# Generation of Comprehensible Hypotheses from Gene Expression Data

Yuan Jiang, Ming Li, and Zhi-Hua Zhou

National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{jiangy,lim,zhouzh}@lamda.nju.edu.cn

**Abstract.** Machine learning techniques have been recognized as powerful tools for the analysis of gene expression data. However, most learning techniques used in class prediction in gene expression analysis during the past years generate black-box models. Although the prediction accuracy of these models could be very well, they provide little insight into the biological facts. This paper holds the recognition that a more reasonable role for machine learning techniques is to generate hypotheses that can be verified or refined by human experts instead of making decisions for human experts. Based on this recognition, a general approach to generate comprehensible hypotheses from gene expression data is described and applied to human acute leukemias as a test case. The results demonstrate the feasibility of using machine learning techniques to help form hypotheses on the relationship between genes and certain diseases.

## 1 Introduction

DNA arrays consist of a large number of DNA molecules spotted in a systemic order on a solid substrate. When the diameter of the DNA spot is less than $250\mu m$, DNA arrays can be categorized as microarrays [14][19]. With the development of microarray technology, the simultaneous measurement of gene-expression levels for thousands of genes is now possible. Analyzing gene expression data could be helpful for medical treatment. For example, systematic and unbiased approaches could be developed for cancer classification through analyzing gene expression data, which is very important for cancer treatment [8]. However, as keeping track of thousands of measurements and their relationships is overwhelmingly complicated, gene expression data is difficult to analyze without the help of computers.

During the past years, machine learning techniques have been recognized as powerful tools for gene expression analysis [16]. Machine learning [15] is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. It is closely related to pattern recognition and statistical inference, and has data mining as its engineering application aspect. Machine learning techniques such as clustering, neural networks, hidden Markov models, and nonlinear regression have already been widely used in the practice of engineering, business, and science.

In analyzing gene expression data, most work had primarily been descriptive rather than analytical and had focused on cell culture rather than primary patient material, in which genetic noise might obscure an underlying reproducible expression pattern. Since Golub et al.'s work [8], learning techniques such as neural networks [1][8], support vector machines [2][7], ensemble learning methods [1][2][4][5][22], nearest neighbor classifiers [2], logistic regression [13], linear discriminant analysis [5][17], emerging patterns [12], etc. have been applied to gene expression data, where the goal is to classify cases into diagnostic or prognostic categories. However, almost all the learning techniques used in gene expression analysis during the past years generate black-box models. Although the prediction accuracy of these models could be very well, they provide little insight into the biological facts and can hardly provide explicit explanations for their predictions.

Imagine the scenario where a patient asked the doctor why he made a specific diagnosis, the doctor said he could not explain because he did not know either. Of course such a diagnosis is unacceptable. This somewhat reflects the problem of prediction with black-box models. In fact, it is not reasonable to anticipate computers replace experienced human medical experts. Therefore a more reasonable role for machine learning techniques is to generate hypotheses that could be verified or refined by human experts, which might be the basis for further understanding of the relationship between specific genes and diseases. It is evident that black-box models are helpless to this purpose.

In fact, there are many works devoted to the improving of the comprehensibility of black-box models [24], some of which has already been applied to medical diagnosis [10][21]. Unfortunately, as mentioned before, in the analysis of gene expression data, almost all the methods used before generate black-box models.

In this paper, a general approach to generate comprehensible hypotheses from gene expression data is described, which is based on a recent achievement of machine learning, i.e. the C4.5Rule-PANE method [25]. This paper applies this approach to human acute leukemias as a test case. The results demonstrate the feasibility of using machine learning techniques to help disclose the relationship between genes and certain diseases.

The rest of this paper is organized as follows. Section 2 briefly introduces the C4.5Rule-PANE method. Section 3 describes the case study on generating comprehensible hypotheses for human acute leukemias from gene expression data. Section 4 concludes.

## 2 The Method

Although the main purpose of this paper is to report on the application of the C4.5Rule-PANE method to human acute leukemias, for the self-containness of this paper, here a brief introduction on the method is given. Interested readers can refer [25] for more details.

