

# Prediction vs. Description: Two Data Mining Approaches to the Analysis of Genetic Data

Christopher A. Shoemaker    Michael A. Sao Pedro    Sergio A. Alvarez\*    Carolina Ruiz

Department of Computer Science  
Worcester Polytechnic Institute  
Worcester, MA 01609. U. S. A.  
{cshoemak, falcor, ruiz}@cs.wpi.edu, salvarez@wellesley.edu

## Abstract

We analyzed the simulated disease data for the founders of the general population in the GAW12 dataset with two techniques rooted in the machine learning and data mining literature. We used Artificial Neural Networks (ANN) to infer a predictive model of the simulated disease, and Association Rule Mining (ARM) to search for a descriptive model. These two approaches are quite different in that ANN can quickly achieve very good predictive performance but the resulting models are difficult for humans to interpret, while classical ARM algorithms are extremely slow but any resulting association rules are fairly easy to understand. We describe these techniques below, together with some of the experimental results obtained by applying them to the GAW12 dataset.

## 1 Introduction

This paper addresses Problem 2 of the 12th Genetic Analysis Workshop (GAW12). “Answers” were not available to the authors of this study when the study was performed. Both genotypic and phenotypic information were employed; the former includes single nucleotide polymorphism (SNP) data and microsatellite marker data, and the latter refers mainly to quantitative trait data. The environmental factors in the GAW12 dataset were also considered. We restricted attention to the pedigree founders in the general population. Some additional details regarding data selection are given below; full details are given in an accompanying GAW12 abstract by Sao Pedro et al.

We analyzed the simulated disease data for the founders of the general population in the GAW12 dataset with two techniques rooted in the machine learning and data mining literature. We used Artificial Neural Networks (ANN) to infer a predictive model of the simulated disease, and Association Rule Mining (ARM) to search for a descriptive model. These two approaches are quite different in that ANN can quickly achieve very good predictive performance but the resulting models are difficult for humans to interpret, while classical ARM algorithms are extremely slow but any resulting association rules are fairly easy to understand. We describe these techniques below, together with some of the experimental results obtained by applying them to the GAW12 dataset.

There has been a significant amount of work employing machine learning methods in the analysis of genetic data. Clustering methods [4, 3] attempt to find meaningful groupings of the input data. Neural networks have been shown to be useful in a wide variety of data analysis problems, including determining consensus sequences [2]. They are particularly good at implicitly identifying functional relationships in data.

---

\*Current affiliation: Department of Computer Science, Wellesley College, Wellesley, MA 02481

## 2 Data Representation and Statistical Analysis

A data format encompassing information about phenotypic data, marker genotypes for all chromosomes, and sequence data for all genes was developed to join all relevant information for the entire population. Certain changes to the representation of the data were made to successfully perform the merge. SNP data were incorporated by changing their data representation so that each person had the same number of data elements for SNPs, and by changing the naming scheme so that all the SNPs from different genes could be incorporated in the same file. Satellite marker data were incorporated directly, but the satellite marker naming scheme was changed so that the names contained the number of alleles present in the population for that satellite marker. For more information on the data conversion, see the GAW12 abstract by Sao Pedro et al. Of the 9515 various SNPs present in the general population, we selected only SNPs for which at least one person in each replicate varied from the normal allele. There were 715 such SNPs. We employed a chi-square test of independence so that we could later select the most significant subset of SNPs and markers. One important result of this ranking was that when we later took the top 100 SNPs, we noticed that only SNPs in genes 1, 2 and 6 were present. The distribution of the top 100 SNPs among genes is shown below. For comparison, the distribution of all 715 SNPs among genes is also shown.

	gene 1	gene 2	gene 3	gene 4	gene 5	gene 6	gene 7
Breakdown of top 100 SNPs by gene	69	20				11	
Breakdown of 715 SNPs by gene	157	90	99	121	37	34	177

## 3 Association Rules

Association rules, independently introduced in [5] and [1], identify collections of data attributes that are statistically related in the underlying data. We mined associations among SNPs, markers, environmental factors, quantitative traits, and affection status in the general population.

