Naive Bayes Based Firefly Algorithm for Gene Selection and Classification of Microarray Data

Naive Bayes Tabanlı Firefly Algoritması ile Gen Seçimi ve Mikrodizi Verilerinin Sınflandırılması

Taliha Sunduri, Prof. Dr. Nizamettin Aydın
Computer Engineering Department
Yıldız Technical University
İstanbul, Turkey
talihasunduri@gmail.com
nizamettin@ce.yildiz.edu.tr

***Abstract*—Selection of relevant genes for classification is an important task in most gene expression studies. In this study a novel approach based on Random Forest Ranking, Firefly Algorithm and Naive Bayes classifier is proposed for gene selection and classification of microarray data. In proposed approach, Random Forest and Firefly Algorithm are used to perform gene selection to remove irrelevant and redundant genes. To evaluate the effectiveness of our proposed method, the experiments are done on three benchmark microarrays and one real microarray which is obtained from NCBI.**

Random Forest is an ensemble classifier that uses many decision trees. Accuracy and variable importance information is provided with the results. It gives very successful results of finding the most important predictors.

Naive Bayes is a classification technique based on Bayes Theorem with an assumption of independence among predictors. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Firefly algorithm is one of the recent swarm intelligence methods. It is inspired by the flashing lights of fireflies in nature. It can be applied for solving the hardest optimization problems.

Keywords — firefly algorithm; binary firefly algorithm; random forest ranking; naive bayes classifier; gene selection; classification of microarrays.

***Özetçe*— Sınıflandırma işlemi için en iyi bilgiyi veren genlerin seçimi birçok gen ifadesi çalışmalarında büyük önem taşımaktadır. Bu çalışmada Random Forest tabanlı sıralama, Firefly algoritması ve Naive Bayes sınıflandırıcısı kullanılarak gen seçimi ve mikrodizi verilerinin sınıflandırılması işlemi gerçekleştirilmiştir. Firefly algoritması gen seçimi aşamasında fazla ve katkısı olmayan genleri çıkarmak için kullanılmıştır. Yöntemin etkinliğini değerlendirebilmek amacıyla deneyler üç mikrodizi benchmark'ında ve NCBI'dan alınan gerçek bir mikrodizi üzerinde yapılmıştır.**

Random forest birçok karar ağacı kullanarak çalışan bir toplu sınıflandırıcıdır. Random forest sınıflandırıcısı kullanılarak sınıflandırma doğruluğu ve değişkenlerin önem değeri bilgileri elde edilebilir. En önemli tahmin değişkenlerinin bulunmasında başarılı sonuçlar vermektedir.

Naive Bayes algoritması Bayes teoremini ve bağımsızlık önermesini kullanan bir sınıflandırıcıdır. Naive Bayes modelinin kullanımı kolaydır ve özellikle büyük veri setleri için kullanışlıdır. Basit olmasının yanında daha karmaşık sınıflandırma yöntemlerinden bile daha büyük başarı gösterebilmektedir.

Firefly algoritması güncel sürü zekası algoritmalarından biridir. Doğadaki ateş böceklerinin yaydığı ışıktan esinlenilerek geliştirilmiştir. En karmaşık optimizasyon problemlerinin çözümünde bile kullanılabilir.

Anahtar Kelimeler — firefly algoritması; binary firefly algoritması; random forest tabanlı sıralama; naive bayes sınıflandırıcısı; gen seçimi; mikrodizilerin sınıflandırılması

- ***Introduction:*** The true diagnosis of the diseases is crucial for the successful treatments. DNA microarray technology provides great opportunities for the studies in this area. Microarrays are the datasets that allow you to process thousands of genes at the same time. Along the many studies done with microarrays, gene expression microarrays are attracting attention in data mining,

machine learning and statistics. The main purpose of the classification of gene expression microarrays is to construct a classifier from gene expression microarray data and classify the new data with this constructed classifier. In practice, gene expression microarrays are widely used in the classification and prediction of clinical cancer outcomes in cancer research. It is also applied in cancer diagnosis.

Classification is an important process in the field of machine learning and data mining, where each instance of the data is classified into different groups. Microarrays contain high-dimensional and noisy data. Often, it may contain genes that are both redundant and irrelevant. The presence of some genes reduces performance and the quality of the entire data set. Therefore, it is very important to use gene selection as a pre-processing tool in solving the classification problems.

Existing gene selection methods can be classified as filter methods and wrapper methods. In the wrapper methods, the fitness value of the selected subset of genes is obtained through a classification algorithm. In filter methods, gene selection is independent of any classification algorithm. Appropriate features are derived from the statistical properties of the data. In this study a filter and a wrapper method is used together for gene selection and classification of microarray data.

Despite many studies on this area, there are many problems due to the high-dimensional data. The most common problems are sticking to the local minimum/local maximum and the high computational time of the methods. Evolution algorithms are powerful general search algorithms that are successfully used in many areas. The Firefly algorithm (FA) can be considered to be a newer swarm intelligence based evolutionary method than other methods [1]. Firefly algorithm is an appropriate algorithm for gene selection. It gives the best or almost the best results. In this study a novel approach based on Random forest ranking, Firefly algorithm and Naïve Bayes classifier is used for gene selection and classification of microarray data. Gene selection for cancer identification based on this approach has not considered yet. Gene selection is showed in Figure 1.

