Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability

Charles Edeki

Mercy College, Mathematics and Computer Science Department,
Broadway, Dobbs Ferry, NY, USA

Shardul Pandya

School of Business and Technology,
Capella University, Minneapolis, Minnesota, USA

Doi:10.5901/mjss.2012.v3n14p49

Abstract

Huge efforts are being made by computer scientists and statisticians to design and implement algorithms and techniques for efficient storage, management, processing, and analysis of biological databases. The data mining and statistical learning techniques are commonly used to discover consistent and useful patterns in a biological dataset. These techniques are used in a computational biology and bioinformatics fields. Computational biology and bioinformatics seeks to solve biological problems by combining aspects of biology, computer science, mathematics, and other disciplines (Adams, Matheson & Pruim, 2008). The main focus of this study was to expand understanding of how biologists, medical practitioners and scientists would benefit from data mining and statistical learning techniques in prediction of breast cancer survivability and prognosis using R statistical computing tool and Weka machine learning tool (freely available open source software applications). Six data mining and statistical learning techniques were applied to breast cancer datasets for survival analysis. The results were mixed as to which algorithm is the most optimal model, and it appeared that the performance of each algorithm depends on the size, high dimensionality of data representation and cleanliness of the dataset.

Keywords*: Data Mining, WEKA, R tool, Computational Biology, Bioinformatics*

Introduction

The advancement of medicine now relies upon the collection, management, storage, and analysis of large biological datasets. Data mining, statistical and machine learning techniques are the process by which new knowledge is extracted from a dataset. According to Mitchell (1997), the definition of machine learning is as follows: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measure by P, improves with experience E" (p. 2). Data mining, statistical and machine learning are based on inductive inference, a process of observing a phenomenon, then building a model for that phenomenon and making predictions using the model.

In this study, the results of a comprehensive comparative study of the following data mining, statistical and machine learning algorithms was examined:, Support Vector Machines (SVM);, RandomForest;, AdaBoost, Bagging;, Boosting;, Decision Trees and Artificial Neural Networks (ANN) classifiers algorithms. The main focus of this research was to study the effective classification

learning techniques for prediction of breast cancer survivability. In other words, can one algorithm or techniques be more effective at predicting survivability over others.

There are two main aspects in prediction of cancer survivability: accuracy (how true is the algorithm's prediction), and efficiency (how fast can the algorithm execute the prediction task). Data reduction technique was applied to the dataset and obtained a reduced representation of the breast cancer dataset. The resulting data set was much smaller in volume, yet closely maintained the originality of the data (Han, & Kamber, 2006). The R PCA function was used to reduce the large dataset (the patients in this case) to smaller components of objects related according to their expression patterns with tumor size.

Classification algorithms are the most common data mining and machine learning algorithms, often used for data analysis in both industry and academia. Classification is a supervised learning algorithm used to map a dataset into predefined groups or classes. The biological datasets from the National Cancer Institute (NCI) biological database system was used to find the prediction rate of each algorithm and comparative studies of the algorithms were performed in order to find the optimal classification model.

R and Weka software were used to analyze the breast cancer dataset. R is open source statistical analysis software (R Development Core Team, 2010), and Weka is open source machine learning application software that can be used to normalize and analyze datasets.

Methods

The exponential growth of the amount of biological data available raises two problems: on one hand, efficient information storage and management, and on the other hand, the extraction of useful information from these data. The second problem is one of the main challenges in computational biology, which requires the development of an effective computational analysis tool and is the problem that was presented in this study.

For many studies in medicine, researchers are interested in assessing the time it takes for an event to happen. Very often, the event is an outcome, such as diagnosis or death, but the outcome may also be other measurable parameters, such as onset of disease or relapse of disease. There is a term that describes the period leading to the event, called survival time. Furthermore, survival analysis is the term used to describe the investigation into the patterns of these events that occur within one or more cohorts in a study (Thongkam et al. 2007). In dealing with the analysis of survival data, researchers are interested in the length of time it takes a patient to reach an event rather than simply the fact that the event has or has not occurred.

