Applications of Neural Network and Genetic Algorithm Data Mining Techniques in Bioinformatics Knowledge Discovery - A Preliminary Study

Richard S. Segall

Arkansas State University
Department of Computer & Information Technology, State University, AR 72467-0130
rsegall@astate.edu

Qingyu Zhang

Arkansas State University
Department of Computer & Information Technology, State University, AR 72467-0130
qzhang@astate.edu

ABSTRACT

This paper presents preliminary research in the area of the applications of modern heuristics and data mining techniques in knowledge discovery. Specifically applications of data mining for neural networks using NeuralWare Predict® software and genetic algorithms using Biodiscovery GeneSight® software were selected for bioscience data sets of continuous numerical valued abalone fish data. Conclusions and future directions of the research are presented.

1.0 Introduction

Large volume of data and complexity in problem solving inspire research in data mining and modern heuristics. Data mining (i.e., knowledge discovery) is the process of automating information discovery. It is the process of analyzing data from different perspectives, summarizing it into useful information, and finding different patterns (e.g., classification, regression, and clustering). Many problems are difficult to be solved analytically in a feasible time. Therefore, researchers are trying to find search techniques or heuristics to get a good enough or satisfactory solution in a reasonable time (Agard and Kusiak, 2004; Ghosh and Nath, 2004; Li and Wang, 2004; Soransen and Janssens, 2003).

This paper is the starting point of performing research that was described in a grant proposal written by Segall and Zhang (2004) and selected for funding as a 2005 Summer Faculty Development Grant by the College of Business of Arkansas State University. The objective of the funded research proposal of Segall and Zhang (2004) was to investigate the applications of modern heuristics and data mining techniques to problems of information systems and technology management. The proposed research of Segall and Zhang (2004) was to both survey the current state-of-the-art of the algorithms in data mining and knowledge discovery and also pursue the application of selected algorithms to databases available on the web as a starting point of additional research. The selected algorithms for this paper are those for neural networks and genetic algorithms.

2.0 Background

This section discusses some recent advances in bioinformatics, microarray data analysis, functional genomics, and genetic and evolutionary algorithms.

2.1 Bioinformatics

Bioinformatics describes any use of computers to handle biological information. Specifically, bioinformatics refers to the use of computers to store, compare, retrieve, analyze or predict the

composition or the structure of biomolecules. Texts in bioinformatics computing have been written by Bergeron (2003), Brown (2000), Jones and Pevzner (2004), and Krawetz and Womble (2003). A series of articles concerning bioinformatics are published in the July 2005 issue of the International Journal on Artificial Intelligence Tools (Bourbakis and Karypis, 2005; Lindlöf et al., 2005). Coppin (2004) presented an entire text on artificial intelligence with individual chapters on fuzzy reasoning, genetic algorithms, intelligent agents, and machine vision.

A text by Krawetz and Womble (2003) contains a collection of separately authored chapters on theoretical and practical approaches to Bioinformatics. These include two chapters by Singh (2003a, 2003b) on statistical modeling of DNA sequences and patterns that include markov chain models and that of statistical and data mining of the matrix attachment regions (MARs) in Genomic sequences. Pevsner (2003) authored an entire text on bioinformatics and functional genomics. Hoppe (2005) discusses the findings of a study on bioinformatics and its potential applications in the risk assessment of complex diseases. Bar-Or et al. (2005) consider the problem of inducing decision trees in a large distributed network of genomic databases. Motivated by the existence of distributed databases in healthcare and in bioinformatics, they present an algorithm that sharply reduces the communication overhead by sending just a fraction of the statistical data.

2.2 Microarray Data Analysis

Microarrays are a new technology to investigate millions of genes simultaneously, which present new statistical problems because the data is very high dimensional with very little replication. Therefore, methods for multiple testing become very important. Microarrays offer an exciting entry point for statistical analysts into areas of bioinformatics. Books that have written that pertain to the exploration and analysis of microarray database using data mining related data mining tools include those by Amaratunga and Cabrera (2004), Baldi and Hatfield (2002), Draghici (2003), Hardiman (2003), Kohane et al. (2003), McLachland et al (2004), Speed (2003), Schena (2003), Stekel (2004), and Wit and McClure (2005).

