IUG Journal of Natural and Engineering Studies
Vol.22,No.1, pp 151-165 **2014**, ISSN 1726-6807, http://www.iugaza.edu.ps/ar/periodical/

Evaluation of Data Mining Classification Models

Abdalla M. EL-HABI, in Applied Statistics
Associate Professor of Statistics
Head of Department of Applied Statistics, Faculty of Economics and Administrative Sciences; Al-Azhar University, Gaza - Palestine.
and
Mohammed El-Ghareeb
M.sc. in Statistics

Abstract: This paper aims to identify and evaluate data mining algorithms which are commonly implemented in supervised classification task. Decision tree, Neural networks, Support Vector Machines (SVM), Naive Bayes, and K-Nearest Neighbor classifiers are evaluated by conducting a simulation study and then assigned to three different datasets to classify and predict the class membership of binary (2-class) and multi-class categorical dependent variables present in these datasets, these datasets were different among each other regarding their size (relatively large and small), and type of predictors (ordinal, numeric, and categorical), as well as number of classes associated with the categorical dependent variable presents in each datasets. Classification performance of these models obtained from a hold-out and 10-fold cross-validation, and empirically evaluated regarding to their overall classification accuracy. We concluded that, there are some differences between the classifiers accuracies, validated by using Hold out and 10-fold cross validation methods assigned to classify a binary categorical dependent variable presents in relatively large dataset, a (3-class) categorical dependent variable presents in relatively small dataset, and a (7-class) categorical dependent variable presents in relatively small dataset, SVM classifier gave the highest averaged rate of classification accuracy in the both methods of validation assigned to these different datasets.

Therefore, we can conclude that the SVM, Neural networks, and k-Nearest Neighbor gave the highest averaged rate of classification, and 10-fold cross validation increased the classifiers accuracies. And this result is approximately matching the conducted simulation results.

Key words: Data mining classification - Decision tree - Neural networks - Support Vector Machine (SVM) - Naive Bayes – k Nearest Neighbor - Hold-out validation - 10-Fold cross-validation – Bootstrapping - Confusion Matrix.

تقييم نماذج مصنّفات التنقيب عن البيانات

ملخص: تهدف هذه الورقة البحثية إلى التعرف على أشهر تقنيات التنقيب عن البيانات وتقييمها، عادةً ما تسمى هذه التقنيات بالمصنّفات، وذلك لاستخدامها في تصنيف المتغيرات التابعة الوصفية أو الفئوية. حيث تم تقييم خمس من هذه المصنّفات وهي : شجرة القرار، والشبكات العصبية، وآلة دعم المتجه، و مصنف بيز، والجار الأقرب من خلال طرق قياس فعالية هذه المصنّفات والحصول منها

على نماذج تنبؤية للتصنيف، تم عمل دراسة محاكاة وتبين منها أن الشبكات العصبية، وآلة دعم المتجه، والجار الأقرب أعطت نتائج عالية نوعا ما في عملية التصنيف، وكذلك تم في هذه الدراسة تطبيق هذه المصنِّفات على ثلاث مجموعات مختلفة من البيانات من حيث الحجم، فمنها الكبيرة نسبياً ومنها متوسطة الحجم، وكانت هذه البيانات مختلفة أيضاً من حيث نوع المتغيرات المستقلة (كمية، أو ترتيبية، أو وصفية)، وأيضاً من حيث عدد فئات المتغير التابع، (ثنائية الفئة، متعددة الفئات). ولتقدير دِقة التصنيف لهذه النماذج قد تم تطبيق هذه المصنفات على طريقتان من طرق قياس الفعالية وهي: Hold out validation و 10−Fold Cross validation وكانت المعيار الرئيسي في تقييم هذه المصنِّفات والنماذج التنبؤية هو التقييم من حيث الدِقة الكلية في التصنيف. تبين لنا من هذه الدراسة وجود بعض الاختلافات بين هذه المصنِّفات في دِقة تصنيفها للمتغير التابع ثنائي الفئة في مجموعة البيانات الكبيرة نسبياً، و تصنيف المتغير التابع ذو الثلاث فئات في مجموعة البيا نات المتوسطة الحجم، وايضا في تصنيف المتغير التابع ذو السبع فئات في مجموعة البيانات المتوسطة الحجم عند استخدامنا كل من Hold out validation و10−Fold Cross validation لقياس فعالية هذه المصنِّفات. كما تبين لنا أن نموذج آلة دعم المتجه هو الأفضل من بين هذه المص نِّفات المتنافسة، حيث أظهر أعلى دِقة في تصنيف تلك المتغيرات باستخدام كلا الطريقتان في قياس الفعالية كما أظهرت هذه الدراسة أن طريقة 10−Fold Cross validation قد زادت من فعالية و دِقة هذه المصنِّفات وقد كانت النتائج في حالتي دراسة المحاكاة والبيانات الحقيقية متقاربة نوعا ما.