There are at least two ways to motivate why particular data mining and statistical learning techniques were suitable for a particular learning task (Joachims, 2001). One way was through comparative studies and the other was through benchmarking (Joachims, 2001). This research study was based on comparative study of data mining and statistical learning techniques. Each of the data mining and statistical learning techniques is briefly discussed below.

Support Vector Machine (SVM) was mainly developed by Vladimir Vapnik and is based on the structural risk minimization principle from statistical learning theory. SVM algorithm uses a nonlinear mapping to transform original training data into higher dimensions. Then SVM searches for the linear optimal separating hyperplane within the new dimension. The hyperplane is the decision boundary separating the datasets of one class from another. The SVM finds this decision boundary using training sets or support vectors, and margins defined by the support vectors. SVM is very accurate due to its ability to model complex nonlinear decision boundaries and is, less prone

to overfitting problem, but according to Han & Kamber (2006), SVM is very slow when compared with other classification algorithms (Vapnik, 1998, Han & Kamber, 2006).

The decision tree algorithm is the most popular algorithm in data mining classification technique because it is easy to understand how it makes predictions. There are many decision tree algorithms for constructing a decision tree, such as ID3, C4.5, SLIQ, Scalable Parallelizable Induction of Decision Tree (SPRINT), etc. There are two phases in generating or creating a decision tree, namely the tree-growing phase and tree-pruning phase. In the tree-growing phase the algorithm starts with the whole data set at the root node. The data set is partitioned according to a splitting criterion into subsets. This procedure is repeated recursively for each subset until each subset contains only members belonging to the same class or is sufficiently small. In the tree-pruning phase, the decision tree is reduced in order to improve time complexity and prevent overfitting (Kleissner, 1998, Sattler & Dunemann, 2001).

AdaBoost is one of the most powerful learning ideas introduced in the past twenty years. It was originally designed for classification problems, but has been extended to regression as well (Hastie, Tibshirani, & Friedman, 2001). AdaBoost is a popular ensemble method and has been shown to significantly enhance the prediction accuracy of the base learner (Thongkam, Xu, Zhang, & Huang, 2007). It is a learning algorithm used to generate multiple classifiers and to utilize them to build the best classifier (Schapire & Singer, 1999). The process of boosting is to combine the outputs of many weak classifiers to produce a powerful classifier. The predictions from the weak classifiers are then combined through a weighted majority vote to produce the final prediction (Hastie, Tibshirani, & Friedman, 2001). The advantage of this algorithm is that it requires less input parameters and needs little prior knowledge about the weak learner (Thongkam, Xu, Zhang, & Huang, 2007).

The study of artificial neural networks (ANN) was inspired by attempts at mimicking the brain functionality (Tan, Steinbach, & Kumar, 2006). Neural networks represent an alternative computational paradigm, which has received much attention in the past few decades (Hertz, Krogh & Palmer, 1991). Neural networks are capable of predicting new classes based on past examples after executing a process of learning. There are two phases in the processes of training the artificial neural network: learning and recalling. Networks are trained by inputting a training dataset with the target data. Weights are adjusted until the outputs reach the desired training outputs. The goal is to minimize the error, which is the difference between the target output and desired output. After learning, the testing dataset would be applied to the artificial neural network to estimate the desired output and determine the performance of learning.

The general approach that was used for predictive model building in this research is as follows:

1. Create training and testing datasets.
2. Apply a data mining/statistical learning technique to the training set.
3. Generate the predictive model.
4. Evaluate model using testing dataset.
5. Repeat step# 2 with other techniques.
6. Compare performance between techniques.