Kohane et al. (2003) wrote an entire text on microarrays for an integrative genomics that provides a systematic introduction to the use of DNA microarrays an investigative tool, and discusses the foundations for analyzing the foundations for analyzing microarray data sets and genomic data mining. Kohane et al (2003, p.10) introduces in the first chapter of their book why we need new techniques of microarray data analysis that are “not amendable to standard biostatistical techniques” by illustrating graphically a major difference between classic clinical and microarray analysis. This difference is explained by Kohane et al. (2003, p.11) as “the high dimensionality of genome data in contrast to the relatively small number of samples typically obtained results in a highly underdetermined system.” A typical genomic study could include a number of variables from 10,000s to 100,000s with only the number of cases being only in the tens or hundreds, when a typical clinical study would include a number of variables only in the tens or hundreds and the number of cases in the range of thousands to millions. This dimensionality difference is illustrated by Figure 1 of this paper as appeared in Kohane (2003).

2.3 Functional Genomics

As various genome sequencing projects have already been completed, genome researchers are shifting their focus to functional genomics *(*Kuramochi and Karypis, 2005). Kohane et al. (2003, p. 4) describes functional genomics as “the overall enterprise of the deconstruction of the genome to assign biological function to genes, group of genes, and particular gene interactions.” Functional genomics represents the next phase that expands the biological investigation to studying the functionality of genes of a single organism as well as studying and correlating the functionality of genes across many different organisms.

Parmigiani et al. (2003) discusses the methods and software for the analysis of gene expression data and microarrays. Parmigiani et al. (2003) is a book that is a collection of chapters each written by different authors about their statistical software for microarray data analysis. Some of the chapter titles include:

"DRAGON: Methods for the Annotation, Analysis, and Visualization of Large-Scale Gene Expression Data," "SNOMAD: Biologist-Friendly Web Tools for the Standardization and Normalization of Microarray Data", "MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments", and "POE: Statistical Methods for Qualitative Analysis Gene Expression."

Mitchell (1999) wrote an entire text as an introduction to genetic algorithms, and Stekel (2003) an entire text on microarray bioinformatics that includes a detailed section and glossary on MIAME (Minimal Information about a Microarray Experiment) that is intended to assist the exchange the microarray information between researchers.

Claverie and Notredame (2003) authored a book that provides a basic understanding of bioinformatics for those getting started that also includes several chapters labeled as "A Survival Guide to Bioinformatics" as well as several other chapters labeled "Becoming a Specialist: Advanced Bioinformatics Techniques."

2.4 Genetic and Evolutionary Algorithms

Evolutionary computation techniques are commonly used to address problems with large search spaces, where an exhaustive search is not applicable in practice (Giráldez et al., 2005). Due to this fact, many machine learning and optimization tasks have been solved by using genetic algorithms (GAs) or evolutionary algorithms (EAs). Eiben and Smith (2003) wrote an entire text on evolutionary computing that includes an entire chapter each on "What is an Evolutionary Algorithm?", "Genetic Algorithms," "Evolutionary Programming," and "Parameter Control in Evolutionary Algorithms." Other texts that have been recently written that pertain to evolutionary computation and algorithms include those by Bäck et al. (2000a, 2000b), Coello et al. (2002), and Deb and Kalyanmoy (2001). Giráldez et al. (2005) categorize and briefly summarize some of the incorporation of knowledge techniques for evolutionary algorithms and present a novel data structure which helps the evolutionary algorithm to provide decision rules using less computational resources.

Estimation and modelling problems often turn out to be intractable by standard numerical methods. An approach to solving the above problems consists in applying optimization heuristics such as evolutionary algorithms (simulated annealing), neural networks, genetic algorithms, tabu search, hybrid methods, etc., which have been developed over the last two decades. Winker and Gilli (2004) introduce the computational complexity of problems encountered in the fields of statistical modeling and econometrics as well as an overview and classification of the optimization heuristics used. Guan and Zhu (2005) explore an incremental approach to genetic-algorithms-based classification rather than neural networks to cope with learning tasks where the learning environment is ever changing or training samples become available over time.