1. INTRODUCTION

Data mining is an extension of traditional statistical methods, allows development of new techniques to deal with more complicated data type to satisfy and match the needs for advanced data analyzing. Data mining methods and algorithms serves statistics in several tasks, such as tasks of classification, prediction, clustering, and etc. There are several data mining techniques and predictive models are available for classification task, these techniques are called classification models or classifiers. This study will concentrates on identifying and evaluating of the five techniques commonly used for classification: Decision tree, Neural networks, Support Vector Machines (SVM), Naive Bayes, and k-Nearest Neighbor classifiers with the famous methods of validation: Hold-out validation, 10-fold cross-validation, and Bootstrapping. These classifiers will be validated and evaluated according to their empirical performance through a comparative case study. One of the main studies with similar approach that have been done is by Aftarczuk (2007). It was shown that it is very difficult to name a single data mining algorithm to be the most suitable for the medical data. Kiang (2003) considered data mining classification techniques neural networks and decision tree models and three statistical methods (linear discriminate analysis, logistic regression analysis and k-nearest-neighbor) to investigate

how these different classification methods performed when certain assumptions about the data characteristics were violated; he showed that data characteristics considerably impacted the classification performance of the methods. Lim (2007) evaluated the performance of different classification methods using five microarray datasets and simulation datasets. She showed that the performance of the proposed method (developing an ensemble-based classifier with logistic regression models on each of the subsets in a random partition of the parameter space) is consistently good in terms of overall accuracy. Efron and Tibshirani (1983) conducted five sampling experiments and compared leave-one-out cross-validation, several variants of bootstrap, and several other methods. The results indicated that leave-one-out cross-validation gives nearly unbiased estimates of the accuracy. Breiman et al. (1984) conducted experiments using cross-validation for decision tree pruning. They chose ten-fold cross-validation and claimed that it was satisfactory for choosing the correct tree. Jain, Dubes, and Chen (1987) compared the performances of the bootstrap and leave-one-out cross-validation on nearest neighbor classifiers using artificial data and claimed that the confidence interval of the bootstrap estimator is smaller than that of leave-one-out. Weiss (1991) compared stratified cross-validation and two bootstrap methods with nearest neighbor classifiers and concluded that stratified two-fold cross validation has relatively low variance and superior to leave-one-out. The paper is organized as follows: in section two we will recall the theoretical concepts, in section three we will conduct a simulation study, in section four we will do numerical examples in order to see if there is a matching result, and in section five we will conclude.

2. THEORETICAL CONCEPTS

2.1 Data Mining Classification

Data mining (DM) is the process of extracting meaningful information from large datasets. Classification is a data mining task of predicting the value of a categorical variable (target or class). The most five famous data mining techniques used for classification are:

a. Decision Tree

Decision trees are one of the most common data mining techniques invented by (Quinlan 1986), he developed a decision tree algorithm known as 1D3 (Iterative Dichotomiser). As its name implies, a decision tree is a data mining predictive model that can be viewed as a flow chart like a tree structure. Specifically, each branch of the tree is a classification question, and the leaves of the tree are the predicted classification holding a class label.

Decision trees classify instances by sorting them down the tree from the root node to a leaf node, which provides the classification of these instances. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the decision tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. This process is then repeated at the node on this branch and so on until a leaf node is reached, which holds the class prediction for that instance. Decision trees can easily be converted to classification rules (Mitchell, 1997).

b. Neural Networks

Neural networks, also called artificial neural network (ANNs), are one of the most famous predictive models used for classification. An artificial neural network is a system based on the operation of biological neural networks, in other words, is an emulation of biological neural system. Artificial neural networks born after McCulloc and Pitts introduced a set of simplified neurons in 1943. These neurons were represented as models of biological networks into conceptual components for circuits that could perform computational tasks. Artificial neural network is developed with a systematic step-by-step procedure which optimizes a criterion commonly known as the learning rule. The input/output training data is fundamental for these networks as it carries the information which is necessary to discover the optimal operating point. There are various neural networks architectures, the most successful applications in classification and prediction have been multilayer feed forward networks. The layer where input patterns are applied is the input layer; the layer from which an output response is desired is the output layer. Layers between the input and output layers are known as hidden or transfer layers, because their outputs are not readily observable.