### 3.1 Basic concepts

Association rules can be formally defined as follows. Let  $I$  be a set of binary attributes called items. Given a dataset  $D$  of transactions, where each transaction  $T$  is a set of items, an association rule is a rule of the form  $X \Rightarrow Y$ , where  $X$  and  $Y$  are sets of items. Such a rule is annotated with the statistical information that indicates the association of those items in the data. Usually two numbers are used for this purpose: the *confidence* of the rule, which is the percentage of data points that contain the items in  $Y$  among those that contain items in  $X$ ; and the *support* of the rule, which is the percentage of data points that contain all the items in  $X$  and  $Y$  in the dataset. Relating these notions to the genetic problem at hand,  $I$  represents all of the different phenotypic information, marker genotypes, and sequence data a person can have;  $T$  corresponds to a person; and  $D$  represents all of the people in a population.

We are specifically interested in those rules which have only “healthy” or “sick” on the right-hand side of the rule. Those rules are called *classification rules* since they predict the presence or absence of the disease.

### 3.2 Converting the GAW12 dataset into transactions

In order to mine association rules, each GAW12 variable was first mapped to a binary set. In order to map discrete variables such as SNPs and satellite markers to a binary set, we broke up all the possible values that an attribute could have into different variables and assigned a “yes” to the value associated with that variable, and a “no” for all the other variables. To map the continuous variables such as the quantitative factors to a binary set, the continuous variables were first broken into discrete intervals, and the procedure described above was followed. The continuous intervals were determined by finding the splits which maximized the information gain with respect to the affection status. In order to make the system search for the presence of either allele for a particular satellite marker, first a number of binary variables for that marker equal to the maximum number of values ii or jj could attain for all satellite markers was created, and then each of the two values were checked as “yes” for their appropriate variable, and the others “no”.

### 3.3 Summary of experimental results

We used two available tools for mining association rules, IBM’s Intelligent Miner for Data, and CBA (Classification Based Associations), an academic tool produced at the National University of Singapore by [8], in order to mine rules from the general population from all 50 replicates. Intelligent Miner, which uses the Apriori algorithm [1], scales linearly in the number of people but exponentially in the number of attributes. Therefore, it could not effectively mine rules due to the high dimensionality of the dataset. CBA, which specifically mines classification rules, also could not mine rules due to the number of attributes in the dataset.

To reduce the dimensionality so that CBA could handle the data, we mined classification rules for replicate 42 using only the 100 most significant SNPs and 100 most significant markers. After this dimensionality reduction, CBA discovered 18,996 rules when mining with a low support (above 1%) and a high confidence (above 80%). Some interesting rules that were discovered are shown in Figure 1. For example, the first rule states that people having a value of 08 in marker D10G108 and a value of 22 for SNP 11911 in gene 1 are predicted to be healthy with a support of 5.7% and a confidence of 98.246%.

Association Rule				Support	Confidence
(marker D10G108 = 08)	and	(SNP 11911 gene 1 = 22)	$\Rightarrow$ healthy	5.7%	98.246%
(marker D16G038 = 02)	and	(SNP 14061 gene 1 = 22)	$\Rightarrow$ healthy	3.7%	100%
(SNP 2540 gene 2 = 11)	and	(SNP 5542 gene 1 = 11)	$\Rightarrow$ healthy	3.4%	100%
(marker D16G038 = 02)	and	(SNP 13074 gene 1 = 11)	$\Rightarrow$ healthy	5.1%	98.039%
(SNP 14839 gene 1 = 11)	and	(SNP 1553 gene 1 = 11)	$\Rightarrow$ healthy	58%	85.69%

Figure 1: Sample classification rules mined from replicate 42.

One problem found when mining classification rules was that the system only mined rules that implied a person would be healthy. No rules were mined that implied a person would be sick. This was because there were fewer sick people than healthy people in the dataset (26% sick vs. 74% healthy). Therefore, there was less support for rules that implied that a person would be sick. We will address this problem by implementing our own association rule mining algorithm that effectively handles high dimensional data, and uses a class-based concept of support.

