International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

# A Comparative Study of Statistical and Artificial Intelligence based Classification Algorithms on Central Nervous System Cancer Microarray Gene Expression Data

**Mustafa Turan Arslan*[1], Adem Kalinli[2]**

*Accepted 3rd September 2016*

***Abstract:*** A variety of methods are used in order to classify cancer gene expression profiles based on microarray data. Especially, statistical methods such as Support Vector Machines (SVM), Decision Trees (DT) and Bayes are widely preferred to classify on microarray cancer data. However, the statistical methods can often be inadequate to solve problems which are based on particularly large-scale data such as DNA microarray data. Therefore, artificial intelligence-based methods have been used to classify on microarray data lately. We are interested in classifying microarray cancer gene expression by using both artificial intelligence based methods and statistical methods. In this study, Multi-Layer Perceptron (MLP), Radial basis Function Network (RBFNetwork) and Ant Colony Optimization Algorithm (ACO) have been used including statistical methods. The performances of these classification methods have been tested with validation methods such as v-fold validation. To reduce dimension of DNA microarray gene expression has been used Correlation-based Feature Selection (CFS) technique. According to the results obtained from experimental study, artificial intelligence-based classification methods exhibit better results than the statistical methods.

***Keywords:*** *DNA microarray, Feature selection, Classification, Statistical methods, Artificial intelligence methods, Gene expression data*.

## 1. Introduction

Gene expression analysis of thousands of genes can be performed with a technique called microchip thanks to innovations in technology and research [1]. DNA microarray is high intensity gene array and it makes possible to examine thousands of gene expression profile [2]. Microarray technology provides abundance of knowledge on expression levels of thousands genes that has been used for diagnostic and prognostic purposes for various types of diseases. Microarray technology is an invention that allows for detection of too many genes. This technique is also used in many fields including medicine primarily [3]. The activity of genes in patient and healthy cells that get same tissues can compare through DNA microarray technology. This technique can help finding genes associated with a disease. For example, it can be used to identify a gene associated with a disease thanks to compare gene expression level of healthy and diseased cells [3].

The interest in working with the rapid advancement of DNA microarray technology is increasing day by day. These studies, is a comprehensive technology used in molecular biology and medicine. DNA microarray data analysis plays an important role the identification of genes associated with diseases such as cancer. It can be calculated in high probability that any individual is ill or healthy by identifying disease-related genes. Therefore, high performance classification methods are very important for analyzing large-scale data such as microarray gene expression data.

Statistical methods such as Support Vector Machines, Decision Trees and Bayesian Network are the most frequently used methods in microarray classification. However, these methods can often be inadequate to solve problems which are based on especially large-scale data. Therefore, it is important to develop methods that can be effective to solve such problems.

In the last few decades, artificial intelligence techniques are methods which are commonly used and preferred to solve difficult problems. That’s why, we are interested in classifying central nervous system microarray cancer gene expression by using artificial intelligence based classification methods including statistical methods in this study.

In the experimental analysis, the results of classification obtained by artificial intelligence methods are compared with the results of classification obtained by statistical methods. The rest of this paper is organized as follow: We describe the relevant methods in our comparison study part in Section 2. In Section 3, we introduced our experimental dataset called central nervous system cancer microarray gene expression dataset. We report the results of our experiments which are followed by statistical analysis and discussions in Section 4. Finally, we conclude the paper with an outlook to our future work about the control parameter optimization of classification algorithms.

## 2. Methods

A classification function consists of two parts. First part is about selecting important features by using “feature selection” methods. Second part is about classifying data thanks to classification

[1] *Kirikhan Vocational School, Mustafa Kemal University, Hatay, Turkey*
[2] *Department of Computer Engineering, Erciyes University, Kayseri, Turkey*
** Corresponding Author: Email: mtarslan@mku.edu.tr*

*Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.*

methods. System component used in this study are described below.

### 2.1. Feature Selection

The number of features is usually very high in gene expression dataset. Therefore, we need to reduce dimension on dataset to make better classification. Feature selection is very important process to make classification with high accuracy on microarray cancer datasets. We examined various feature selection methods and preferred Correlation based Feature Selection (CFS) because it made a successful choice among all features.

