An Empirical Comparison of Machine Learning Algorithms for the Classification of Anthracis DNA Using Microarray Data

Michael Doran, Daniela Stan Raicu, Jacob D. Furst, Raffaella Settimi, Matthew Schipma and Darrell Chandler

***Abstract*— The widespread use of microarrays to discover information about DNA samples has produced a large amount of data as well as the associated challenges associated with volumes of data. The goal of the microarray evaluated in this study is to provide a unique fingerprint for a diverse collection of genomic samples. A wide variety of machine learning algorithms are available for prediction and classification, so the question is which ones are most appropriate? A collection of algorithms are empirically compared based on their ability to discriminate between 26 strains of *Bacillus anthracis* using a general purpose microarray. *B. anthracis* strains are known to have very little variation between strains and can be used as a gold standard for evaluating the capabilities of the microarray and analysis approaches. Support vector machines are found perform significantly better than other approaches and offer several characteristics that make them an attractive solution.**

***Index Terms*—Microarray, Machine Learning**

I. Introduction

MICROARRAYS provide a tooling for conducting hundreds or thousands of parallel tests on a DNA sample to learn about its genetic composition. This capability makes them invaluable amounts of biological insight while simultaneously causing analytical complications by providing orders of magnitude more parallel tests than test cases, which is commonly referred to as the curse of dimensionality. A study on the variability of microarrays shows that there can be substantially different results between replicates of the same test [5]. While this study recommends always using at least three replicates so that it is possible to quantify the variance between trials, this is still a very small number of cases compared to the number of tests that can be performed by each microarray. Each probe on a microarray represents a single test, so an experiment with over 7,000 probes, such as [3] conducts over 7,000 tests in parallel. Prior work based on a similar data set used a statistical analysis of variance with a correction for false discovery [10]. An alternate approach commonly used in data mining is to use a machine learning algorithm and estimate the error in the model based on empirical test results. Some machine learning approaches may have advantages or disadvantages compared to each other and traditional statistical tests. This paper describes some of the relative merits and compares them based on their ability to classify bacillus strains in a test data set. The molecular differences between the genomic compositions of the *B. anthracis* strains are expected to be one of the hardest possible test cases for these techniques. Performance results from this set should generalize well.

The data set is derived from an experimental oligonucleotide microarray with 390 randomly selected 9-mer probes. The probes are not long enough to target highly distinguishing sequences but the collection of hybridization levels may be enough to uniquely identify a diverse set of classes. The ratio of experiment replicates to tests performed is exceptionally low. There are only 9 samples of each isolate while there are over 43 times as many attributes (probes) for each sample. The intent of the microarray design is to develop a single test that can be used to create fingerprints that new samples can be compared to.

The collection of machine learning algorithms used includes Naïve Bayesian, Bayesian Belief Networks, C4.5 decision trees, k-Nearest Neighbor and Support Vector Machines. A typical assumption in data mining applications is that there is an abundance of data. This is very much not the case with microarray data. One way to reduce the impact of the case to attribute ratio is to use only a subset of the available which contain most of the information. The information gain criterion is used to reduce the data dimensionality and compare the relative impact this has on the various classification models. Some algorithms may benefit from having more or fewer attributes. A single criterion is sufficient to briefly compare the algorithms and there is a strong argument for not using feature selection in practice so arriving at the precise balance is not beneficial, so it is not the focus of our paper. If a model is applied to a larger more general data set then any attribute may be important for discriminating between samples. To scale well the models would ideally be able to perform well using the complete feature space.

The microarray image data is processed using the AMIA

Updated 7/30/2006.

Michael Doran, Daniela Stan Raicu, Jacob D. Furst and Raffaella Settimi are with the Intelligent Multimedia Processing Laboratory in the School of Computer Science, Telecommunications and Information Systems at DePaul University, Chicago, IL 60604.

Matthew Schipma and Darrell Chandler are with the Biochip Technology Center at Argonne National Laboratory, Argonne, IL 60439.

tool kit [9] to extract intensity data. The log ration of the forground to background intensity for each spot is used as the intensity for the spot. Quantile normalization is applied to the entire data set to make intensity values comparable between slides [1].