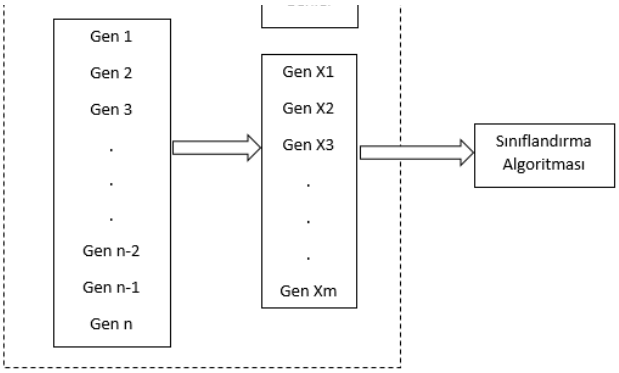

Figure 1. Gene selection

- ***Literature Review:*** Bio-inspired method called Firefly Algorithm hybridized with NB classifier has been used for solving web page classification [2]. There has been a very rapid growth in the amount of data due to internet usage and improvements in communication technologies. Therefore, processing this high amount of data online is a difficult process. To solve this problem, many new methods used by search engines have been developed. This method extracts features from web pages as first step. The FA is applied after once a value has been assigned for each combination of features. Once Naive Bayes (NB) classifier is used at classification stage, results are obtained from the error matrix. The method performed better than the existing methods on this subject. Fetal risk anticipation is another study that used FA and Naïve Bayes methods [3]. The opposition-based FA is used to get relevant features and to increase the accuracy of the SVM classifier. It is seen that the opposition-based firefly algorithm is better than the standard firefly algorithm. It is possible for obstetricians to obtain more accurate results through CTG data by this method. Also there is an email spam classification study that used Firefly algorithm with Naïve Bayes classifier [4]. Spam e-mails are a problem for almost every computer user. They take time of the user and reduce the productivity. In this study, spam e-mails were classified using FA algorithm and NB classifier. Detection of spam e-mails has been tested. The method was compared with existing studies such as PSO, NN. The classification of spam e-mails with NB based FA gave better results than other methods.
- ***System Design:*** Because of the high size of the microarrays, the random forest ranking method was used as a pre-processing tool before the Naïve Bayes based FA was applied to shorten the processing time. As a result of the training of this

classifier, the importance value of the genes is also obtained. This method is mentioned in Breiman's original paper [5]. The first step of determining the importance of genes is to fit the data to the randomly populated random forest or decision trees. The rest of the training data set, also called out of-bag (OOB) data is used to testing. Errors are saved and averaged. In order to calculate the importance value of the ith gene, the values of the ith gene are changed on the training set. The OOB error value is again calculated over the permuted data set. The importance value is obtained by averaging the difference of the OOB error value before and after the permutation over all trees. The standard deviation of these values is taken so the obtained importance values can be normalized. Thus, the genes are ranked according to their importance value and the gene subset is selected.

NB classifier with "stratified" 10-fold cross-validation is used to evaluate the gene subset while classifier is learning. NB classifier is used in this study because it is easy to use and suitable for large data sets. The error rate is used in objective function.

FA is one of the recent swarm intelligence methods developed by Yang [1] in 2008. It is inspired by the flashing lights of fireflies in nature. These lights have two known fundamental functions: to attract mating partners and to warn predators. The light intensity I decreases as the distance r increases according to $I \propto 1/r^2$. There are three idealized rules for the original FA:

- All fireflies are unisex, so one firefly will be attracted to other fireflies regardless of their sex.
- Attractiveness is proportional to a firefly's brightness. Thus for any two flashing fireflies, the less bright one will move toward the brighter one. If there is no brighter one than a particular firefly, it will move randomly.
- The brightness of a firefly is affected or determined by the landscape of the objective function.

Position of a firefly represents a candidate solution. In gene selection the binary representation is more appropriate to use. Thus selected/unselected genes can be easily represented. So in this study binary firefly algorithm (BFA) is used instead of standard firefly algorithm. The main problem in the BFA is to get positions as binary values after the fireflies movement. The movement equation of fireflies gives real values. So in order to convert the real values to binary values tanh function is applied. After the *tanh* function is applied, the genes are selected for values larger than a random value. The pseudo code of FA is showed in Figure 2.

Define light absorption coefficient γ.
while (t <MaxGeneration),
for $i = 1 : n$ all n fireflies
  for $j = 1 : n$ all n fireflies (inner loop)
  if ($I_i < I_j$)
    Move firefly i towards j.
    end if
    Vary attractiveness with distance r via $\exp[-\gamma r^2]$.
    Evaluate new solutions and update light intensity.
  end for j
end for i
Rank the fireflies and find the current global best g_*.
end while
Post-process results and visualization.

Figure 2. Pseudo code of firefly algorithm

The distance measure is crucial in firefly algorithm. The normalized Euclidean distance is used as distance measure.