3.0 Neural Data Modeling

The research presented in this paper is the initiation of research performed for a funded proposal by Segall and Zhang (2004). The first application of data mining techniques presented in this paper is for neural data modeling that utilizes NeuralWorks Predict® software that is manufactured and sold by NeuralWare of Carnegie, PA. According to the website for NeuralWare (2005),

> "NeuralWorks Predict® is an integrated, state-of-the-art tool for rapidly creating and deploying prediction and classification applications. Predict® combines neural network technology with genetic algorithms, statistics, and fuzzy logic to automatically find optimal or near-optimal solutions for a wide range of problems."

The NeuralWorks Predict® software was used on data sets posted on the web by the University of California at Irvine (UCI) Machine Learning Repository (2005) that includes data sets for a variety of

applications ranging from wine recognition database to Challenger USA Space Shuttle O-Ring Databases. The databases selected for the research presented in the paper are databases of biosciences interest both because the Arkansas Biosciences Institute (ABI) funded the purchase of the NeuralWorks Predict® software and also because Bioinformatics computing was investigated in the background section of this paper. Specifically, the databases investigated in this paper are for abalone fish and mushrooms.

3.1 Neural Network Modeling of Numerical Abolone Data

The Abolone data was originally owned by the Marine Research Laboratories and the Department of Primary Industry and Fisheries in Tasmania, Australia. The mushroom data was donated by Jeff Schlimmer as used in his doctoral dissertation research completed at the University of California at Irvine (UCI) in the Department of Information and Computer Science. The mushroom data set was drawn from mushroom records in the Audobon Society Field Guide and is described in terms of physical characteristics as nominal-valued attributes as opposed to numerical values for the Abalone database. Some of the important nominal-valued attributes include classification as either poisonous or edible. The mushroom dataset includes 8124 instances. The Abalone database predicts the age of abalone form physical measurements and includes 4177 instances for eight (8) attributes and has no missing attribute values.

Table 1 of this paper provides for the abalone data set the definition of the eight input variables and the one output variable of the number of rings which is used as the target variable in the neural network algorithm modeling performed by the NeuralWorks Predict® software. Table 2 of this paper provides the statistics for the numeric domains for each of these variables as provided by the Marine Research Laboratories in Tasmania, Australia the original owners of the database.

According to NeuralWorks Predict® Getting Started Guide for Windows written by NeuralWare (2003) and provided with the software purchase, NeuralWorks Predict® builds a network that

> "consists of dividing the input data into train, test and validation sets, and automatically selects input variables that are the most influential in predicting target values. … A genetic algorithm is used to search through the space of combination of input variables. … Predict chooses one of two available learning rules in order to achieve the best results for training the neural network."

Figures 2(a) and 2(b) provide the output generated by NeuralWare Predict® for the complete training which yielded R values of 76.03% for training and 74.49% for testing with accuracies of 97.3% and 96.8% respectively when the variable of number of rings was selected as the output variable. The evaluating model consisting of 4177 records for the abalone data was subdivided into 2923 records (i.e., 70%) for neural network training and 1254 records (i.e., 30%) for testing.

For the prediction phase of the abalone data using NeuralWare Predict® there were five input variables used with the prediction variable of number of rings as indicated by Figures 2(a) and 2(b), and eight variables used in the classification phase. For the classification phase of the abalone data as shown in Figure 3, there were three outputs that included the sex variable that yielded a training accuracy of 55.5% with 2923 records of the 4177 data set, and a testing accuracy of 56.1% for the remaining 1254 records.

