer, 67(1), 58—65, 1993.

[21] Y. Hinoda, *et al.*, "Induction of nonspecific cross-reacting antigen mRNA by interferon-gamma and anti-fibronectin receptor antibody in colon cancer cells." Journal of Gastroenterology, 32(2), 200--205 (1997).

[22] J. Shaik, "Differentially expressed genes for Colon cancer data, " http://www. biomedcentral.com/content/supplementary/1471-2105-8-347-S6.pdf

[23] C.K. Xu, Liu, and D. Huang, "The analysis of microarray datasets using a genetic programming," IEEE, 2009.

[24] Nishioka, K., *et al.*, "Genetic variation of the mitochondrial complex I subunit NDUFV2 and Parkinson's disease," Parkinsonism & Related Disorders, vol. 16, pp. 686—687, 2010.