Generally error rate of the classifier is used as objective function. However the aim is to find the optimal value with less number of genes. So in this study a different objective function is used which is also considering the number of genes in evaluation. This objective function is shown in Equation 1.

$$f(x) = (1-\omega) \times \frac{\sum_i x_i}{N} + \omega \times \frac{E}{E_N} \qquad (1)$$

where *f(x)* is the fitness function given a vector *x* sized *N* with 0/1 elements representing unselected/selected features, *N* is the total number of features in the data set, *E* is the classifier error rate and ω is a constant controlling the importance of classification performance to the number of features selected. E_N is the classifier error rate with using all predictors in the dataset.

- ***Experimental Study:*** To evaluate the effectiveness of this project, Central Nervous System (CNS), SRBCT, Colon Tumor [7] datasets and one real dataset GSE25136 from NCBI [8] are used as datasets. The number of selected genes by Random Forest Ranking is 100 for Colon Tumor and SRBCT, 500 for CNS and GSE25136. The optimal selected genes and their best and average accuracies are evaluated by Naive Bayes with 10-fold cross validation for 100 times. 100 iterations are used in FA for Colon Tumor and SRBCT dataset. 200

iterations are used in FA for Central Nervous System and GSE25136 dataset.

The differents parameters and equations are tested for higher accuracies. The parameters that give most successful results are α=0.7, beta0=2 and γ=1. The normalized Euclidean distance function gave better results than standard Euclidean distance. Also the objective function gave successful results. After the 100 iterations, as iterations increases the program gives almost same accuracies with less number of genes. So this is the success of the objective function that is used in this project.

The selected gene subsets and their accuracies are shown in Table 1. Also the comparison of this method with different classifiers is shown in Table 2.

Datasets	Number of Selected Genes	Best Accuracy	Average Accuracy
Colon Tumor	5	91.94	91.65
	6	91.94	91.74
	7	91.94	91.87
	9	98.39	96.74
SRBCT	7	100	100
	7	100	98.99
	9	100	99.64
	10	100	99.95
CNS	21	96.67	93.12
	18	96.67	94.48
	12	100	99.40
	19	96.67	95.3
GSE25136	17	98.73	96.15
	19	100	97.19
	19	100	98.86
	18	98.73	97.90

Table 1. Optimal selected gene subsets and their best and average accuracies

Datasets	NB	AdaBoostM1	IBK	Random Forest	RFR-FA-NB
Colon Tumor	%56.45	%77.02	%76.21	%81.46	%93
CNS	%68.08	%59.17	%59.58	%63.33	%95.58
SRBCT	%98.49	%56.33	%82.23	%98.80	%99.65
GSE25136	%69.94	%54.11	%58.54	%65.19	%97.53

Table 2. Comparison of method with different classifiers

- ***Result and Discussion:*** The highest accuracy for Colon Tumor dataset is %98.39 and the lowest is %91.65. The highest accuracy for SRBCT dataset is %100 and the lowest is %98.99. The highest accuracy for Central Nervous System dataset is %99.40 and the lowest is %93.12. The highest accuracy for GSE25136 dataset is %100 and the lowest is %96.15. In general this method give high accuracies for all dataset that is used in this project.

  In comparison with different classifiers for all datasets this method ourperformed the all classsifiers with less selected genes.

- ***Conclusion:*** A novel approach based on Random Forest Ranking, Firefly Algorithm and Naive Bayes classifier is implemented to obtain higher classification accuracies with less genes for classification of the microarrays. Naive Bayes with 10-fold cross validation is used as classifier. The performance of the method is evaluated on three benchmarks and one real dataset from NCBI. This method gives high accuracies on all datasets that are used in this project.

REFERENCES

[1] X.-S. Yang, "Firefly algorithms for multimodal optimization," in *International Symposium on Stochastic Algorithms*, Springer, 2009, pp. 169–178.

[2] K. Bhatt, A. Singh, and D. Singh, "An Improved Optimized Web Page Classification using Firefly Algorithm with NB Classifier (WPCNB)", *International Journal of Computer Applications,* vol. 146, pp. 15-21, 2016.

[3] V. Subha and D. Murugan, "Opposition Based Firefly Algorithm Optimized Feature Subset Selection Approach for Fetal Risk Anticipation," *Machine Learning and Applications: An International Journal (MLAIJ)*, vol. 3, pp. 55-64, 2016.

[4] D. K. Renuka, P. Visalakshi, "Blending firefly and bayes classifier for email spam classification", *International Review on Computers and Software*, vol. 8, pp. 118-130, 2014.

[5] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] S. Mirjalili and A. Lewis, "S-shaped versus v-shaped transfer function for binary particle swarm optimization," *Swarm and EvolutionaryComputation*, vol. 9,pp. 1–14, 2013.

[7] Y. S. O. Zexuan Zhu and M. Dash. (2007). Microarray datasets in weka arff format, [Online]. Available: http://csse.szu.edu.cn/staff/zhuzx/ Datasets.html (visited on 11/25/2010).

[8] N. C. for Biotechnology Information. (2010). Geo dataset browser, [Online].Available:https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc= GDS4109 (visited on 12/25/2016).