Suppose there is a gene expression data set $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l)\}$, where $\mathbf{x}_i$ is a feature vector coding the gene measurements on a case, $y_i$ is the known outcome of the case corresponding to $\mathbf{x}_i$, $l$ is the number of cases with known outcomes.

At first, bootstrap sampling [6] is employed to produce $N$ data sets with the same size as $|S|$, i.e. the number of cases in $S$. A neural network classifier is then trained from each of these new data sets. Therefore a neural network ensemble [27] is obtained.

After that, the neural network ensemble is used to judge the cases in $S$. In detail, all the feature vectors $\mathbf{x}_i$ $(i = 1, 2, \cdots, l)$ are fed to the neural network classifiers. For $\mathbf{x}_i$, $N$ predictions will be provided, each by one neural network classifier. Then, the majority of these $N$ predictions is regarded as the outcome of the neural network ensemble on $\mathbf{x}_i$, which is denoted as $y'_i$. Therefore, after processing all the cases in $S$, a new data set $S' = \{(\mathbf{x}_1, y'_1), (\mathbf{x}_2, y'_2), \cdots, (\mathbf{x}_l, y'_l)\}$ is generated.

Moreover, a set of random vectors $\mathbf{x}^*_j$ $(j = 1, 2, \cdots, m)$ can be generated, where the $k$-th element of $\mathbf{x}^*_j$ is a value randomly chosen from the values that could appear on the $k$-th gene measurement. These vectors are fed to the neural network ensemble. Let the outcome of the neural network ensemble on $\mathbf{x}^*_j$ be denoted as $y^*_j$. Then a data collection $S^* = \{(\mathbf{x}^*_1, y^*_1), (\mathbf{x}^*_2, y^*_2), \cdots, (\mathbf{x}^*_m, y^*_m)\}$ is generated. Through combining $S'$ and $S^*$, a new data set $S^{**}$ is obtained. The size of $S^{**}$ can be controlled by the parameter $\mu = m/l$. Note that $S^{**}$ could be far bigger than the original data set $S$ because $m$ could be far bigger than $l$.

Then, a C4.5 decision tree [20] is trained from $S^{**}$ and every path from the root to a leaf is converted to an initial propositional rule by regarding all the test conditions appearing in the path as the conjunctive rule antecedents while regarding the classification held by the leaf as the rule consequence. All the initial rules are generalized by removing antecedents that do not seem helpful for distinguishing a specific class from other classes. Rules that do not contribute to the accuracy of the rule set are also removed. A default rule is created for dealing with cases that have not been covered by any of the generated rules, which has no antecedent and the consequence is the biggest class among these cases. The resulting rules could finally be organized into an 'IF-THEN-ELSE' format with embedding 'IF-THEN-ELSE' structures, which corresponds to a hypothesis generated from the gene expression data.

The details of C4.5Rule-PANE can be found in [25], whose computational cost is very close to that of the neural network ensemble. It has been proven [26] that training a neural network ensemble and then using it to predict the training data, which is then given to another learning approach, could be beneficial. The premise is that the original training data set contains much noise and has not fully captured the target distribution, and the neural network ensemble is more accurate than the model directly trained from the original training data set by the second learning approach. It is evident that the first condition is very easy to meet because in rare applications the training data set does not contain much noise and does fully capture the target distribution.

Note that the C4.5Rule-PANE algorithm is not an algorithm which simply extracts faithful rules from neural network ensembles [24]. Instead, some mistakes made by the neural network ensemble could be corrected during the learning of the rules. Thus, the generalization ability of C4.5Rule-PANE can be even higher than that of the neural network ensemble [25][26].

It is also worth noting that in analyzing gene expression data, an often encountered problem is that there are only a small number of cases with known outcomes. Thus, models built by machine learning techniques are not very reliable since the training set is too small. C4.5Rule-PANE has two keys to relax this limitation. The first is the generation of $S^*$ with the help of neural network ensemble, which could greatly enlarge the training set for the rule generation process. The second is the comprehensible rules it generated, which could be verified by human experts instead of could only be used alone.

