Cancer Data Classification using a Fuzzy Classifier Based on Bio-Inspired Algorithms

Jamshid Pirgazi
University of Zanjan
Department of Computer Engineering
Zanjan, Iran
j.pirgazi@znu.ac.ir

Ali Reza Khanteymoori
University of Zanjan
Department of Computer Engineering
Zanjan, Iran
khanteymoori@znu.ac.ir

Abstract— **Build classifier based on fuzzy rules for high-dimensional data sets, such as genetic data, are faced with great difficulties. An effective approach to this problem using feature selection techniques and dimension reduction methods. Hence, in this paper, using five different feature selection methods, size of data is reduced and the based on accuracy of the support vector machines classifier to this data a five dimensional feature vector extracted .then using frog leaping algorithm and genetic algorithm, With the aim of minimizing the number of rules and optimize the parameters of its a set of fuzzy rules for data classification are extracted. The proposed method was tested on five gene expression datasets. The experiments results show that the proposed method achieves higher accuracy than existing.**

Keywords— ***gene expression data, fuzzy classifier, frog leaping algorithm, genetic algorithm***

I. INTRODUCTION

Gene dataset have high dimensional, small sample size and is Unbalanced. High dimensional of a data set increases the search space and reduce the power of generalization and computational complexity. Feature selection methods based on the evaluation of a set of attributes; select the optimal set of features.

In the filter techniques, feature selection methods is independent of classification and learning algorithms and features searched based on the intrinsic properties of the data, such as distance, consistency, Dependence. Wrapper methods embed the model hypothesis search within the feature subset search. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated [13,19]. After the reduction features there are many ways to classify gene data. A method of classification is fuzzy classifier. Fuzzy expert system by a set of if –then rules and membership functions, shape inaccurate knowledge and approximate. The most common form of fuzzy rules is as follows:

$$R_j :\ if\ \ x_1 is\ A_{j1} and\ \ \dots\ and\ x_n is\ A_{jn} then\ class\ C_j \qquad (1)$$

Where $A_{j1}, \dots, A_{jn}$ are language Value for $x_1, \dots, x_n$ and C_j is language Value for output Variable class. In a fuzzy expert system, fuzzy rule base that is the main component of fuzzy inference system (FIS), be created by a set of rules to. In [9] a method based on genetic algorithm to select the proper rules of the larger set of rules for fuzzy classifier system has been introduced. In [2] a method based on Ant colony optimization algorithms for extracting fuzzy rules are proposed for the diagnosis of diabetes. Ant colony optimization is used to create rules for fuzzy classifier [1, 4, 14]. Combination particle swarm optimization and genetic algorithm to obtain fuzzy rules have been evaluated [5]. In this paper all required parameters of fuzzy classifier for classification of genetic cancer data calculated based on hybrid algorithm frog leaping and genetic algorithm (SFLA_GA).

II. DIMENSION REDUCTION METHODS

In this paper for dimension reduction, five commonly feature selection is used. Based on these methods, features are ranked according to the ranking of each feature in each method a collection of the best features is selected. Feature selection methods used in this paper are explained in this section.

Fisher Score: This method is base on assigin value to the samples and select samples with similar value [16]. This method evaluation features base on following formulated:

$$SC_F = (f_i) = \frac{\sum_{j=1}^{C} n_j(\mu_{i,j} - \mu_i)^2}{\sum_{j=1}^{C} n_j \sigma^2_{i,j}} \qquad (2)$$

Where μ_i is the mean of the feature f_i, $\mu_{i,j}$ and $\sigma_{i,j}$ are the mean and the variance of f_i on class j, respectively and n_j is the number of samples in the jth class.

T-Score: The T-score is the relevant measure for binary problem. If dataset have unequal sample sizes and samples have unequal variance t-score can be calculated as [16]:

$$R_t = \frac{\mu_1 - \mu_2}{\frac{\sigma^2_1}{n_1} - \frac{\sigma^2_2}{n_2}} \qquad (3)$$

Kruskal Wallis: This method is a non-parametric method. In this method samples Ranked base following formulated [10]:

$$k=(N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r_i}-\bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij}-\bar{r})^2} \qquad (4)$$

Where N is the total number of samples, n_i is number of samples in group 'i' and r_{ij} is rank of sample 'j' in the group 'i'.