## 4 Neural Networks

Artificial neural networks (ANN) are a class of models of distributed computation inspired by biological ideas. Because of their roots in biology, it is especially important to note that ANN as considered here are not intended to be realistic models of computation in biological systems of neurons. Rather, ANN can be viewed as providing a mechanism for semi-automatically inferring implicit functional relationships in data. This section provides a brief overview of the ANN concepts that are most relevant to our analysis of the GAW12 dataset, together with a summary of the experimental results obtained using this technique as a means for inferring a predictive model of the simulated disease among the founders of the general population.

### 4.1 Basic concepts

A simplified form of ANN is defined by a weighted directed acyclic graph  $(G, E, w)$ , where  $G$  is the set of nodes (*processing units*) of the network, corresponding roughly to either individual neurons or functional groups of neurons;  $E$  is the set of edges of the graph, representing neuronal interconnections; and  $w$  is a real-valued function defining weights on the edges of the network. The state of each processing unit  $i$  is described by its *activation*, modeled as a real-valued function of time. The weight  $w_{i,j}$  on the edge  $E(j, i)$  (from unit  $j$  to unit  $i$ ) quantifies the degree of influence of unit  $j$  in determining the activation of unit  $i$ . The activation value  $a_i(t)$  of node  $i$  at time  $t$  is an instantaneous function of the activation values  $a_j(t)$  at the same instant of time  $t$  of all nodes  $j$  that feed into node  $i$ . More specifically, we take the activation  $a_i(t)$  to be the result of applying a nonlinear *activation function*  $f$  to a linear combination of the activations  $a_j(t)$ . Neural networks as described above are capable of a rudimentary form of “learning”, associated with changes in

the weights between processing units. Training may be performed by using the well-known method of error backpropagation (e.g. [9]). This supervised learning algorithm requires that a set of training pairs  $(I_k, O_k)$  be presented to the system; here,  $I_k$  is an input vector and  $O_k$  is the desired output vector corresponding to  $I_k$ . The name of the method is derived from the recursive form of these equations: the errors  $E_k$  at the output layer are “propagated back” through previous layers in order to adjust the connection weights.

## 4.2 Network architecture

We employed a feedforward network architecture with two distinct layers of processing units. The inputs feed directly into the units of the first, or “hidden” layer. The hidden units feed into the units of the second, or “output” layer. The network topology is “fully connected”, meaning that all pairs of (input, hidden) units are connected, as are all (hidden, output) pairs. Inputs were extracted from data for the founders of the general population in the GAW12 dataset. Some of the experiments used data for a single replicate while others used several replicates concurrently. Different combinations of input features were considered, including SNP data only, SNP and microsatellite marker data, and the latter together with data for the environmental factors and quantitative attributes. A single output unit representing the disease affection status was used. Weights of the connections between network units were initialized to small random values. The network was then trained using the method of error backpropagation ([9]) as described above. Convergence to an equilibrium state generally took place after a few hundred training iterations. The time required for training varied depending on the size of the input attribute set; the longest runs took on the order of 1 hour on a 400 MHz Pentium family PC with 128 MBytes of RAM.

## 4.3 Summary of experimental results

The results obtained using ANN show good predictive accuracy, with increasing accuracy as the amount of information available increases. For example, SNP-only accuracy was lower than that obtained when marker data was incorporated. During training, prediction error decreased from an initial value of 25%–30% corresponding roughly to the fraction of diseased individuals among the founders of the general population, to significantly lower values, less than 1% for some of the experiments. These results should be cautiously interpreted, as the error rate alluded to above was obtained when testing the networks on the same data used for training. When the available data was split into disjoint training and testing sets, the error rate obtained after training was in the 10%–15% range – still significantly lower (better) than what would be obtained by predicting the majority class (healthy) in all cases.