In fact, different rules can be produced if C4.5Rule-PANE is run for several times. These rules can be validated if there are cases with known outcomes that have not been used in training, and the rule with the highest validating accuracy can be regarded as the final hypothesis which might be helpful in disclosing the relationship between certain genes and diseases.

## 3 Case Study

Chemotherapy regimens for acute lymphoblastic leukemia (ALL) generally contain corticosteroids, vincristine, methotrexate, and L-asparaginase, whereas most acute myeloid leukemia (AML) regimens rely on a backbone of daunorubicin and cytarabine [3][18]. Although remissions can be achieved using ALL therapy for AML and vice versa, cure rates are markedly diminished, and unwarranted toxicities are encountered. So, distinguishing ALL from AML is critical for successful treatment.

The data were taken from [8]. The training data set consists of 11 AML cases and 27 ALL cases, while the test data set consists of 14 AML cases and 20 ALL cases. Each case is composed of 7,129 gene expressions.

Since the number of gene measurements is very large compared to the number of cases, feature selection in the context of gene expression analysis has been investigated and found beneficial [1][4][9][11]. In a previous work where seven different feature selection methods were applied to the concerned data, it was found that among the 7,129 measurements, there were only 30 being chosen by more than two methods [4]. Therefore, here only these 30 measurements are used. The results reported below show that such a feature selection scheme is quite effective. Note that among these 30 measurements only 16 measurements finally appear in the hypotheses generated by C4.5Rule-PANE, as shown in Table 1.

Through setting the parameter $\mu$ to 1, 2, 3 and 4, respectively, four different hypotheses have been generated by C4.5Rule-PANE, as listed in Table 2. Here $(x/y)$ at the end of each line indicates that among the 34 test cases, $y$ cases were judged by this line while $x$ cases were correctly judged.

**Table 1.** Measurements appear in the hypotheses

| ID# | Gene accession number | Gene description |
|---|---|---|
| 5 | AFFX-BioC-3_at | AFFX-BioC-3_at (endogenous control) |
| 8 | AFFX-CreX-5_at | AFFX-CreX-5_at (endogenous control) |
| 13 | AFFX-BioC-5_st | AFFX-BioC-5_st (endogenous control) |
| 22 | AFFX-DapX-3_at | AFFX-DapX-3_at (endogenous control) |
| 461 | D49950_at | Liver mRNA for interferon-gamma inducing factor (IGIF) |
| 1834 | M23197_at | CD33 CD33 antigen (differentiation antigen) |
| 1882 | M27891_at | CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage) |
| 2020 | M55150_at | FAH Fumarylacetoacetate |
| 2242 | M80254_at | PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR |
| 2402 | M96326_rnal_at | Azurocidin gene |
| 2759 | U12471_cds1_at | Thrombospondin-p50 gene extracted from Human thrombospondin-1 gene, partial cds |
| 3258 | U46751_at | Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA |
| 3320 | U50136_rnal_at | Leukotriene C4 synthase (LTC4S) gene |
| 4847 | X95735_at | Zyxin |
| 5039 | Y12670_at | LEPR Leptin receptor |
| 6201 | Y00787_s_at | INTERLEUKIN-8 PRECURSOR |

The test accuracy of any of these hypotheses is 97.1%. In fact, each hypothesis made only one misclassification. Note that endogenous control measurements appear in Hypothesis 3 and Hypothesis 4, and therefore Hypothesis 1 and Hypothesis 2 are more preferable.

Since the data used in this case study has been widely investigated, the test accuracy of the above hypotheses can be compared with the best accuracy reported by different researchers, as shown in Table 3. Note that since the machine learning techniques used in most previous research on human acute leukemias generate black-box models, the comprehensibility of the hypotheses generated by C4.5Rule-PANE is much better.