Gini Index: Gini index is a criterion for measuring ability of feature in distinctive between classes [6]. Gini index of feature f can be calculated as

$$GiniIndex(f)=1-\sum_{i=1}^{C}[p(i\mid f)]^2 \qquad (5)$$

where C is number of class.

BLogReg: This method is based on eliminates the regularization parameter (λ) from Logistic Regression [7]. If we assume that the original model is as follows:

$$M = E_D + \lambda E_\alpha \qquad (6)$$

In the BLogReg this equation will be as follows:

$$\mathrm{Q} = E_D + N\log E_\alpha \qquad (7)$$

III. Shuffled Frog Leaping Algorithm

In this paper we have *sfla_p* frog. Each frog represented a fuzzy classifier. Population of frogs is partitioned into subsets called memeplexes. *sfla_m* is the number of memeplex. Therefore there are *sfla_n* frogs in each memeplex. The different memeplexes are considered as different cultures of frogs, each performing a local search. Within each memeplex, there is a sub-memeplex. In each sub-memeplex there are *sfla_q* frogs are randomly selected according to the following probability function. Sub-memeplex causes the algorithm does get seldom stuck in a local optimum.

$$P_j=\frac{2(sfla\_n+1-j)}{sfla\_n(sfla\_n+1)} \quad , j=1,2,\ldots,\mathrm{sfla\_n} \qquad (8)$$

Where P_j is the probability of select j-th frog. After a number of mimetic evolution steps, frog of memeplexes shuffling in a shuffling process. The local search and the shuffling processes continue until pre-defined convergence criteria are satisfied. In each iteration, within each sub-memeplex of memeplexs, the frogs with the best and the worst fitness's are identified as P_b and P_w, respectively. The frog with the global best fitness is identified as P_g. In each iteration only the worst fitness frog will be modified. Therefore, the position of the frog with the worst fitness is adjusted as follows:

$$S_G=\begin{cases}\min\{\mathrm{int}(rand.[P_G-P_w]),S_{\max} & for\ positive\ leap\\ \max\{\mathrm{int}(rand.[P_G-P_w]),-S_{\max} & for\ negative\ leap\end{cases} \qquad (9)$$

$$P_w^{'}=P_w+S_B \qquad (10)$$

Where S_{max} The maximum length allowed for leap. If new frog (P_w') is better than worst frog (P_w) it will be replaced by the worst frog. Otherwise the position of the worst frog is modified as follows according to the position of the frog with the global best fitness:

$$S_G=\begin{cases}\min\{\mathrm{int}(rand.[P_G-P_w]),S_{\max} & for\ positive\ leap\\ \max\{\mathrm{int}(rand.[P_G-P_w]),-S_{\max} & for\ negative\ leap\end{cases} \qquad (11)$$

$$P_w^{''}=P_w+S_B \qquad (12)$$

The same state before the new frog (P_w'') was better than the worst frog (P_w), it will replace the worst frog. If no improvement becomes possible in this case a random frog is generated which re-places the worst frog in sub-memeplex. This steps are repeated several times (*IT_{mem}*), again all frog shuffling together and again be divided into *sfla_m* memeplex. This operation will continue until the termination conditions are satisfied.

Pseudo-code of SFLA is shown in Table (I).Based on this algorithm; the worst frog can leap to the better frog. By repeating this operation, mean fitness of population increase in the mimetic evolution steps. The best solution found during the search process can be considered as the output of the algorithm [11,12].

TABLE I. Pseudo-code of SFLA

```
1. Create an initial population of SFLA_P frogs generated randomly.
2. Divide the frogs into afla_mmemplexeseachholdingsfla_n frogs.
2.2.  i= 0
 2.3. while I <$IT_{mem}$
  2.3.1. create a submemeplex for each memeplex
  2.3.2. the position of the worst frog $P_w^{'}$ for the memplex is adjusted such as (3)
  2.3.3. if (fitness($P_w^{'}$) < fitness($P_w$))
     the position of the worst frog $P_w^{''}$ for the memplex is adjusted such as (5)
  2.3.4. if (fitness($P_w^{''}$) < fitness($P_w$))
     a random frog is generated which replaces the worst frog.
  2.3.5. otherwise
     $P_w = P_w^{''}$
  2.3.6. otherwise
     $P_w = P_w^{'}$
  2.3.7. i = i + 1
 2.4. frog shuffling together
3. Check the convergence. If the convergence criteria are satisfied stop, otherwise return to the   step 3.
4.finish
```

IV. Hybrid fuzzy algorithm

The proposed algorithm in this paper consists of three basic steps.

