Artificial Immune Recognition System for DNA Microarray Data Analysis

Chuanliang Chen, Chuan Xu, Rongfang Bie
College of Information Science and Technology
Beijing Normal University
Beijing 100875, P. R. China
Corresponding Author: rfbie@bnu.edu.cn

X. Z. Gao
Department of Electrical Engineering
Helsinki University of Technology
Otakaari 5 A, FI-02150 Espoo, Finland
gao@cc.hut.fi

Abstract

Artificial Immune Systems (AIS) are emerging information processing methods, which embody the principles of biological immune systems for tackling complex real-world problems. The Artificial Immune Recognition System (AIRS) is a new kind of supervised learning AIS. The development of microarray technology has supplied a large volume of data for the prediction and diagnosis of cancer. Many popular machine learning techniques have been used in the microarray data analysis. In this paper, we apply AIRS to perform the microarray data classification based on an improved version of the information gain feature selection method. Three traditional classifiers have also been employed in our experiments for performance comparison. The results demonstrate the promising ability of AIRS in the microarray data analysis.

1. Introduction

The biological immune system is a robust, complex, adaptive, and elaborate system, which can defend the body from foreign pathogens. Artificial Immune Systems (AIS) are novel machine-learning algorithms that embody some of the natural immune principles and take advantage of the benefits of the biological immune system in tackling challenging real-world problems [1]. The AIRS (Artificial Immune Recognition System) is an emerging immune-inspired learning algorithm [2]. It has shown significant successes in various classification areas.

Recently, the rapid development of DNA microarrays technology has offered a global view of cells by the measurement of the expression levels of thousands of genes [3]. With the DNA expression microarray technology, we can classify different cancers according to the corresponding expression levels in the normal and tumor cells, discover the relationship among the genes, and identify the critical genes in the causes of diseases [4]. The typical applications of the DNA microarray technology include cancer classification [5, 6, 7], clinical diagnose [8, 9], gene function identification [10], and drug discovery [11]. However, classification of tissues on the basis of the DNA microarray data is a demanding problem, because the microarray data always has high dimensions. To deal with the high-dimension microarray data, numerous feature selection algorithms have been proposed.

In this paper, we employ the AIRS as an efficient classifier to analyze the microarray data. Three well-known data classification techniques are used to compare with the AIRS. The information gain as a feature selection approach is also investigated to improve the performances of the AIRS and other data classifiers.

The remainder of this paper is organized as follows. In section 2, we demonstrate the information gain feature selection method, and propose a modified version. We briefly present our AIRS-based microarray data analysis method in section 3. The three data classifiers for comparison are explained in section 4. Section 5 discusses the results of the experiments in details. Finally, some remarks and conclusions are drawn in section 6.

2. Gene Selection

It is well known that the expression levels of thousands of genes (or attributes of entries) have been measured. However, not all of them are always useful for classification. That is to say, many of the genes are irrelevant, while only a few marker components of the genes subsets can identify the type of a tissue. Too many genes used are indeed impeditive for the performance of classifiers. Therefore, extracting relevant genes and reducing the dimensions of microarray data are usually necessary. In this paper, we apply the information gain to evaluate the genes of the microarray datasets.

Information gain is the expected entropy reduction caused by partitioning the data according to an attribute [12]. Let S be a set of microarray samples, and the entropy of the entire set S is:

$$Ent(S) = -\sum_{i=1}^{n} \frac{|C_i|}{|S|} \log_2 (\frac{|C_i|}{|S|}), \qquad (1)$$

where $|C_i|$ is the number of training samples in cancer class C_i (for i=1,2,…,n, n is total number of cancer classes) , and $|S|$ is the cardinality of the entire set S. The lower the entropy, the purer the set is.

The information gain $InfoGain(S,t)$ of a gene (or data feature) t is defined as:

$$InfoGain(S,t) = Ent(S) - \sum_{v \in V(t)} \frac{|S_v|}{|S|} Ent(S_v), \qquad (2)$$

where S_v is the subset of S, for which gene has value v, and $V(t)$ is the set of all the possible values of gene t. However, the information gain as a metric to evaluate the importance of every gene can only give the entropy of a gene. The number of the genes selected to perform the latter classification task is difficult to choose. In this paper, we propose a metric—η, a so called 'representative ability', to evaluate the quality of a subset of genes. More precisely, let T be the set of genes of S and M a subset of T, and M's representative ability is defined as:

$$\eta = \frac{\sum_{t \in M} Infogain(S,t)}{\sum_{t \in T} Infogain(S,t)}, \qquad (3)$$

where t is an element of T. When η is 0.99, we can intuitively conclude that M preserves 99% information for classification of T. Since the information gain of genes/features always varies quiet differently among different datasets, it is hard to build up a quantitative relation between η and the number of the selected genes. In this paper, we also investigate the effects of η on the performances of the data classifiers.