II. Feature Selection

Feature selection is considered as a method of mitigating the analytical problems introduced by the attribute to class ratio. One method for selecting attributes, or probes in the language of microarrays, is to use a rule of thumb metric such as a signal to noise ratio or to look for intensity values that vary between isolates several times the standard deviation of the normal range as in [12]. To avoid selecting an arbitrary threshold the information gain metric can be used to quantify how well each attribute can separate the set by class (isolate) as measured by entropy. The relative entropy before and after the data set is split is compared to calculate the information gain of an attribute as outlined in (1).

$$Entropy(S) = \sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

$$Information\ Gain(S, A) = Entropy(s) - \sum \frac{|S_v|}{|S|} Entropy(S_v)$$

S is the set of data containing all of the classes. n is the number of classes in the set. p_i is the probability that class i is in the subset. S_v is a subset of S that can be created by splitting the set using the attribute being evaluated.

The ideal machine learning algorithm will function well without feature selection, but comparing results using the full and reduced feature sets will provide insight into how much they are affected by the use of all dimensions. Only 67 of the probes had an information gain above 0 as shown in Figure 1. For this study all algorithms are evaluated using the entire set of attributes and a reduced set consisting of these 67 attributes.

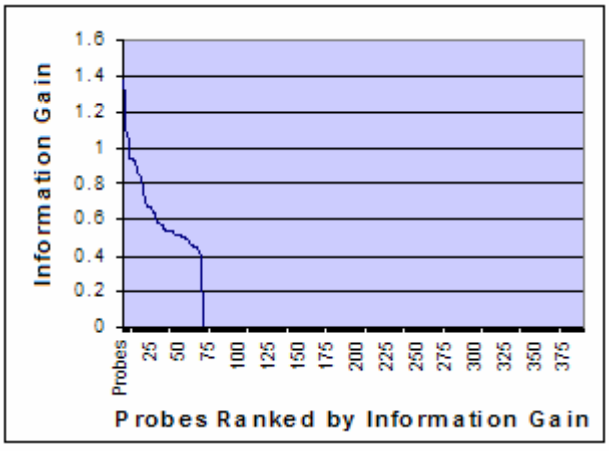

Figure 1: Relative Information Gain per Probe

III. Machine Learning Algorihtms Overview

A complete discussion of each of machine learning algorithms is outside of the scope of this paper, but a brief note is included about each. Specifically, strengths and weaknesses of each are compared in the context of the problem domain.

A. *Naïve Bayesian*

Naïve Bayesian is the simplest implementation of Bayes' rule and assumes that all attributes are independent. A downfall of the Bayesian approach is that if there are no examples of an attribute having a particular value then it will determine that there is a 0% chance of a new example having that value. With the possibility of outliers and the few samples available this may be a problematic assumption. The assumption of independence may be valid for this microarray given the small size of probes. In practice Naïve Bayes can perform well even when this assumption is violated [7]. Another strength of the Naïve Bayesian approach is that it produces a probability for each classification which could be used to give a confidence level to each prediction. Just as importantly, a low confidence level could indicate that there is not sufficient evidence to make a prediction. The classifier predicts that a new case x consisting of attributes $\{x_1,x_2,\ldots,x_l\}$ is a member of class ω based on a set of M classes using formula (2).

$$\omega_m = \arg\max_{\omega_i} \prod_{j=1}^{l} p(x_j \mid \omega_i), i = 1,2,...,M \qquad (2)$$

B. *Bayesian Belief Networks*

The potentially problematic assumption of attribute independence of Naïve Bayesian is solved with Bayesian Belief Networks. Acyclic digraphs are created using the K2 algorithm [4] that compute conditional probabilities for class membership. If the independence assumption is not valid then this may perform better than Naïve Bayesian while still providing the advantage of assigning a probability to each prediction.

C. *C4.5 Decision Trees*

Decision trees provide easily interpreted classification rules [6]. The clarity of the rules makes the approach favorable. At each node in the decision tree a specific probe determines how a sample is classified. Figure 2 shows an example of a decision tree where a sample where probe 1 and 2 both have intensity values greater than 2 is classified as class B. This simplicity may also be the decision tree's primary weakness if the data is noisy. If there are not enough samples to provide examples of enough errors then a single incorrect reading in a new sample may be classified incorrectly.