The breast cancer dataset consists of five categories of patient data, as shown in Table 1, that exist for more than 62,000 breast cancer patients diagnosed in the United States between 1990 and 1997. Thus, all files contain variable data for the same group of patients. The dataset originated from The Surveillance, Epidemiology, and End Results (SEER) Program of the NCI. Most of the data, including pathology, diagnosis, and treatment, are real and excellent biomedical dataset. The demographic data, however, was partially artificial due to patient's privacy, as the original dataset

from SEER is completely anonymous. This identifier acts like a hospital record number of a patient but is purely fictitious, as the original data is anonymous. Variables for the complete patient dataset are shown in Table 1.

Table 1. *Patient Dataset Variables*

Table Name	Attribute Name	Attribute Description
Demographic data	patientid	unique patient identifier (artificial)
	dateofbirth	patient date of birth (artificial)
	maritalstatus	marital status at diagnosis
	race	patient ethnicity
	ageatdiagnosis	age at diagnosis
	alivestatus	patient alive or dead
	survivaltime	survival time from date of diagnosis
Diagnosis data	patientid	
	yearofdiagnosis	year of diagnosis
	histology	histologic type of tumor
	primarysite	site of primary tumor
	numberofprimaries	number of primary tumor
Pathology data	patientid	
	Grade	tumor grade
	Nodesexam	number of lymph nodes examined
	Nodespos	number of positive lymph nodes
	Extent	extent of disease
	Nodalstatus	status of lymph node involvement
	Size	size of tumor
	Pgr	progesterone receptor status
	Er	estrogen receptor status
Staging	patientid	
	Stage	stage of tumor
Treatment	patientid	
	Surgery	surgery regime received
	Radiotherapy	radiotherapy received

There are a number of methods that can be used to transform data variables into forms that are usable by data mining algorithms. The Weka data-mining tool was used for the preparation of the breast cancer datasets for mining.

The PCA data reduction method (prcomp( ) function) in R statistical program was used to reduce the dataset. PCA is a statistical method routinely used to analyze interrelationships within a large set of data, revealing common underlying factors or components. PCA examines the correlations between the original data values and condenses the information contained within objects into smaller group of components with minimal loss of information.

According to Thongkham et al. (2007), stratified 10-fold cross-validation is a common validation method used to minimize bias and variance associated with random sampling of the training and test datasets. Also, it is a common method for data selection in machine learning related to medical and biological research. The stratified 10-fold cross-validation process was used

in this study in evaluating and validating the predictive model. The process consists of four steps as follow (Thongkam et al. 2007):

1. Divide the dataset into a set of subclasses.
2. Assign a new sequence number to each set of subclasses.
3. Randomly partition the subclass into 10 subsets or folds.
4. Combine each fold of each subclass into a single fold.

The Weka data mining tools support automatic splitting of a data set into training and test sets using either a straight percentage splits or through k-fold cross validation.

Results

This section discusses analysis of the breast cancer dataset by various methods.

Analysis was begun by performing logistic regression on the complete 10-year survival dataset. The *summary( )* function was used and *length* on the *alivestatus* factor to determine the number of rows for each outcome, as well as the total number of patients as shown in table 2.

Table 2. *Total Number of Patients*

Total Number of Rows in the Dataset		
0	Number of live patients	11,714
1	Number of dead patients	3,480
	Total number of patients	15,194

The number of patients alive after 10 years (row 0) is more than three times the number of patients that have died (row 1). To create a logistic regression model, glm( ) function is called, which provides a model that is an equation to predict whether a patient will survive 10 years. To evaluate the predictive ability of the model, we used the predict( ) function to predict the probability of outcome for all cases in the dataset. The classification result of the logistic regression was 12,080 (11,301 + 779) correct predictions (true positive and true negative), and 3,114 (2,697 + 417) incorrect predictions, resulting in the overall accuracy of 79.5% (12,080/15,194). The precision was 80.7% (11,301/13,998). The recall was 96.4% (11,301/(11,301+417)).

Logistics Regression with Holdout: We repeated the logistic regression approach using the holdout method that contained lesser dataset to evaluate the model; the result was 1,482 (816 + 666) correct predictions (true positive and true negative), and 604 (392 +212) incorrect predictions, resulting in the overall accuracy of 71% (1,482/2,086). The precision was 67.5% (816/(816+392)). The recall was 79.4% (816/(816+212)).