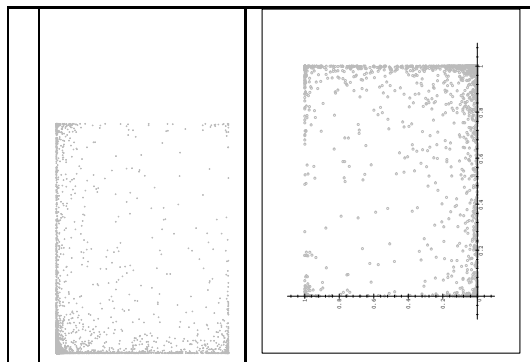


Figure 2: Learned features for healthy (left) and sick (right) individuals

Fig. 2 provides a graphical display of the internal representations of the healthy and diseased subpopulations as learned by an ANN trained on the GAW12 data. As the figure shows, the ANN has learned a two-dimensional “feature space” in which the two subpopulations correspond to different (but overlapping) regions; healthy individuals are clustered toward the lower left corner, while sick individuals tend to lie closer to the upper right corner. Further processing of the feature space within the ANN predicts affection status by determining a suitable decision boundary in this space. Because the regions occupied by the two

subpopulations overlap, some prediction error remains. Still, our results show that the ANN performs quite well on the prediction task. Difficulties arise, however, if one intends to extract a descriptive model from the ANN, as it is not easy to determine from an inspection of the trained ANN what combinations of the input attributes contribute most significantly to the network's strong predictive performance.

## 5 Discussion

Our results obtained on the GAW12 data for the pedigree founders of the general population illustrate the use of two very different machine learning / data mining approaches to the analysis of genetic data. Artificial neural networks (ANN) trained using the method of error backpropagation provide accurate predictive models; indeed, our experiments show great reductions in the error rate of ANN relative to that of guessing the majority class for the task of predicting the disease status. However, it is not easy for humans to extract descriptive information from such ANN. In particular, it is difficult to single out significant genomic sites on the basis of the connection weights of an ANN trained to predict the disease status of individuals. For this reason, we have also applied association rule mining (ARM) techniques to the GAW12 dataset. ARM is intended to produce descriptive models that humans can readily interpret. However, a major disadvantage of standard ARM algorithms is that they spend an inordinate amount of time attempting to find all association rules with support and confidence above minimum levels specified *a priori*. We are currently developing ARM algorithms that are better suited to genetic analysis problems such as GAW12 Problem 2. This work builds on previous work by two of the authors of the present study [7, 6].

## 6 Acknowledgements

The authors wish to thank Julia Krushkal, Elizabeth Ryder, Kimberly Ferzoco and other members of the Center for Research in Exploratory Data and Information Analysis (CREDIA) at Worcester Polytechnic Institute for helpful discussions. Support for this research was provided by the Research and Development Council at Worcester Polytechnic Institute.

## References

- [1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 207–216, Washington, D.C., May 1993. ACM.
- [2] C.F. Alex, J.W. Shavlik, and F.R. Blattner. Neural network input representations that produce accurate consensus sequences from dna fragment assemblies. *Bioinformatics*, 15(9), 1999.
- [3] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.
- [4] M.B. Eisen, P.T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95:14863–14868, Dec. 1998.
- [5] P. Hájek, I. Havel, and M. Chytil. The guha method of automatic hypotheses determination. *Computing*, 1:293–308, 1966.
- [6] W.-Y. Lin, S. A. Alvarez, and C. Ruiz. Collaborative recommendation via adaptive association rule mining. In *WebKDD-2000 Workshop on Web Mining for E-Commerce, Sixth International Conference on Knowledge Discovery and Data Mining*, Aug. 2000.
- [7] W.-Y. Lin, C. Ruiz, and S. A. Alvarez. A new adaptive-support algorithm for association rule mining. Technical Report WPI-CS-TR-00-13, Department of Computer Science, Worcester Polytechnic Institute, May 2000.
- [8] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 80–86, New York, August 1998.
- [9] D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by backpropagation. In *Parallel Distributed Processing*, volume I. MIT Press, 1986.