#### 2.1.1. Correlation based Feature Selection (CFS)

CFS is a simple filter algorithm that ranks feature subsets and discovers the merit of feature or subset of features according to a correlation based function. According to this approach, subsets which has the best attributes consist of attributes which have a high correlation with the corresponding class label and have low correlation with each other. The rest of features should be ignored. CFS feature subset evaluation function is shown as follows:

$$G_s = \frac{k\overline{r_{ci}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \qquad (1)$$

where $G_S$ is the heuristic merit of a feature subset S containing k features, $r_{ci}$ is the mean feature–class correlation, and $\overline{r}_{ii}$, is the average feature-feature intercorrelation. This equation is, in fact, Pearson's correlation, where all variables have been standardized [4].

### 2.2. Classification

In this study, we used six algorithms to classify central nervous system cancer gene expression data. In six algorithms, 3 of them belong to statistical algorithms and the rest of algorithms belong to artificial intelligence-based algorithms. These algorithms are given below.

#### 2.2.1. Support Vector Machines (SVM)

SVM is a statistical algorithm found by V. Vapnik [5] in the late 1960s. It is a method which is used particularly in classification microarray gene expression levels. SVM is a supervised classification algorithm based on statistical learning theory. SVM initially had designed for two-class classification of linear data and then was generalized to classify non-linear and multi-class data. The working principle of SVM is based on predicting the optimal decision function that can separate two classes from each other. In other words, the most appropriate way to define SVM is the hyper-plane, which can separate two classes each other [6]. An infinite number of non-optimal hyper-plane can be drawn to split the two sets from each other. However, SVM try to find optimal hyper-plane that provides the maximum margin to separate the two sets from each other.

If samples can't separate as linear, then the samples are moved to higher dimension with the help of kernel functions with different characteristics.

Radial basis kernel function is frequently used in classification applications. SVM is an important classification method because it is fast and especially perform good results on large-scale data like microarray gene expression.

#### 2.2.2. One Rule (OneR)

One-R or "One Rule (a Rule)" is a simple algorithm proposed by Holt R.C [7]. This algorithm produces a rule in training data for each feature and then rule which has minimum error rate according to One-R is selected.

#### 2.2.3. J48 Decision Tree

J48 Decision Tree is the Weka implementation of the C4.5 algorithm, based on the ID3 algorithm. The main idea of this method is to generate decision trees by using the information entropy. The method divides dataset by calculating the information gain of each attribute and attribute which provides the most benefit is used to make a decision [8].

#### 2.2.4. Multi-Layer Perceptron (MLP)

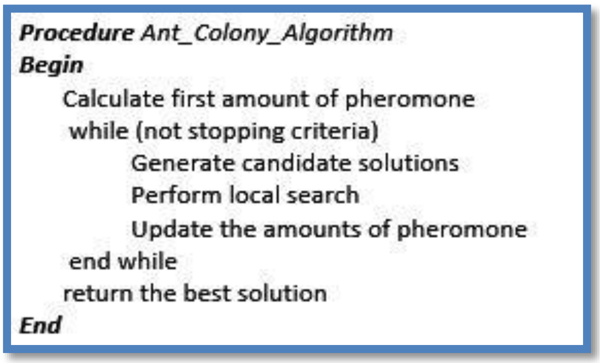
This model on which Rumelhart and his friends worked together is called error propagation model or back-propagation model (backpropagation of network) [9]. There are one or more hidden layer excluding an input and output layer in this model. Neurons in layers is associated with other layers. Information flow direction is forward and there is no feedback on the network in MLP. Therefore, it is known as feed forward neural network. Data is not processed in neurons in the input layer. The number of neurons in this layer depends on the number of dimensions of the problem to be applied to the network. The number of hidden layer and neuron are randomly determined. The number of neurons in the output layer depends on the type of problem [10].

#### 2.2.5. Radial basis Function Network (RBFNetwork)

RBFNetwork was revealed in 1988 by inspiring the behaviour of biological neurons [11]. Training of this model can be compared to curve fitting approach in multidimensional space [12]. It is used radial basis activation functions the transition from the input layer to the intermediate layer unlike other neural network structure in RBFNetwork and a non-linear clustering (cluster) analysis is performed. There are three layers called input, hidden and output on RBFNetwork like conventional ANNs structure. The structure between the intermediate and output layer is also same in other types of Artificial Neural Networks, and training is performed among neurons which are intermediate and output layer.