Decision Tree Algorithm: The Weka's J48 decision tree learner, based on C4.5 decision tree algorithm was used with default parameter setting to build a decision tree model for a 10-year survival dataset. The function was called J48 and is already implemented in RWeka. The precision for the model is 79.9% (2,485/(2,485+631)). The decision tree model was evaluated using the 10-fold cross-validation.

The multilayer perceptron learner algorithm in Weka with default parameter settings was modified such that it could serve as a Neural Network. The hidden layers parameter was set to one hidden layer with five nodes to build the artificial neural network model for a 10-year survival dataset. The function is called multilayerPerceptron and is already implemented in RWeka. The model was evaluated using 10-fold cross-validation and the original train.full_1 dataset was used to

build the model. The results was 72.94% accuracy in classification. The correct prediction was 4926/(4926+1827), which was 72.94% and incorrect prediction was 1827/(4926+1827), which was 27.1%. The kappa statistics was 0.523.

The next modeling approach was a support vector machine (SVM). The SVM algorithm implemented in Weka is called SMO (sequential minimal optimization). A significant factor in the SVM model-building process is parameter adjustment. The SVM model was generated using RWeka's built-in function, SMO( ). Ten-fold cross validation of the SVM model was performed and the model was evaluated using the 200-instance test set.

The SVM model accuracy result on the full dataset was 68.4%, the correct prediction was 4620/(4620+2133), and incorrect prediction was 2133/(4620+2133), which was 31.6%. The kappa statistics was 0.3683 and the ROC area was 0.684.

We applied boosting to the breast cancer dataset using J48 decision tree as our model-building algorithm. To implement AdaBoost.M1, we called the AdaBoostM1( ) function and set the classifier algorithm parameter (W) to "J48" using Weka_control( ). We evaluated the model by performing 10-fold cross-validation; the boosted model is then evaluated on the small test set. The boosting model accuracy result on the full dataset was 69.5%, the correct prediction was 4694/(4694+2059) and incorrect prediction was 2059/(4694+2059), which was 30.5%. The kappa statistics was 0.3902 and the ROC area was 0.759. The boosting model accuracy result on the 200_test data was 73%. We applied bagging to the breast cancer dataset using the J48 decision tree. The bagging( ) function in Weka was called and set the classifier algorithm parameter (W) to "J48". The model was evaluated by performing 10-fold cross-validation, the bagged model was evaluated on the small test set (200 instances).

The bagging model accuracy result on the full dataset was 68.84%, the correct prediction was 4649/(4649+2104), which was 68.84% and incorrect prediction was 2104/(4649+2104), which was 31.16%.

The RandomForest model was built using Weka's RandomForest ( ) function, which is based on the same concept as the original Random Forest algorithm developed by Breiman (Breiman, 2001). Like boosting and bagging, the Random Forest model was created using the Weka's RandomForest( ) classifier and evaluated the model by performing 10-fold cross-validation. Using Weka_control() function, the RandomForest( ) function created 1,000 trees by setting the parameter I to 1000.

The Random Forest model accuracy result on the full dataset was 75%, the correct prediction was 5064/(5064+1689), which was 74.99% and incorrect prediction was 1689/(5064+1689), which was 20.01%.

The summary of the prediction results of the data mining and statistical learning algorithms are shown in Table 3. The SVM classifier is the only algorithm that did not improve when applied to the independent dataset with 200 records. The rest of the algorithms showed slight improvement when applied to the independent dataset.