Table 3 shows that the accuracy of the hypotheses generated by C4.5Rule-PANE is very comparable to the best result reported before. In fact, since the test data set is very small (only 34 cases), the reliability of a model is not guaranteed even though its test accuracy reaches 100%. Different to the black-box models used before [1][2][4][5][7][8][13][17][22], the hypotheses generated by C4.5Rule-PANE can be verified by human experts. Therefore, it can be anticipated that they are of greater value in prediction than black-box models. Moreover, it is worth noting that the value of these hypotheses are beyond pure prediction, because they are comprehensible and might help human experts to disclose the relationship between certain genes and diseases.

**Table 2.** Hypotheses generated on the human acute leukemias data set

| | |
|---|---|
| Hypothesis 1: | |
| IF (M27891_at > 820) AND (U12471_cds1_at > 145) THEN class = AML | (11/11) |
| ELSE IF (X95735_at ≤ 1380.257218) THEN class = ALL | (18/18) |
| ELSE IF (U50136_rna1_at ≤ 1464.485389) THEN class = ALL | (2/3) |
| ELSE class = AML | (2/2) |
| Hypothesis 2: | |
| IF (Y00787_s_at ≤ 523) AND (M23197_at ≤ 472.769728) THEN class = ALL | (16/17) |
| ELSE IF (D49950_at > 58.110118) AND (M23197_at > 191.728789) AND (U12471_cds1_at > 56.992413) THEN class = AML | (13/13) |
| ELSE IF (U46751_at > 5402.227657) AND (M96326_rna1_at ≤ 17475.565446) THEN class = AML | (0/0) |
| ELSE class = ALL | (4/4) |
| Hypothesis 3: | |
| IF (M27891_at ≤ 1358) AND (Y00787_s_at ≤ 11858.081462) THEN class = ALL | (19/19) |
| ELSE IF (U12471_cds1_at > 173.296678) THEN class = AML | (9/9) |
| ELSE IF (M27891_at > 7090.800758) AND (M80254_at ≤ 693) AND (AFFX-BioC-3_st ≤ -225) AND (M27891_at ≤ 14708.236206) THEN class = ALL | (0/0) |
| ELSE IF (Y12670_at ≤ 2017.153487) THEN class = AML | (5/6) |
| ELSE IF (AFFX-CreX-5_at ≤ -393.142106) THEN class = AML | (0/0) |
| ELSE class = ALL | (0/0) |
| Hypothesis 4: | |
| IF (M27891_at > 851.428976) AND (U12471_cds1_at > 139.014026) AND (D49950_at > 19.992492) THEN class = AML | (12/12) |
| ELSE IF (Y12670_at > 1233.788534) AND (M55150_at > 284.216049) THEN class = AML | (0/0) |
| ELSE IF (M23197_at ≤ 288) THEN class = ALL | (19/19) |
| ELSE IF (AFFX-DapX-3_at ≤ -94.34693) THEN class = ALL | (1/2) |
| ELSE IF (AFFX-BioC-5_st ≤ -264.931499) THEN class = AML | (1/1) |
| ELSE class = ALL | (0/0) |

## 4 Conclusion

Machine learning techniques have been introduced into gene expression analysis recently. However, almost all the previous works emphasize on constructing models with high prediction accuracy, nevertheless the models are black-boxes. This paper claims that an important role for machine learning techniques to play in gene expression analysis is to help human experts grasp the laws behind biological facts, therefore comprehensible hypotheses might be more helpful than black-

**Table 3.** Comparing the test accuracy of the hypotheses generated by C4.5Rule-PANE with the best accuracy reported by different researchers

| Authors | Year | Method | Accuray |
|---|---|---|---|
| Golub et al. [8] | 1999 | Self-Organizing Map | 85.3% |
| Ben-Dor et al. [2] | 2000 | AdaBoost | 95.8% |
| Furey et al. [7] | 2000 | Support Vector Machine | 94.1% |
| Li & Yang [13] | 2001 | Logistic regression | 94.1% |
| Cho & Ryu [4] | 2002 | Ensemble of heterogeneous classifiers | 100% |
| Dudoit et al. [5] | 2002 | BoostCART | 95.0% |
| Li & Wong [12] | 2002 | Emerging Patterns | 91.2% |
| Nguyen & Rocke [17] | 2002 | Logistic discriminant | 97.1% |
| Albrecht et al. [1] | 2003 | Ensemble of perceptrons | 100% |
| Tan & Gilbert [22] | 2003 | Ensemble of decision trees | 91.2% |
| Yun & Keong [23] | 2005 | Discrete Function Learning | 94.1% |
| this paper | now | C4.5Rule-PANE | 97.1% |