#### 2.2.6. cAnt-Miner

Ant Colony Optimization algorithm (ACO) is a proposed algorithm inspired by the behaviour of real ants in nature [13].This algorithm was proposed by utilizing the ability to find the shortest path between nest and food source of the ants [14]. The general steps of this algorithm is shown Figure 1.

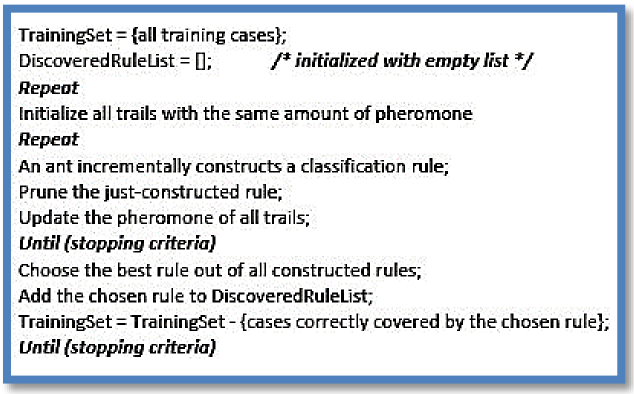


```
Procedure Ant_Colony_Algorithm
Begin
    Calculate first amount of pheromone
    while (not stopping criteria)
        Generate candidate solutions
        Perform local search
        Update the amounts of pheromone
    end while
    return the best solution
End
```

**Figure 1.** Ant colony optimization algorithm

The use of this algorithm for solution of classification problems has been just used and Ant-Miner has been proposed by

Parpinelli and his friends [15] to classify data recently. After that, Ant-Miner was developed by Liu and his friends [16] by the same logic, but different heuristic and pheromone update strategy. cant-Miner algorithm was also developed for data which has continuous values. High-level pseudocode of cAnt-Miner algorithm is shown Figure 2.



```
TrainingSet = {all training cases};
DiscoveredRuleList = [];          /* initialized with empty list */
Repeat
    Initialize all trails with the same amount of pheromone
    Repeat
        An ant incrementally constructs a classification rule;
        Prune the just-constructed rule;
        Update the pheromone of all trails;
    Until (stopping criteria)
    Choose the best rule out of all constructed rules;
    Add the chosen rule to DiscoveredRuleList;
    TrainingSet = TrainingSet - {cases correctly covered by the chosen rule};
Until (stopping criteria)
```

**Figure 2.** High level pseudocode of cAnt - Miner algorithm

We used cAnt-Miner algorithm because central nervous system microarray gene expression dataset we use in this study has continuous values.

## 3. Data Description

Central Nervous System cancer dataset, provided by Pomeroy [17], contains the expression levels of 7129 genes. Each sample was obtained from brain tissues and was analysed using Affymetrix microarrays. This dataset contains two subtypes of cancer, namely classic medulloblastomas (CMD) and desmoplastic medulloblastomas (DMD). After data pre-processing, 857 genes remain. The source of the 857 gene expression measurements is publicly available at [18]. Central Nervous System cancer dataset is available at Schliep lab bioinformatics Repository of Rutgers University contains 857 genes with one class attribute. The dataset includes numeric attributes. The class shows two subtypes of cancer named CMD and DMD. The dataset contains 34 samples belonging to two different target class. In the 34 samples, 25 of them belong to CMD class and 9 samples belong to DMD class.

**Table 1.** Description of Central Nervous System microarray cancer data

| Dataset | Comparison | Class | Gene | Sample |
|---|---|---|---|---|
| Central Nervous System Cancer | Tumour Subtypes (CMD,DMD) | 2 | 857 | 34 |

Table 1 shows the summary of the characteristics of the microarray cancer dataset.

## 4. Results and Discussion

In this study, we compared the efficiency of the classification methods including; SVM, OneR, J48 Decision Tree, MLP, RBFNetwork and cAnt-Miner methods for the prediction of cancer risks. We used improved and modified Weka [19] software for applying classification on the experimental dataset. Classification accuracy was used as performance measure. The percentage accuracy is defined as the ratio of correctly classified samples to the total number of samples. The percentage accuracy is given by (Equation.2).