Table 3. *Prediction Results of the Algorithms*

Type	Overall Accuracy – Full Dataset	Overall Accuracy – 200 Independent dataset	Precision – full dataset	Precision – 200 Independent dataset
Logistics Regression	71%	72.5%	67.5%	68.3%
Decision Tree – J48	70.17%	71.5%	71.7%	74.2%
ANN MultilayerPerceptron( ) function	72.94%	73.04%	74%	74.7%

Support Vector Machine (SVM) using Weka's Sequential Minimal Optimization (SMO)	68.414%	66.5%	69.7%	69.4%
Boosting- AdaBoostM1	69.5%	73%	70.2%	71.7%
Bagging - Weka's Bagging( ) function	68.84%	72%	67.3%	71.6%
Random Forest - Weka's RandomForest function	75%	76.6%	72%	73.1%

Discussion

The prediction of cancer survivability has been a major issue in medicine and biology. In this study, we have explored six different statistical and machine learning methods for generating predictive models for datasets with either binary or continuous response variables. It is critical that one does not apply classification or regression methods to datasets without having confidence that the methods are indeed suitable for data.

For the binary outcome survival status dataset, we generated six models from diverse statistical learning and data mining techniques. This was useful because it gave us a choice of models and indicated which model is superior by assessing the accuracy and precision. From the accuracy perspective, the best model is RandomForest (75.0%). We did, however, express concern about cost of predicting patients to survive 10 years but who actually die (False–Negative). If this is more important than overall accuracy or precision, our best model is produced by bagging (26.5% error) and the worst is the decision tree (33.3% error). The second best error rate for false-positive is Random Forest (30% error). Clearly there is much to think about even after we have generated the models, from this study, we can say the result of each model depends on the quality of the biological dataset, the size of the dataset and the representation of the dataset.

Conclusion

Medical institutions looking to undertake a data mining approach to solve biological problems could be well-served by including statistical learning and data mining processes in their analytical and intervention efforts. Computer scientists, medical researchers and statisticians need to look at their own biological data availability for variables that might potentially link to prediction of cancer survivability. The selection of variables in this study was based on computational biology and bioinformatics literatures, breast cancer dataset available and domain knowledge of the researcher.

Data preparation (data quality) could be the difference between a successful machine learning project and a failure and takes between 60 – 80% of the whole data mining or machine learning effort or process (Witten & Frank, 2005).

Findings indicate that none of the data mining and statistical learning algorithms applied to the breast cancer dataset outperformed the others in such a way that it could be declared the optimal algorithm. Additionally, none of the algorithm performed poorly as to be eliminated from future prediction model in breast cancer survivability tasks.

References

Adams, J., Matheson, S., & Pruim, R. (2008). BLASTED: Integrating biology and computation, Journal of Computing Sciences. 24(1), 47-54.

Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufman.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York, NY: Springer.

Hertz, J., Krogh, A., & Palmer, R. (1991). Introduction to the Theory of Neural Computation. New York, NY: Addison-Wesley.

Kleissner, C. (1998). Data mining for the enterprise. 1060-3425/98 IEEE. Retrieved from the Institute of Electrical and Electronics Engineers (IEEE) Digital Library.

Joachims, T. (2001). A Statistical learning model of text classification for support vector machines. SIGIR, New Orleans, LA. Retrieved from ACM Digital Library.

Mitchell, T. (1997). Machine Learning. San Francisco, CA: McGraw-Hill.

Sattler, K., & Dunemann, O. (2001). SQL database primitives for decision tree classifiers. Proceedings of the 2001 CIKM Conference. Atlanta, Georgia. Retrieved from ACM Digital Library.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. Journal of Machine learning, 37(3), 297-336.

Sriraam, N., Natasha, V., & Kaur, H. (2006). Data mining approaches for kidney dialysis treatment. Journal of mechanics in medicine and biology. 6(2), 109-121.

Tan, P., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. Boston, MA: Addison Wesley.

Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2007). Breast cancer survivability via AdaBooost algorithms. Australian workshop on health data and knowledge management, Wollongong, NSW, Australia. Retrieved October 9, 2010, from the Association for Computing Machinery (ACM) Digital Library.

Vapnik, V. N. (1998). Statistical Learning Theory. Chichesrter, GB: Wiley.

Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. San Francisco, CA: Morgan Kaufma