For the Abalone data, Figure 4 shows the greatest accuracy was for the second and sixth variables of length and shucked weight respectively, and the least accuracy was for the third and seventh variables of diameter and viscera weight respectively. Figure 5 illustrates a plot of the target and predicted values obtained by NeuralWare Predict® for the target variable of number of rings associated with the abalone fish data set. The presence of many big differences between the curves of Figure 5 for the target values

provides indications that the model did not perform well for these data values, while the presence of small differences between the curves of Figure 5 for data values between approximately 916 and 2000 provide indications that the model worked quite well for these values. Hence the model performed well for some data values of the abalone data set such as these and others, but also did not perform well for many other values. Hence there were distinct intervals where the model performed well or not so well as illustrated by this Figure 5 for the abalone data. In summary, training rules using neural network algorithms was a method of determining those attributes of importance for abalone fish with varying levels of accuracy.

4.0 Genetic Data Modeling

The second application of data mining techniques presented in this paper is for genetic algorithm data modeling that utilizes GeneSight® software that is manufactured and sold by Biodiscovery of El Segundo, CA. According to the documentation link available on the Biodiscovery.com website GeneSight® is "an efficient data mining visualization, and reporting tool that you can use to analyze the massive gene expression data generated by microarray technology." According to the overview link for GeneSight® on Biodiscovery website:

> "GeneSight® allows the researcher to explore large data sets from multiple experimental groups using advanced normalization, visualization, and statistical decision support tools. These tools include GenePie™ visualization, 2-D scatter plots; interactive ratio histogram plotting, hierarchical, K-means, and neural network clustering, principal components analysis, and conditional analyses such as time series. The confidence analyzer tool can use replicated gene expression data for identifying genes having expression patterns which can differentiate between classes of experimental conditions such as disease states."

The GeneSight® software was able to be applied to the abalone fish data set because this data is numerical data, but was not able to be applied to the mushroom data set because all of this data is text data.

4.1 Genetic Modeling of Numerical Abolone Data

Figures 9 thru 18 of this paper provide outputs of GeneSight® software as applied to the abalone fish data. Figure 9 shows the window of GeneSight® with all of the eight input variables and output of NeuralWare predictions.

Figure 10 uses the same window of GeneSight® as Figure 9 but the two input variables of length and height respectively are selected as indicated by the blacked variable tabs for these variables respectively. Using these two selected variables of length and height of the abalone fish clusters were formed with five partitions with dimensions as indicated in the bottom of this GeneSight® window of Figure 10. Cluster5 of the abalone fish data has dimension of 1305, which is the largest cluster. By clicking on each of these clusters, a complete Gene list is enumerated. The bottom of Figure 10 shows the first several genes in the listing of 300 total for cluster1.

Figure 11 presents the Report view of GeneSight® for the abalone data, which includes all of the input, outputs, cluster identification for each of the 4177 instances of gene related data. As checked in boxes on left of this window, plots for both K-Means clustering and scatterplots of the abalone data were requested as described below.

Figure 12 shows the scatterplot generated by GeneSight® for the abalone fish data for the two selected variables of height and length. Figure 12 clearly indicates a strong linear relationship between the height and length measurements of the abalone fish in the sense that as the length increases so does the height. Figure 13 illustrates a histrogram of the frequency distribution of the length measurements of the abalone fish data. Figure 13 illustrates a skewed normal distribution of the length measurements of the abalone

fish with the most frequent length measurement given by the peak of Figure 13 at slightly above 0.5 in this figure.

Figure 14 illustrates the results of the K-Means clustering for the abalone fish data using the Euclidean distance metric for the selected variables of length and height. Figure 14 illustrates the K-Means clustering for the five (5) gene clusters using one experimental condition of the clusters. The shadings illustrated in the bands presented in Figure 14 for the selected conditions of length and height clearly indicate the five (5) clusters as distinct using the Euclidean distance metric.

Figure 15 shows the hierarchical clustering for the three selected variables of length, diameter, and height of abalone fish using the Euclidean distance metric. Figure 15 illustrates on the left of condition1 column an extensive set of branches for the hierarchical clustering created for this selection of variables, distance metric, and cluster linkage of division for the abalone fish data.