1- First step: using feature selection methods described in section 2, five feature vectors are created for training five support vector machine classifiers.

2- Second step: based on five feature vectors from before step, support vector machine classifiers trained. After trained classifiers five models obtained. In this step for each data obtained a vector with five dimensions.

3. Third stage: using feature vector calculated from before step and hybrid algorithm frog leaping and genetic algorithm to train the new fuzzy classifier. Thus a fuzzy inference system to find a class of test data obtained.

In the third stage, SFLA_GA algorithm is used to obtain the membership rules and membership functions based on training data. Rules are in the form of relation (1) and membership functions also include a triangular function, trapezoid, Z-shaped, S-shaped and Gaussian. In this paper, the Mmdany fuzzy model is used also used multiplication for AND operator, sum for fuzzification and COA for defuzzification. The population of frogs that each one is fuzzy inference system randomly generated. Then the set of fuzzy rules in the form of relation (1) are randomly generated for each frog. Number of rules for each frog calculate base on follows equation.

$$R_j = \frac{2(R_{\max} + 1 - rand)}{R_{\max}(R_{\max} + 1)} \qquad (13)$$

Where R_{max} is the maximum number of rules allowed. After the creation of the population, population should be evaluated. For this purpose we used of accuracy rate. The memeplexs and sub-memeplex for each memeplex will be created. The worse frog in each sub-memeplex based on the equation in section 3 do leaping.To improve worse frog base on better frog, first, the number of rules that should be added or removed from the rule base worse frog are calculated by equation (13).

$$S_B = \begin{cases} \min\{\text{int}(rand.[P_b - P_w]), S_{\max}\} \text{if } SP_b > SP_w \\ \max\{\text{int}(rand.[P_b - P_w]), -S_{\max}\} \text{else} \end{cases} \qquad (14)$$

Where SP_b and SP_w are respectively number rules of better frog and worse frog and S_{max} the maximum number allowed for changes in the rule base. Then of S_B is positive, S_B rules will be add to rule base of worse frog and if of S_B is negative, S_B rules will be remove to rule base of worse frog. Then from each of the worse and better frogs a rule randomly is selected. In order to create new rules, crossover and mutation operator of genetic algorithm is used. To use these operators problem be displayed in the form of Chromosomes. For each rule, we can use the six genes in the chromosomes (five genes for the input variables and one gene to output variable).

For display fuzzy functions required chromosomes with 17 gene, that each gene is a vector of length 5. (For five input variable three language term and for output variable two language term) After creating the chromosome crossover and mutation operator is applied. With this operator wores frog leaping towards better frog. These steps are executed repeatedly. After running the algorithm several times, frog with the best accuracy rates is chosen as solution. This fuzzy system is used to classify test data.

V. Results

A. Data set

We chose five common microarray data sets to evaluate the accuracy of our proposed method. Summary of the data sets are shown in Table II.

TABLE II. MICROARRAY DATA SETS USED IN THE EXPERIMENTS

Data Set	#Samples	#Gene	#classes	#class1	#class2
Leukemia	72	7129	2	47	25
Colon	62	2000	2	40	22
Prostate	136	12600	2	77	59
DLBCL	77	11226	2	58	19
CNS	60	7129	2	39	21

The data sets include leukemia dataset [8], colon dataset [3], prostate tumor dataset [17], Diffuse Large B-Cell Lymphoma dataset (DLBCL) [18] and Central Nervous System dataset (CNS) [15]. Leukemia dataset contains expression levels of 7129 genes taken over 72 samples which contain 47 Acute Lymphoblastic Leukemia (ALL) samples and 25 Acute Myelogenous Leukemia (AML) samples. The colon dataset contains expression levels of 2000genes taken in 62 samples. For each sample it is indicated whether it came from a colon cancer or not. Prostate dataset contains expression levels of 12600 genes taken over 136 samples. For each sample it is indicated whether it came from a tumor or not. DLBCL dataset contains expression levels of 11226 genes taken over 77 samples which contain 58 diffuse large b-cell lymphoma samples and 19 Follicular lymphoma samples. The CNS dataset contains expression levels of 7129 genes taken over 60 samples.

B. Evaluation

To evaluate the proposed method do following Experiment and the Experiment.

Hypothesis1. Do high dimensional of data Causes over fit?