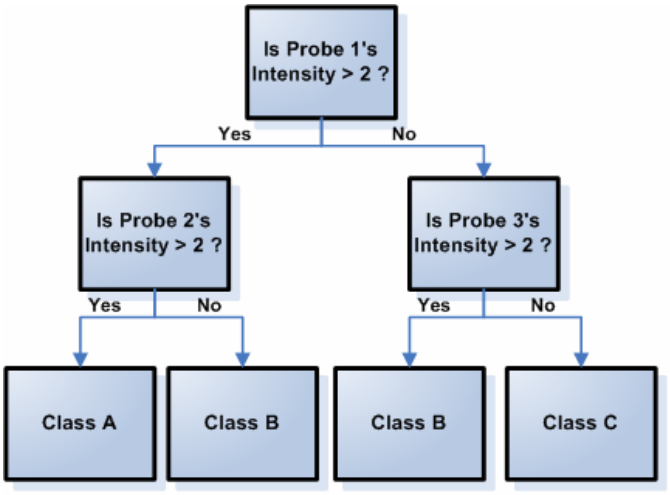

Figure 2: Example Classification Decision Tree

D. *k-Nearest Neighbor*

k-NN is described as instance based learning. Instead of extracting rules from training data, new samples are compared to existing samples using a distance matrix [11]. The majority class of the closest *k* neighbors is used to the predicted class of the new case. Two components of applying k-NN are the distance metric used and the value of k. k-NN can be computationally expensive since new case is compared to every case in the training set. For this comparison the Euclidean distance is used and values of 1 and 3 are used for k.

E. *Support Vector Machines*

An optimal margin classifier called SVMs (Support Vector Machines) [8] provide a statistical learning approach. SVMs are known to perform well on sparse data sets [8] and microarray data sets in particular [2]. While this approach is expected to perform well, a disadvantage is that the model is difficult to interpret. The SVM attempts to locate the optimal hyper plane to separate the classes. With more than 2 dimensions this model is impossible to display graphically. The idea of margins is easy to interpret with microarrays as the amount of space that exists between hybridization levels. This does not assume normality of the data. SVMs also use a complexity parameter that allows it to remove some noise training data to decrease the influence of outliers on the model. The model does not assume a Gaussian normal distribution so heteroscedacity does not impact the correctness of the model. This could be an issue with ANOVA based models when significance levels may appear elevated if there is lower variance in a specific attribute.

IV. Model Evaluation

If an abundance of data were available the preferred way to evaluate classification models is to use two independent data sets. A training set is used to construct the model. The model is applied to a separate test set. The predicted classes for the test set are compared to the known classes to estimate the positive predictive power of the model. In the data set studied here, and in many microarray experiments in general, there are very few cases available. To cope with the shortage *m*-fold cross validation is used to evaluate each model. Briefly, the data is divided into *k* segments with an equal number of cases from each class in each fold. The testing process is repeated *m* times. In each iteration, the m^{th} fold is used as the testing data and all other folds are combined to form the training set. The average number of correct predictions from each of iteration is averaged together to form an overall estimate of the model performance. A *k* value of 9 is used since that is the number of cases available for each class, or bacillus strain.

V. Results

The comparative results for each classifier are shown in table 1. The decision tree proved to have the lowest performance on this particular data set, probably because of how susceptible it is to a single outlier. The overall number of attributes available did not improve the classifier, probably because it focused only on attributes that had the highest relative information gain. *k*-NN, with *k*=1, performed relatively well with a reduced set of attributes, but classified fewer than half of the cases correctly using the full data set. Of the two approaches that used Bayes' rule, the Bayesian Belief Network is probably preferable. Using the reduced data set

Classifier	Probes Used	Correctly Classified
Naïve Bayesian	390	49.4%
Naïve Bayesian	67	58.7%
Bayesian Belief Network	390	53.6%
Bayesian Belief Network	67	54.5%
C4.5 Tree	390	33.9%
C4.5 Tree	67	33%
k-NN 1	390	49.4%
k-NN 1	67	60.5%
k-NN 3	390	40.8%
k-NN 3	67	56.8
Support Vector Machine	390	78.1%
Support Vector Machine	67	69.9%

Table 1: Classification Model Comparison

Naïve Bayesian performed slightly better, but only about 5% better. The Bayesian Belief Network offered similar performance using the complete data set which may make it

more robust when applied to a data set with a greater variety of DNA samples. The SVM correctly classified the greatest number of cases and interestingly did not benefit from reducing the number of attributes. SVMs weight attributes according to the margin between classes and are know to perform well on sparse data sets, so it is not surprising that the model is able to perform well with the full feature set [8].

A. *Significance of the Classification Results Variation*

To evaluate the significance of the positive predictive power of the SVM compared to the Bayesian Belief Network, the cross validation results from each fold are considered matched pairs of results. This approach follows a generally accepted method for comparing data mining methods as described by [11]. The distance *d* between the positive predictive power of each model is computed as the difference between the results for each fold. The t statistic in equation (2) is computed to see if there is a significant difference between d and 0. A value of 0 indicates that the models performed similarly. Table 2 shows the paired results for each fold.