box models. Based on this recognition, this paper presents a general approach to generate comprehensible hypotheses from gene expression data, which utilizes a recent proposed machine learning technique, i.e. the C4.5Rule-PANE method. Case study show that this approach work well on human acute leukemias. It is evident that such an approach can be applied to generate comprehensible hypotheses from other gene expression data, which may help enlarge the benefit from microarray technology. In the future the authors expect to work with experts on genes and diseases, wishing that the power of the C4.5Rule-PANE method can be really utilized.

## Acknowledgement

We want to thank the anonymous reviewers for their constructive comments. This work was supported by the National Science Foundation of China (60505013), the Jiangsu Science Foundation (BK2004001, BK2005412), and the Fok Ying Tung Education Foundation (91067).

## References

1. Albrecht, A., Vinterbo, S.A., Ohno-Machado, L.: An epicurean learning approach to gene-expression data classification. Artificial Intelligence in Medicine **28** (2003) 75–87
2. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z.: Tissue classification with gene expression profiles. Journal of Computational Biology **7** (2000) 559–584
3. Bishop, J.F.: Adult acute myeloid leukaemia: update on treatment. Medical Journal of Australia **170** (1999) 39–43

4. Cho, S.-B., Ryu, J.: Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. Proceedings of the IEEE **90** (2002) 1744–1753
5. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association **97** (2002) 77–87
6. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)
7. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16** (2000) 906–914
8. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537
9. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46** (2002) 389–422
10. Hayashi, Y., Setiono, R., Yoshida, K.: A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders. Artificial Intelligence in Medicine **20** (2000) 205–216
11. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine **7** (2001) 673–679
12. Li, J., Wong, L.: Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics **18** (2002) 725–734
13. Li, W., Yang, Y.: How many genes are needed for a discriminant microarray data analysis. In: Lin, S.M., Johnson, K.F. (eds.): Methods of Microarray Data Analysis. Kluwer, Boston, MA (2001) 137–150
14. Maughan, N.J., Lewis, F.A., Smith, V.: An introduction to arrays. Journal of Pathology **195** (2001) 3–6
15. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
16. Mjolsness, E., DeCoste, D.: Machine learning for science: state of the art and future prospects. Science **293** (2001) 2051–2055
17. Nguyen, D.V., Rocke, D.M.: Tumor classification by partial least squares using microarray gene expression data. Bioinformatics **18** (2002) 39–50
18. Pui, C.H., Evans, W.E.: Acute lymphoblastic leukemia. New England Journal of Medicine **339** (1998) 605–615
19. Quackenbush, J.: Computational analysis of microarray data. Nature Reviews Genetics **2** (2001) 418–427
20. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA (1993)
21. Setiono, R.: Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence in Medicine **18** (2000) 205–219
22. Tan, A.C., Gilbert, D.: Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics **2** (2003) S75–S83
23. Yun, Z., Keong, K.C.: Identifying simple discriminatory gene vectors with an information theory approach. In: Proceedings of the 4th IEEE Computational Systems Bioinformatics Conference. Stanford, CA (2005) 13–24
24. Zhou, Z.-H.: Rule extraction: using neural networks or for neural networks? Journal of Computer Science & Technology **19** (2004) 249–253

25. Zhou, Z.-H., Jiang, Y.: Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble. IEEE Transactions on Information Technology in Biomedicine **7** (2003) 37–42
26. Zhou, Z.-H., Jiang, Y.: NeC4.5: neural ensemble based C4.5. IEEE Transactions on Knowledge and Data Engineering **16** (2004) 770–773
27. Zhou, Z.-H., Wu, J., Tang, W.: Ensembling neural networks: many could be better than all. Artificial Intelligence **137** (2002) 239–263