3. Artificial Immune Recognition System

The AIRS algorithm is one of the widely applied artificial immune systems for classification problems. It has also been used in data mining [13,14,15]. In this section, we present a modified AIRS1 proposed in [15], AIRS2. The effectiveness of the AIRS2 is further demonstrated in section 5.

3.1. Initialization

The initialization of the AIRS can be considered as a data pre-processing phase with parameter discovery [15]. During initialization, all the samples in a microarray dataset are normalized so that the Euclidean distance used in the AIRS between the feature vectors of any two items is within [0, 1]. Besides the Euclidean distance, other distance measures can be employed to evaluate affinity as well.

The affinity threshold, the average affinity value over all the training data, is calculated after normalization. The affinity threshold is obtained by using:

$$Affiniy\ Threshold = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} affinity(ag_i, ag_j)}{\frac{n(n-1)}{2}}, \qquad (4)$$

where n is the number of the training samples (or antigens), ag_i and ag_j are the i^{th} and j^{th} training antigen training vector, and $affinity(x, y)$ represents the affinity (Euclidean distance in this paper) between two antigens.

The seeding of the memory cells and initiation of the ARB population are at the last step of initialization.

3.2. Memory Cell Identification and Antigen Training

The goal of this stage in the AIRS is to train the antigens. The first step is to identify the memory cells and generate the ARB. The memory cells for a given antigen are discovered by calculating a matching metric—mc_{match}:

$$mc_{match} = \arg\max_{mc \in MC_{ag.c}} stimulation(ag, mc), \qquad (5)$$

where $stimulation(x, y)$ is defined in (6):

$$stimulation(x, y) = 1 - affinity(x, y). \qquad (6)$$

If $MC_{ag.c} \equiv \varnothing$, $mc_{match} \leftarrow ag$, and $MC_{ag.c} \leftarrow MC_{ag.c} \cup ag$. This memory cell is used to generate new ARBs to be included into the population of the previous ARBs.

3.3. Competition for Limited Resources

At this stage, the amount of the total resources is defined by the user. In the procedure of resource allocation, the resources are allocated to an ARB in accordance with its clone rate and normalized stimulation value. If this allocation of resources results in more resources being allocated than allowed, the resources are removed from the weakest (least stimulated) ARBs until the total number of resources is below the desired level. Finally, those ARBs with zero resources are removed. When the mean of the normalized stimulation level is above a preset threshold, this ARB refinement process is terminated, and the ARB with the greatest normalized stimulation value is selected to be the memory cell candidate.

3.4. Memory Cell Construction

This final training stage selects the newly-developed candidate memory cells, $mc_{candidate}$, into the set of the existing memory cells MC. A candidate memory cell is added to the set of memory cells, only if it is more stimulated than mc_{match}. The replacement of $mc_{candidate}$ with mc_{match} in the set of memory cells is determined by whether the affinity between $mc_{candidate}$ and mc_{match} is less than the product of the *affinity threshold* and *affinity threshold scalar*.

3.5. Data Classification

After the training process is completed, the evolved memory cells become the core of the AIRS-based data classifier. The process of data classification is based on a k-nearest neighbor approach and majority vote. Note that

the data vectors in the cells need be denormalized in classification.

3.6. A Modified AIRS—AIRS2

The above data classifier—AIRS1 can be modified by introducing the *Memory cell evolution* and *Somatic hypermutation* [15] so that the AIRS2 is proposed. Compared with the AIRS1, it has been shown [15] to be a simpler but more efficient artificial immune algorithm. In this paper, we employ AIRS2 for the microarray data analysis.

4. Classifiers in Performance Comparison

Three traditional classifiers are used to compare with AIRS2 concerning their classification abilities. They are: kNN, OneR, and Naïve Bayes. A brief introduction of these techniques is given below.

kNN: kNN is a memory-based classification method, which is a non-parametric inductive learning algorithm storing the training instances in a memory structure. It can predict the class label of a data point by a majority vote of the k nearest neighbors of this data point with ties broken at random. k is set to be 1 in our experiments.

OneR: OneR is a simple but efficient data classifier, which classifies the instances via generating a one-level decision tree [17]. Although the rules produced by the OneR are slightly less accurate than that of the state-of-the-art classifiers, these rules are relatively easy for experts to interpret [18].

Naïve Bayes: Naïve Bayes is one of the most popular data classifiers, which searches for a class label that maximizes the posterior probability based on the Bayes rule.