Experiment1. Apply gene data without feature selection to SVM and fuzzy classifier.

The results of the experiments are shown in Table1. As the results show that accuracy of SVM and fuzzy classifier are not acceptable. One reason for this bad accuracy is over fit. Because the sample size is small and number of features is large

TABLE III. ACCURACY RATE OF FUZZY AND SVM CLASSIFIER ON A DATASET WITHOUT FEATURE SELECTION

Data Set	SVM	Fuzzy lassifierC	Data Set	SVM	Fuzzy Cl assifier
Prostat	73	69.01	CNS	84.09	79.47
Colon	80.70	73.21	DLBCL	82.21	`80.70
Leukemia	77.42	70.23	--------	--------	---------

Hypothesis2. Do feature select affect to accuracy rate of classifier?

Experiment2. For this purpose, the gene data was then reduced by choosing good features and then Applying this data to SVM and fuzzy and combination fuzzy classifier. The results of the experiments are shown in Table IV. By applying feature selection accuracy rate of fuzzy and SVM classifier on all data sets except Leukemia dataset increased. Also the combination fuzzy classifier accuracy rate is high on all data sets and is better than of fuzzy and SVM classifier.

TABLE IV. ACCURACY RATE OF FUZZY, SVM AND COMBINATION FUZZY CLASSIFIER ON A DATASET WITH FEATURE SELECTION

Data Set	SVM	Fuzzy Classifier	uzzyF Combination Classifier
Prostat	80.01	78	86.72
Colon	90.32	87.89	91.93
Leukemia	76.86	75.33	82.89
CNS	76.46	72.23	84.21
DLBCL	95.71	92.23	95.71

Hypothesis3. Is the combination fuzzy classifier robust to classify gene data?

Experiment3. A parameter affected in the accuracy rate of combination fuzzy classifier and other classifier is number of features. For this purpose, the effect number of features evaluated in the performance of combination fuzzy classifier. In Figure 1 are shown the accuracy rate of the proposed classifier on five datasets with different features. As can be seen, for most data sets to increase the number of features increases the accuracy rate. However, this increase is roughly to number of features is 40 and from 40 to 100 features the accuracy rate is not increased substantially but in some data sets will also be less. Also with more than 100 feature accuracy rate is downtrend

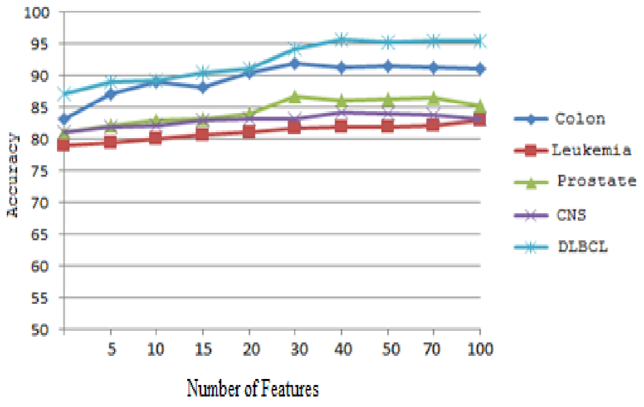

Fig. 1. Evaluation number of features on accuracy rate combination fuzzy classifier

Hypothesis4. What is the process of convergence of the algorithm?

Experiment4. For this purpose, the accuracy rate of the best frog and mean accuracy rate of frogs over 30iteration of the algorithm is show in Figure 2. It is clear mean and best (max) accuracy of frogs is increasing and the systems are learning.

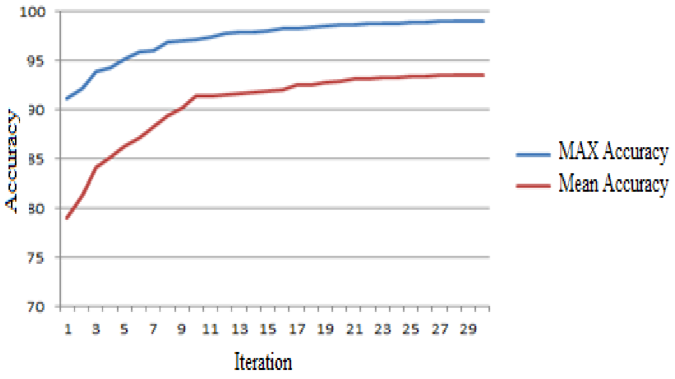

Fig. 2. Mean and maximum accuracy of populations of frogs in the past 30 step training

VI. CONCLUSION

To use fuzzy classifier, requires creating rules, membership functions and parameters of the membership functions. In this paper, these parameters are calculated based on the genetic and Leap Frog algorithms. The results of implementation show that the proposed method does not over fit on the data and accuracy rate is better than of fuzzy and SVM classifier. The effect number of features evaluated in the performance of combination fuzzy classifier and was found that less than 50 feature an acceptable accuracy rate reached. This reduces the risk of over fitting and reduces the time runs out. Since the algorithm is based on the population and the population is divided into different groups do not get stuck in a local optimum and forward to global optimum and finally converges.