Figure 16 shows the PCA plot in 3-D for the abalone data with the five clusters color-coded in different levels of shading on the Axis1. Figure 17 provides the preliminary data for each of the three axes respectively that was used collectively to construct the clusters in the 3-D plot of Figure 16. Figure 18 shows the Box plot for the three selected variables of length, diameter, and height of the abalone fish. The Box plots shown in Figure 18 are in decreasing size and magnitude for each of the variables of length, diameter and height respectively.

5.0 Conclusions and Future Directions of Research

This paper illustrates the useful information that can be obtained using data mining for evolutionary algorithms specifically as those for neural networks and genetic algorithms. The use of NeuralWare Predict® was a very effective method of implementing training rules for neural networks to identify the important attributes of numerical-valued data for abalone fish.

The future directions of the research are data mining of discrete nominal-valued microarray databases for mushrooms and also those items as described more fully in the funded proposal by Segall and Zhang (2004). The latter is to include investigations of modern heuristics of simulated annealing and Tabu search for data mining functionalities of classification and prediction, regression, clustering, and association rules, and continuation of research on modern heuristics of genetic algorithms and neural networks for the data mining functionalities of regression and association rules not discussed in this paper.

6.0 Acknowledgements

The authors wish to acknowledge support of this research from a grant awarded by the Summer Research Grant Committee of the College of Business (CoB) from Arkansas State University for a research proposal submitted. The authors also wish to acknowledge a generous free trail period for use of GeneSight® software that was provided by Biodiscovery Incorporated and also their extremely helpful technical support. The first author also wishes to acknowledge the support provided by the Arkansas Biosciences Institute (ABI) for the purchase of NeuralWare Predict® software, and the technical support that was also generously provided by the staff of NeuralWare and without whose help the corresponding results of this paper would not have been possible.

7.0 References

Agard, B. and Kusiak, A. (2004), Data mining based methodology for the design of product families, International Journal of Production Research, 42(15), 2955-2969.

Amaratunga, D. and Cabrera, J. (2004), Exploration and Analysis of DNA Microarray and Protein Array Data, Wiley-Interscience.

Armstrong, N. and van de Wiel, M. (2004), Microarray data analysis: From hypotheses to conclusions using gene expression data, *Cellular Oncology*, 26(5/6), 279-290.

Bäck, T., Fogel DB, Michalewicz Z., and Beck T. (2000a), Evolutionary Computation 1: Basic Algorithms and Operators, Institute of Physics Publishing, Bristol, UK.

Bäck, T., Fogel DB, Michalewicz Z., and Beck T. (2000b), Evolutionary Computation 2: Advanced Algorithms and Operators, Institute of Physics Publishing, Bristol, UK.

Baldi, P. and Hatfield, GW (2002), DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling, Cambridge University Press, Cambridge, UK.

Bar-Or, A., Keren, D., Schuster, A. and Wolff, R. (2005), Hierarchical Decision Tree Induction in Distributed Genomic Databases, *IEEE Transactions on Knowledge & Data Engineering*, 17(8), 1138-1151.

Bergeron, B. (2003), Bioinformatics Computing, Prentice-Hall, Upper Saddle River, NJ.

Bourbakis, N. and Karypis, G. (2005), Preface in bioinformatics, *International Journal of Artificial Intelligence Tools*, 14(4), 559-560.

Brown, SM (2000), A Biologist's Guide to Biocomputing and the Internet, Eaton Publishing, Natick, MA

Claverie, JM and Notredame, C. (2003) Bioinformatics for Dummies, Wiley Publishing Inc., NY.

Coello, CA, Van Veldhuizen DA, and Lamont, GB (2002), Evolutionary Algorithms for Solving Multi-Objective Problems, Plenum Press, US.

Coppin, B. (2004), Artificial Intelligence Illuminated, Jones and Bartlett Publishers.

Deb, K. and Kalyanmoy D. (2001), Multi-Objective Optimization Using Evolutionary Algorithms, John Wiley & Sons, Inc.