Percentage accuracy (%) =

$$\frac{Correctly\ Classified\ Samples}{Total\ Number\ of\ Samples} * 100 \qquad (2)$$

We transformed data into the format arff for Weka. At first, feature selection method was used to find relevant features in the central nervous cancer data and then, classification algorithms were applied to the selected features to evaluate the algorithms. Thirty features (genes) were selected by the feature selection method. Summary of the experimental work is presented Figure 3.

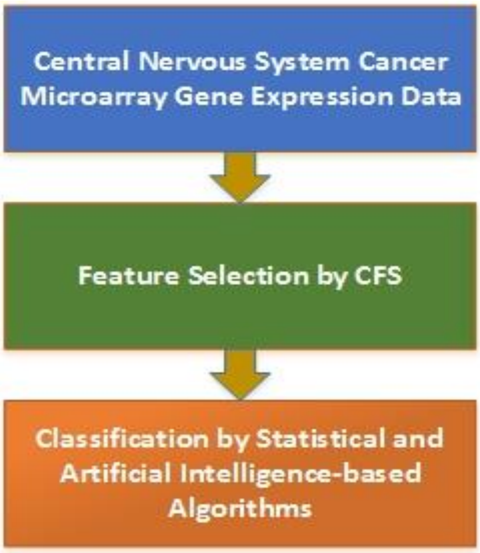

Central Nervous System Cancer Microarray Gene Expression Data
Feature Selection by CFS
Classification by Statistical and Artificial Intelligence-based Algorithms

**Figure 3.** Basic steps of the experimental study

To select an important subset of genes from thousands of genes is pretty arduous. That's why, gene selection becomes the most needed requirement for a diagnostic classifying system. For this reason, many researchers have applied different techniques to select a small subset of informative genes that can classify different subgroups of cancers accurately. The same experiment was repeated for six classifiers. The classification methods called statistical and artificial intelligence-based were applied to the dataset by performing feature selection and we tested the accuracy of our classification methods with 10-fold cross validation. The results of the classification algorithms have been shown Figure 4.

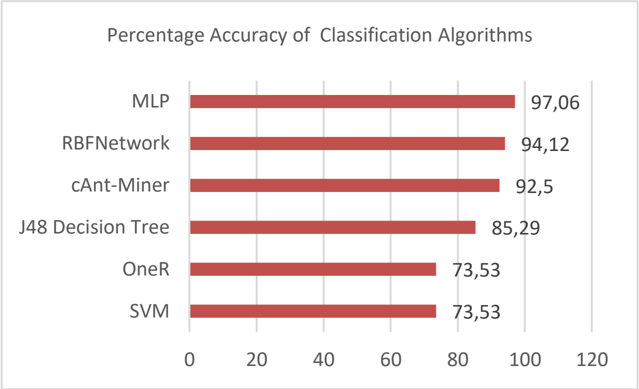


**Figure 4.** Percentage accuracy of 10-fold cross validation of classification algorithms for all genes on Central Nervous System cancer data

According to Figure 4, MLP has the best performance among classification methods which performed on central nervous system microarray gene expression dataset with 97.06%. RBFNetwork and cAnt-Miner have the best results respectively, with 94.12% and 92.50% after MLP. SVM and OneR have the worst performances, with 73.53% in the experimental study.

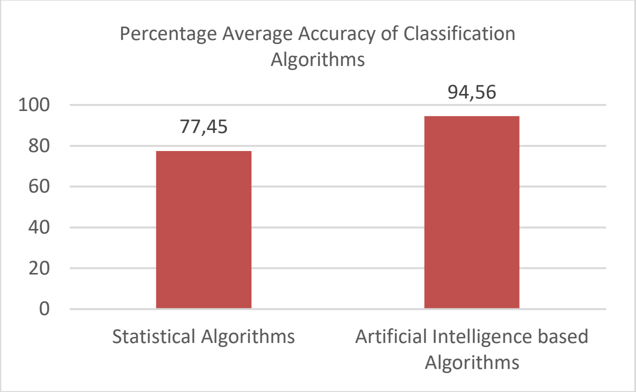


**Figure 5.** Percentage average accuracy of statistical and artificial intelligence based classification algorithms on the experimental data

As it is shown in Figure 5, the average performance of artificial intelligence-based algorithms is better than the average performance of statistical algorithms. Approximately, the average performance of artificial intelligence-based algorithms is 22.09% more than the average performance of statistical algorithms. Artificial Intelligence-based classification methods performed very well on experimental dataset with, 94.56%. However, the average of statistical methods is less than the average of artificial intelligence-based methods, with 77.45%.