5. Experiments and Analysis

In this section, we firstly describe the four microarray datasets used in our experiments, and next present and analyze the result.

5.1. Microarray Datasets

We explore the capabilities of the AIRS on four microarray datasets: Colon cancer, Brain tumor, DLBCL and Nine tumor. The details of these datasets are given as follows.

Colon cancer dataset: Colon adenocarcinoma tissues are collected from patients, and from some of these patients, paired normal colon tissues also are obtained [19]. The data set contains the expression of the 2,000 genes with the highest minimal intensity across the 62 tissues. Every gene intensity has been derived from the about 20 feature pairs that correspond to the gene on the chip by using a filtering process. This dataset consists of 22 normal as well as 40 tumor samples.

Brain tumor dataset: This microarray dataset is actually the version of the one in [20] from a study of brain tumor. It consists of 5 types of brain tumor: Medulloblastoma, Malignant glioma, AT/RT, Normal cerebellum, and PNET. The dataset contains 5,921 genes and 90 samples. There are totally 60 cases of Medulloblastoma, 10 cases of Malignant glioma, 10 cases of AT/RT, four cases of Normal cerebellum, and six cases of PNET.

Nine tumors dataset: This microarray dataset consists of nine various human tumor types, NSCLC, Colon, Breast, Ovary, Leukemia, Renal, Melanoma, Prostate, and CNS. There are 5,726 genes and 60 samples. It contains 9 cases of NSCLC, 7 cases of Colon, 8 cases of Breast, 6 cases of Ovary, 6 cases of Leukemia, 8 cases of Renal, 8 cases of Melanoma, 2 cases of Prostate, and 6 cases of CNS. More details can be found in [21].

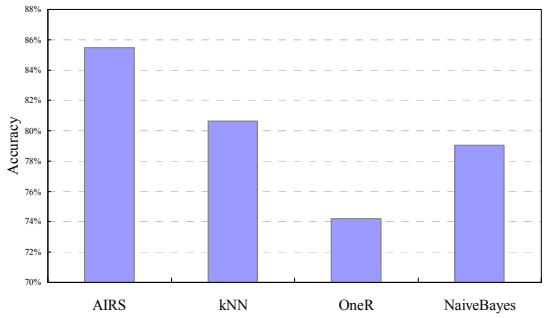
DLBCL dataset: The details of the Diffuse Large B-Cell Lymphoma (DLBCL) microarray dataset are explained in [22]. The version of this dataset used here contains 5,470 genes and 77 samples. Among the 77 samples, there are 58 samples of diffuse large B-cell lymphomas (DLBCL) and 19 samples of follicular lymphoma (FL).

5.2. Results and Analysis

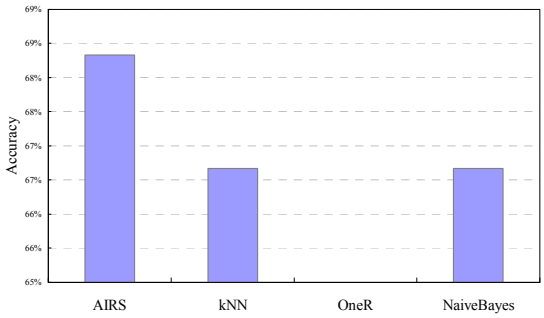
The 10-fold cross-validation on the above four microarray datasets is developed in this paper to evaluate the performances of aforementioned classifiers and improved information gain feature selection method. The classification accuracy is used as the performance index. We first demonstrate the histograms of the four data classifiers in Fig. 1. The results show that the performance of the AIRS is better than those of the other three data classifiers. The η = 0.99 implies that 99% classification information has been preserved. In fact, when η =0.99, the dimensions of the microarray datasets have been reduced significantly: the dimensions of the four datasets are Colon: 134, Brain tumor: 1587, Nine tumors: 163, and DLBCL: 869. The dimension reduction percentages are Colon: 93.30%, Brain Tumor: 73.20%, Nine Tumor: 97.15%, and DLBCL: 84.11%.

We also study the effect of η on the data classification performance. Only two datasets, Colon and DLBCL datasets, are selected in this case. Figure 2 shows the change of the classification accuracy of the four data classifiers when η varies.