Fold	SVM	Bayesian Belief Network	d
1	84.6%	53.8%	30.8
2	80.8%	42.3%	38.5
3	76.9%	57.7%	19.2
4	80.8%	57.7%	23.1
5	69.2%	46.2%	23.1
6	80.8%	61.5%	19.2
7	76.9%	69.2%	7.7
8	73.1%	50.0%	23.1
9	80.0%	56.0%	24.0

Table 2: Model Performance by Fold

$$k = number\ of\ \ folds$$
$$d_i = d\ \ for\ \ fold\ i$$
$$\bar{d} = \frac{1}{k}\sum_{i=1}^{k} d_i \qquad (2)$$
$$t = \frac{\bar{d}}{\sqrt{\sigma_d^2 / k}}$$

The t-test is used to evaluate the following null and alternate hypothesis:

H_0: The SVM and Bayesian Belief Network models performed similarly, so $d = 0$

H_A: The SVM performed significantly better than the Bayesian Belief Network model, so *d* is probably 0 with a high level of statistical significance.

Based on this evaluation, the t-statistic is a very high 8.29 and we can reject H_0 in favor of H_A. This means that the SVM model has a greater positive predictive power at a very high level of statistical significance.

B. *Significance of the Classification Results Variation using Feature Selection*

After reducing the number of attributes there is still a compelling case for choosing support vector machines. The t statistic when comparing the SVM to the k-NN algorithm for k =1 is 2.878, so we can conclude that we are 98% sure that the SVM will perform better than k-NN. The Bayesian Belief Network performed significantly lower than the k-NN model with a t statistic of 2.495.

C. *Misclassifications*

With each algorithm certain classes proved to be challenging to classify. Pairs of isolates commonly confused include 16 and 19, and to a lesser extent 23 and 24 along with 12 and 13. REP-PCR was used to verify that sufficient DNA material was available to conduct the tests. Interestingly, each confused pair had the same number of visible PCR bands. While it generally does not have sufficient discriminatory power to distinguish between *B. anthracis strains*, REP-PCT is often used as one of the steps in the comparison of strains. The consistency suggests that there is a positive correlation between strains that seem similar based on our microarray analysis and the more widely used PCR-Rep approach.

VI. CONCLUSION

Based on a collection of *B. anthracis* samples applied to an experimental microarray of untargeted oligonucleotide probes, the support vector machine approach provides a significant advantage over the other machine learning algorithms. Optimal results are found using the entire data set so minimal consideration is given to attribute selection. The *B. anthracis* data set is considered to be biologically hard because of the similarity between strains so these results are probably applicable to a wider set of microarray classification problems.

REFERENCES

[1] Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance*. Bioinformatics 19(2):185-193

[2] Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M. and Haussler, D., (2000) Knowledge-based analysis of microarray gene expressions data by using support vector machines, *National Academy of Sciences* 97(1),262-267.

[3] Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*. Science 15 October 1999 286: 531-537.

[4] Heckerman, D., D. Geiger, and D.M. Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20(3):197-243.

[5] Lee MLT, Kuo FC, Whitmore GA, Sklar J, *Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations*. Proceedings of the National Academy of Sciences USA 97: 9834-9839 (2000).

[6] Quinlan, J.R. (1996) C4.5: Programs for Machine Learning. Morgan Kaufman, San Mateo, CA.

[7] Theodoridis and Koutroumbas (1999). Pattern Recognition. Academic Press, New York.

[8] Vapnik, V. (1998). Statistical Learning Theory. John Wiley & Sons, Inc., New York.

[9] White, Daly, Willse, Protic and Chandler, (2005) Automated Microarray Image Analysis Toolbox for MATLAB. Bioinformatics. 21: 3578-3579

[10] Willse, Alan, Chandler , Darrell P., White , Amanda, Protic, Miroslava, Daly, Don S., and Wunschel, Sharon (2005) "Comparing Bacterial DNA Microarray Fingerprints", *Statistical Applications in Genetics and Molecular Biology*: Vol. 4: No. 1, Article 19. http://www.bepress.com/sagmb/vol4/iss1/art19

[11] Witten, I.H., and E. Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco

[12] Yu, Liang, Maximilian Diehn, Nathan Watson, Andrew W. Bollen, Ken D. Aldape, M. Kelly Nicholas, Kathleen R. Lamborn, Mitchel S. Berger, David Botstein, Patrick O. Brown, and Mark A. Israel. *Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme*. PNAS, Apr 2005; 102: 5814 - 5819.