c. Support Vector Machines

Support Vector Machines (SVM) is one of the most recent Data mining techniques used for classification, developed by Cortes and Vapnik in 1995 for binary classification, (Cortes and Vapnik, 1995). SVM have been developed in the framework of statistical learning theory (Vapnik, 1998), and have been successfully applied to a number of applications, ranging from time series prediction, to face recognition, to biological data processing for medical diagnosis (Evgeniou, et al., 1999). SVM classification finds the hyper -plane where the margin between the support vectors is maximized. If all classifications contain two-class dependent variables with two predictors, then the points of each class could be easily

divided by a straight line. SVM is an algorithm for the classifications of both linear and nonlinear data.

d. Naive Bayes

The Naive Bayesian classifier is based on Bayes' Theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

e. Nearest Neighbor

Nearest Neighbor (more precisely k-nearest neighbor, also KNN) was first described in the early 1950s, is one of data mining non-parametric algorithms that are very simple to understand, KNN assumes that the data is in a feature space. Since the points are in feature space, they have a notion of distance; this need not necessarily be Euclidean distance although it is the one commonly used (Saravanan, 2010). Unlike the other methods, K-Nearest Neighbor requires no training. It works on the idea that close objects are more likely to be in the same class using a metric for measuring the distance between the query points we want to classify. Euclidean distance is one of the most common metrics to measure this distance. Therefore, KNN classification and prediction of new instances are based on the majority vote of the closest points to our query point.

2.2 Model Validation

Validation of a data mining classification model is the most important phase in data mining process. Validation is the process of assessing how well your data mining models perform against real data. The main concept of model validation is to estimate the model predictive performance. There are several methods regarding the model validation issues. We introduce some basic concepts of famous model validation methods:

a. Cross-Validation

The idea for cross-validation originated in (Larson, 1931), when one sample is used for regression and a second for prediction. Cross-Validation is a statistical method of evaluating and or comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. The end result is a model that has learned how to predict our outcome given new unknown data. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against (karami, et al, 2012).

b. Hold-Out Validation

The hold-out validation method, sometimes called test sample estimation. A natural approach is to split the available data into two non-overlapped parts: one for training and the other for testing. The test data is held out and not looked at during training (Barnali and Mishra, 2011). Partitions the data into two mutually exclusive subsets, it is common to designate 2/3 of the data as the training set and the remaining 1/3 as the test set. (Kohavi, 1995). Hold-out validation avoids the overlap between training data and test data, yielding a more accurate estimate for the generalization performance of the algorithm. The downside is that this procedure does not use all the available data and the results are highly dependent on the choice for the training/test split. The instances chosen for inclusion in the test set maybe too easy or too difficult to classify (Zadeh et. al, 2009).

c. K-fold Cross-Validation

In k-fold cross-validation, the actual data is first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining k - 1 folds are used for learning. In data mining 10-fold cross-validation (k = 10) is the most common method, and serves as a standard procedure for performance estimation and model selection. (Kohavi,1995). It is the basic form of cross validation, the idea of this method is to divide the data set into a number of k folds and evaluate the errors. The most common choice for evaluating a classification task is the accuracy (the percentage of correctly classified cases). All other possible famous names of validation methods are seem to be as special cases of k folds cross validation depending on the choosing value of k. Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross-validation where k equals the number of instances in the data. In other words in each iteration nearly all the data except for a single observation are used for training and the model is tested on that single observation [k = n-1]. An accuracy estimate obtained using LOOCV is known to be almost unbiased but it has high variance, leading to unreliable estimates (Efron, 1983). It is still widely used when the available data are very rare, especially in bioinformatics where only dozens of data samples are available. (Zadeh, et. Al, 2009).

d- Bootstrap

The bootstrap family was introduced by (Efron, 1983) and is fully described in Efron & Tibshirani (1993). Given a dataset of size n, a bootstrap sample is created by sampling n instances from the data (with replacement). Since the dataset is sampled with replacement, the probability of any given

instance not being chosen after n samples is $(1\text{-}1/n)^n \approx e^{-1} \approx 0.368$; the expected number of distinct instances from the original dataset appearing in the test set is thus $0.632n$. (Kohavi, 1995). We mentioned that there are other methods also to be as special cases of k folds cross validation, like Leave-one-out cross-validation (LOOCV), and Bootstrapping, which are beyond of our analysis. So here we just gave the reader a brief idea of them.