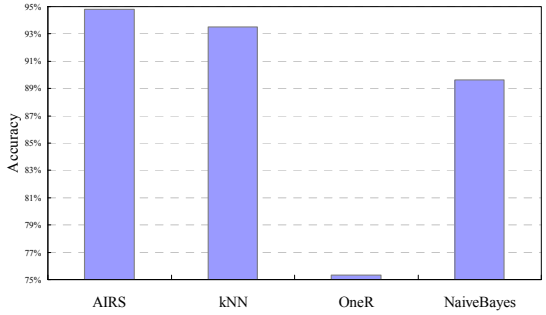
From Fig. 2, we can find out that when η is 1.0, i.e., all the genes are used in the process of classification, the performances of all the four classifiers become worse. Thus, it is concluded that 'many of genes are irrelevant for a particular classification problems, while only a few marker components of genes subsets can identify the type of a tissue'. However, in this case, the AIRS still behaves



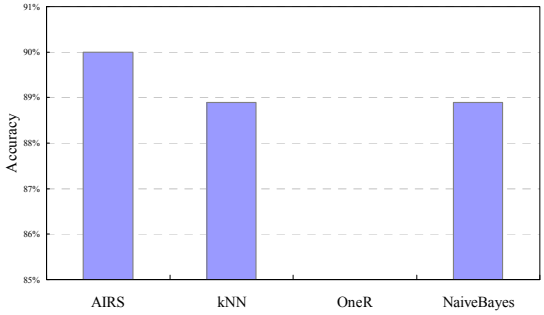
(a) Results on Colon cancer dataset.



(b) Results on Nine tumor dataset.

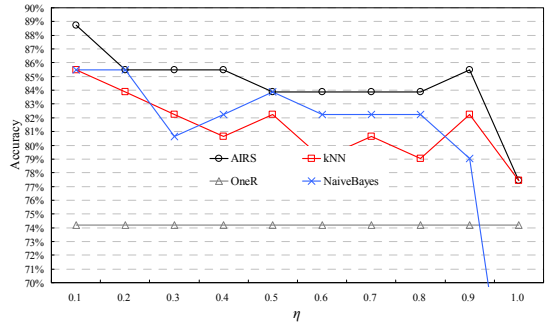


(c) Results on DLBCL dataset.

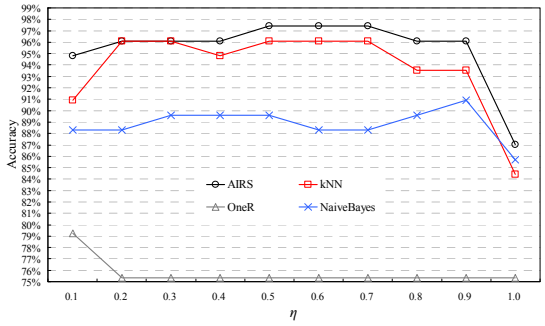


(d) Results on Brain Tumor dataset.

Fig. 1. Accuracy of AIRS and other three classifiers on the four popular microarray datasets. The Information Gain gene selection method is used with η = 0.99.



(a) Changes of accuracy on Colon dataset



(b) Changes of accuracy on DLBCL dataset

Fig. 2. Changes of accuracy of the four classifiers on Colon dataset and DLBCL dataset, when η is set to be 10 different values.

better than the other three classifiers. In other words, the AIRS is a data classifier with superior performances. For the Colon dataset, the performances of the four classifiers achieve the highest accuracy, when η is 0.1, that is, for Colon dataset, a subset of genes with 10% of total classification information are enough to accomplish the task of classification. Concerning the DLBCL dataset, the highest accuracy is achieved, when η is 0.5, 0.6, and 0.7. The number of the selected genes and dimension reduction percentage are summarized in Table 1. In summary, the information gain gene selection method can achieve high data classification accuracy with the irrelevant genes removed. As for a particular classification task, our empirical study shows that η=0.9 is enough for most demands. Moreover, Table 1 shows that there may be not a quantitative relation between η and the number of the selected genes.

Table 1. Summary of results of dimension reduction by Information Gain with various values of Representative Ability—η, RP is Reduction Percentage.

η	Colon dataset		DLBCL dataset	
	Dimensions	RP	Dimensions	RP
0.1	8	99.60%	47	99.14%
0.2	18	99.10%	107	98.04%
0.3	29	98.55%	177	96.76%
0.4	41	97.95%	255	95.34%
0.5	54	97.30%	340	93.78%
0.6	69	96.55%	432	92.10%
0.7	84	95.80%	533	90.25%
0.8	101	94.95%	641	88.28%
0.9	118	94.10%	757	86.16%
1.0	2000	0.00%	5469	0.00%

6. Remarks and Conclusions

In this paper, we employ the AIRS for data classification in the microarray data analysis. The AIRS is compared with three well-known data classifiers: kNN, OneR, and Naïve Bayes. Moreover, the Information Gain with parameter—"Representative Ability" is used to reduce the number of genes for improving the performances of these classifiers. The data classification results on four popular microarray datasets show that the AIRS is a promising classifier for microarray data classification, and the Information Gain gene selection method can enhance the performances of all the four classifiers with reduced dimensions of the microarray data.