Draghici, S. (2003), Data Analysis Tools for DNA Microarrays, Chapman & Hall/CRC, Boca Raton, FL.

Eiben, AE and Smith, JE (2003), Introduction to Evolutionary Computing, Springer-Verlag, Berlin, Germany.

GeneSight , Bioscovery, El Segundo, CA 90245, www.biodiscovery.com/genesight.asp , viewed July 15, 2005.

Ghosh, A. and Nath B. (2004), "Multi-objective rule mining using genetic algorithms", *Information Science: An International Journal*, 163(1-3), 123-133.

Giráldez, R., Aguilar-Ruiz, J. and Riquelme, J. (2005), Knowledge-Based Fast Evaluation for Evolutionary Learning, *IEEE Transactions on Systems, Man & Cybernetics: Part C*, 35(2), 254-261.

Guan, S. and Zhu, F. (2005), An Incremental Approach to Genetic-Algorithms-Based Classification, *IEEE Transactions on Systems, Man & Cybernetics: Part B*, 35(2), 227-239.

Hardiman, G (2003), Microarrays Methods and Applications: Nuts & Bolts, DNA Press, www.dnapress.net.

Hoppe, C. (2005), Bioinformatics: Computers or clinicians for complex disease risk assessment? *European Journal of Human Genetics*, 13(8), 893-894.

Krawetz, SA and Womble, DD (2003), Introduction to Bioinformatics, Humana Press, Totowa, NJ.

Li, R. and Wang, Z. (2004), Mining classification rules using rough sets and neural networks, *European Journal of Operational Research*, 157, 439-448.

Lindlöf, A., Lubovac, Z. and Michael, H. (2005), Simulations of Simple Artificial Genetic Networks Reveal Features in the Use of Relevance Networks, *Silico Biology*, 5(3), 239-249.

McLachlan, GJ, Do KA, and Ambroise C (2004), Analyzing Microarray Gene Expression Data, Wiley-Interscience.

Mitchell, M. (1999), An Introduction to Genetic Algorithms, The MIT Press, Campbridge, MA.

NeuralWare (2003), NeuralWorks Predict® Getting Started Guide for Windows, Carnegie, PA.

NeuralWare (2005), NeuralWare Predict, http://www.neuralware.com/products.jsp, viewed July 1, 2005.

Parmigiani, G., Garrett, ES, Irizarry, RA, and Zeger, SL, Editors (2003), The Analysis of Gene Expression Data: Methods and Software, Springer-Verlag, Inc., New York.

Pevsner, J (2003), Bioinformatics and Functional Genomics. Wiley-Liss. 551-562.

Schena, M (2003), Microarray Analysis, Wiley-Liss.

Segall, RS and Zhang, Q (2004), Applications of Modern Heuristics and Data Mining Techniques in Knowledge Discovery, Proposal submitted to the Summer Research Grant Committee, College of Business, Arkansas State University, State University.

Singh, GB (2003a), "Statistical Modeling of DNA Sequences and Patterns," Chapter 22 in Krawetz, SA and Womble, DD (2003), Introduction to Bioinformatics, Humana Press, Totowa, NJ.

Singh, GB (2003b), "Statistical Mining of the Matrix Attachment Regions (MARs) in Genomic Sequences, Chapter 23 in Krawetz, SA and Womble, DD (2003), Introduction to Bioinformatics, Humana Press, Totowa, NJ.

Soransen, K. and Janssens, G. (2003), Data mining with genetic algorithms on binary trees, *European Journal of Operational Research*, 151, 253-264.

Speed, T. (2003), Statistical Analysis of Gene Expression Microarray Data, Chapman & Hall/CRC.

Stekel, D (2004), Microarray Bioinformatics, Cambridge University Press, New York.

Winker, P. and Gilli, M. (2004), Applications of optimization heuristics to estimation and modelling problems, *Computational Statistics & Data Analysis*, 47(2), 211-223.

Wit, E and McClure, J (2005), Statistics for Microarrays: Design, Analysis and Inference, John Wiley & Sons.

Figures and Tables available from authors upon request.