## 5. Conclusion

In this paper, we have conducted a comparative study of classification algorithms for microarray data analysis on publicly available dataset including microarray central nervous system cancer dataset. Firstly, we applied Correlation-based Feature Selection (CFS) which is method of dimension reduction on the microarray dataset and then we compared the performances of six classification algorithms, namely Support Vector Machine (SVM), OneR, J48 Decision Tree, Multi-Layer Perceptron (MLP), Radial basis Function Network (RBFNetwork) and cAnt-Miner on central nervous system dataset by using control parameters most commonly used in the literature. In conclusion, the experimental results show that the artificial intelligence-based algorithms have higher accuracy than the statistical algorithms. In the future, we will study on control parameter optimization of classification methods and then, we will compare results of classification algorithms with parameters most commonly used in the literature and with optimal control parameters. Furthermore, we will also apply artificial-intelligence based algorithms on different microarray data.

## Acknowledgements

This work was supported by the Research Fund of Erciyes University of Turkey, grant number: FYL-2015-6095.

## References

[1] H. Liu, I. Bebu, and X. Li, “Microarray probes and probe sets.,” *Front Biosci (Elite Ed)*, vol. 2, pp. 325–38, 2010.

[2] H. U. Luleyap, *The Principles of Moleculer Genetics*. Izmir: Nobel Bookstore, 2008.

[3] K. Ipekdal, “Microarray Technology,” 2011. [Online]. Available: http://yunus.hacettepe.edu.tr/~mergen/sunu/s_mikroarrayandecology.pdf. [Accessed: 05-Jul-2016].

[4] M. a. Hall and L. a. Smith, “Practical feature subset selection for machine learning,” *Comput Sci*, vol. 98, pp. 181–191, 1998.

[5] V. Vapnik and V. Vapnik, *Statistical learning theory*. 1998.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines,” *Mach Learn*, vol. 46, no. 1/3, pp. 389–422, 2002.

[7] J. Novakovic, M. Minic, and A. Veljovic, “Genetic Search for Feature Selection in Rule Induction Algorithms,” pp. 1109–1112, 2010.

[8] C. Saylan, “Intelligent method based on new feature selection algorithm on renal transplantation patients,” Kadir Has University, 2013.

[9] E. Oztemel, *Artificial Neural Network*. Papatya Publishing, 2003.

[10] E. Cetin, “The Applications of Artificial Intelligence,” Ankara,Turkey: Seckin Publishing, 2007, pp. 379–401.

[11] D. S. Broomhead and D. Lowe, “Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks,” 1988.

[12] F. M. Ham and I. Kostanic, *Principles of neurocomputing for science and engineering*. McGraw Hill, 2001.

[13] A. C. Marco Dorigo, V. Maniezzo, Alberto Colorni, Marco Dorigo, Marco Dorigo, Vittorio Maniezzo, Vittorio Maniezzo, Alberto Colorni, “Positive Feedback as a Search Strategy,” 1991.

[14] B. Alatas and E. Akin, “The Discovery of Classification Rules by Ant Colony Algorithm,” 2004.

[15] R. S. Parpinelli, H. S. Lopes, and A. A. Freitas, “An ant colony algorithm for classification rule discovery,” in *Data mining: A heuristic approach*, 2002, pp. 191–208.

[16] B. Liu, H. A. Abbass, and B. Mckay, “Classification Rule Discovery with Ant Colony Optimization,” *IEEE Comput Intell Bull* , vol. 3, no. 1, pp. 31–35, 2004.

[17] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. Mclaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Z. I, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. G. Ii, “Prediction of central nervous system embryonal tumour outcome based on gene expression,” vol. 415, no. January, pp. 436–442, 2002.

[18] “Central Nervous System Cancer Dataset,” 2013. [Online]. Available: http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/Affymetrix/pomeroy-2002-v1/pomeroy-2002-v1_database.txt. [Accessed: 12-Jul-2016].

[19] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, “Data mining in bioinformatics using Weka,” *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, Oct. 2004.