Acknowledgements

The research work in this paper was supported by grants from the National Natural Science Foundation of China (Project No. 10601064). X. Z. Gao's research work was funded by the Academy of Finland under Grant 214144.

References

[1] A. Watkins, J. Timmis, and L. Boggess, "Artificial Immune Recognition System (AIRS): An Immune Inspired Supervised Machine Learning Algorithm," *Genetic Programming and Evolvable Machines*, Springer Netherlands, 2004, pp. 291-317.

[2] A. Watkins, "A resource limited artificial immune classifier," *Master's thesis*, Dept. Comp. Sci., Mississippi Sate University, Mississippi, 2001.

[3] A. D. Keller, M. Schummer, L. Hood, W. L. Ruzzo, "Bayesian Classification of DNA Array Expression Data," University of Washington, Washington, D.C., Tech. Rep. UW-CSE-2000-08-01, 2000.

[4] H. Hu, J. Li, A. Plank, H. Wang, G. Daggard, "A Comparative Study of Classification Methods for Microarray Data Analysis," In *Proc. of Australasian Data Mining Conference (AusDM2006)*, Sydney, 2006, pp. 33-37.

[5] T. Golub, D. Slonim, P. Tamayo, et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science* 286, 1999, pp. 531-537.

[6] L. V. Veer, H. Dai, M. V. de Vijver, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature* 415, 2002, pp. 530-536.

[7] E. Petricoin, A. Ardekani, B. Hitt, P. Levine, et al., "Use of proteomic patterns in serum to identify ovarian cancer," *The lancet* 359, 2002, pp. 572-577.

[8] K. Lu, A. P. Patterson, et al., "Selection of potential markers for epithelial ovarian cancer with gene expression arrays and recursive descent partition analysis," *Clin. Cancer Res.* 10, 2004, pp. 291-300.

[9] A. Santin, F. Zhan, S. Bellone, M. Palmieri, "Gene expression profiles in primary ovarian serous papillary tumors and normal ovarian epithelium: identification of candidate molecular markers for ovarian cancer diagnosis and therapy," *International Journal of Cancer* 112, 2004, pp. 14-25.

[10] C. Yeang, S. Ramaswamy, P. Tamayo, et al., "Molecular classification of multiple tumor types," *Bioinformatics 17(Supplement 1)*, 2001, pp. 316-322.

[11] O. Maron, T. Lozano-P´erez, "A framework for multiple-instance learning," *Advances in Neural Information Processing Systems*, 1998, pp. 570-576.

[12] J. Han, M. Kamber, "Data mining concepts and techniques," Morgan Kaufmann Publishers, 2000.

[13] D. Goodman, L. Boggess, A. Watkins, "An Investigation into the Source of Power for AIRS, An Artificial Immune Classification System," in *Proc. of the 2003 International Joint Conference on Neural Networks*, 2003, pp. 1678-1683.

[14] A. Watkins, J. Timmis, "Artificial Immune Recognition System (AIRS): Revisions and Refinements," in *Proc. of the 1st International Conference on Artificial Immune Systems*, Canterbury, UK, 2002, pp. 173-181.

[15] A. Watkins, J. Timmis, L. Boggess, "Artificial Immune Recognition System (AIRS): An Immune Inspired Supervised Machine Learning Algorithm," *Genetic Programming and Evolvable Machines*, 2004, pp. 291-317.

[16] J. Brownlee, "Artificial Immune Recognition System (AIRS): A Review and Analysis," Swinburne University of Technology, Melbourne, Australia, Tech. Rep. No. 1-02, Jan. 2005.

[17] R. C. Holte, "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets," *Machine Learning*, 1993, pp. 63-91.

[18] G. Buddhinath, D. Derry, "A Simple Enhancement to One Rule Classification," University of Washington, Washington, D.C., Tech. Rep., Melbourne, Australia, Tech. Rep., 2006.

[19] A. Alon, N. Barkai, D. A. Notterman, et al., "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," in *Proc. Natl. Acad. Sci. USA*, 1999, pp. 6745-6750.

[20] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature* 415, 2002, pp. 436-442.

[21] A. Bhattacherjee, W. G. Richards, J. Staunton, et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," in *Proc. Natl. Acad. Sci. USA*, 2001, pp. 13790-13795.

[22] M A. Shipp, K N. Ross, P. Tamayo, et al., "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning," *Nature Medicine*, 2002, pp 68-74.