REFERENCES

[1] A. A. Freitas, H. S. Lopes,R. S. Parpinelli,“Data mining with an ant colony optimization algorithm," *IEEE Transactions on Evolutionary Computation,* vol. 6, pp. 321–332, 2002.

[2] A.Khotanzad and E. Zhou,”Fuzzy classifier design using genetic algorithms”, *Pattern Recognition,* pp. 3401–3414, 2007.

[3] Alon.U,Barkai.N,Gishdagger.k,Levine.A.J,Mackdagger.D,Notterman.D. A and Ybarradagger.S, “Broad Patterns of GeneExpression Revealed by Clustering Analysis of Tumor and NormalColon Tissues Probed by Oligonucleotide Arrays,” *Proc. Nat’l Academy of Sciences USA,* vol. 96, no. 12, pp. 6745-6750. June 1999.

[4] B. Liu, B. McKay and H. A. Abbass, “Classification rule discovery with ant colony optimization” *presented at the IEEE/WIC int. conf. on intell.agent techno*, 2003.

[5] B. Baesens ,D. Martens, D. Backe , J. Vanthienen, M. Snoeck, R. Haesenand, “Classification with ant colony optimization”, *IEEE Transactions on Evolutionary Computation,* vol. 11, pp. 651–656, 2007.

[6] C. Gini,“Variabilite e mutabilita”, *Memorie di metodologia statistica*, 1912.

[7] G.C.Cawley and N.L.C.Talbot, “Gene selection in cancer classification using sparse logistic regression with bayesian regularization”, *BIOINFORM ATICS*, 22:2348-2355, 2006.

[8] Golub.T.R et al, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring”, *Science*, vol. 286, no. 5439, pp. 531-537, 1999.

[9] H.Shibuchi and T.Yamamoto,“Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining” , *Fuzzy Sets and Systems,* vol. 141, pp. 59–88, 2004.

[10] L.J.Wei,”Asymptotic conservativeness and e_ciency of kruskal-wallis test for k dependent samples”, *Journal of the American Statistical Association*, 76(376):1006 -1009, December 1981.

[11] katayoun madani, M.T.Vakil Baghmisheh,”A discrete Shuffled frog optimization algorithm”, *artifintell* rew(2011)36:267-284.

[12] Mohammad Rasoul Narimani*,”*A New Modified Shuffle Frog Leaping Algorithm for Non-Smooth Economic Dispatch*”, World Applied Sciences Journal* 12 (6): 803-814, 2011 ISSN 1818-4952.

[13] Minaalibeigi, S. H, “Dbfs: An Effective Density Based Feature Selection Scheme For Small Sample Size And High Dimensional Imbalanced Data Sets”, *Data& Knowledge Engineering*, 2013.

[14] M.F.Ganji and M.S.Abadeh, “A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis”, *Expert Systems with Applications* vol. 38, 2011.

[15] Pomeroy.S.L et a.l, “Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression”, *Nature*, vol. 415, pp. 265-271.2002.

[16] R.O. Duda, et al, ”Pattern Classification”, *John Wiley & Sons*, New York, edition, 2001.

[17] Singh ,D et al, “Gene Expression Correlates of Clinical Prostate Cancer Behavior”, *Cancer Cell*, vol. 1, no. 2, pp. 203-209.2001.

[18] Shipp.M.A et al,“Diffuse Large B-Cell Lymphoma OutcomePrediction by Gene-Expression Profiling and Supervised Machine Learnin”, *Nature Medicine*, vol. 8, no.1, pp. 68-74.Jan. 2002.

[19] Waelawada.T.M, “A Review Of The Stability Of Feature Selection Techniques For Bioinformatics Data”, *Las Vegas, Nevada, Usa IEEE* (P. 8), 2012.