2.3 Evaluation of a Classification Model

Classifiers and predictive models evaluation is one of the key points in any data mining process. The main and frequently evaluation criteria desired in classification perspective is the criteria of overall accuracy obtained by model validation method. One of the famous methods of visualizing the metric predictive performance can be obtained by constructing a confusion matrix. Specific matrix layout that allows visualization of the performance of an algorithm shows the predicted and actual classifications. A confusion matrix is of size L x L, where L is the number of classes. The matrix is a valuable tool because it not only shows how frequently the model correctly predicted a value, but also shows which other values the model most frequently predicted incorrectly.

General form of Confusion Matrix

Let $N_{i\,j}$ be the number of elements in the population, of size N, which are really of type j but are classified as being of type i. The matrix $\{ N_{i\,j} \}$ is usually represented as:

True class, *j*

		N_{11}	N_{12}	. . .	N_{1j}
	Assigned class, *i*	N_{21}	N_{22}	. . .	N_{1j}
.	.				
.	.				
.	.				
		N_{i1}	N_{11}	. . .	N_{ij}

The diagonal elements of this matrix are the counts of the correct classifications of each type. Several metric performance evaluations of a classification model can be obtained by confusion matrix, the main and importance one is the measure of the accuracy of the classification model which is the proportion of correctly classified instance. In contrast the error of classification resulted by the classifier or usually known as misclassification error, can be obtained by confusion matrix, and could be as other criteria of metric performance evaluations of a classification model.

The matrix shows the accuracy of the classifier as the percentage of correctly classified patterns in a given class divided by the total number of patterns in that class. The overall (average) accuracy of the classifier is also evaluated by using the confusion matrix is presented by applying the following formula:

$$\text{Overall Accuracy Rate} = \frac{\sum_{i} N_{ii}}{\sum_{i,j} N_{ij}} \quad (2.1)$$

Error made by the classifier, if a case is of class j but is not classified as such. On the other hand the assignment of a case to class i when it is not of this class. Hence the proportion of cases of type *j* which are misclassified is:

$$\text{Rate of Misclassification} = 1 - \frac{\sum_{i} N_{ii}}{\sum_{i,j} N_{ij}} \quad (2.2)$$

Or simply, misclassification error = 1- overall accuracy rate.

3. SIMULATION STUDY

We performed a simulation study to evaluate the five proposed data mining classifiers described previously: Decision tree, neural networks, support vector machine, naïve Bayes, and k-Nearest Neighbor.

We draw random samples from two different normal distributions with different mean vectors, but equal covariance matrices, we used the identity matrix as the covariance matrix. The two different normal distributions were generated from N(1, 1) for the first group and N(0, 1) for the second. The data sets are generated to observe the impact of changes regarding the sample size, categorization and correlation matrices between the predictors.

Sample size: The samples are simulated from normal distributions with the same covariance matrix and different mean vectors, which are divided equally into 2 classes. These simulations are based on an R function (*mvrnorm*) for simulating from a multivariate normal distribution from R package MASS. Five different sample sizes are generated 100, 200, 400, and 500 to observe the impact of changes related to sample size.

Correlation: We used strong and weak correlation matrices in order to evaluate the performance of the different classifiers in terms of the presence of multicollinearity and to examine the effect of correlation between explanatory variables. 2 simulated samples with two predictors

have correlation coefficients of 0.25 and 0.90 were used for every simulated data set this purpose.

Categorization: After sampling, the normally distributed variables can be categorized, and divided into a certain number of categories of equal size to assess each method in terms of number of categories. Four different number of categories are considered (2, 3, 4 and 10).

All possible combinations between mentioned sample size, correlation and the number of categories are considered. This process is repeated 2000 times to achieve the convergence criterion. For each simulation replication, 10-fold cross validation was performed for evaluating the performance of each classification method.

The average of those 2000 correct classification rates is then obtained to estimate the true classification rate of the different classifiers.

A part of simulations results are presented in the followings:

Tables 3.1, 3.2, and 3.3 show the measures of Data mining (DM) classifiers overall accuracies versus the sample size, correlation, and categorization

Table 3.1: Overall Classification Accuracy versus the Sample Size, with weak correlation, and no. of categories 3.

Sample size	*DT*	*Neural Networks*	*k-Nearest Neighbor*	*Naïve Bayes*	*SVM*
100	0.820	0.845	0.868	0.820	0.840
200	0.840	0.850	0.874	0.830	0.855
400	0.865	0.865	0.880	0.855	0.865
500	0.875	0.884	0.885	0.860	0.870

Table 3.2: Overall Classification Accuracy versus the Correlation, *n*=200, and no. of categories = 4

Correlation	*DT*	*Neural Networks*	*k-Nearest Neighbor*	*Naïve Bayes*	*SVM*
0.90	0.785	0.850	0.820	0.790	0.820
0.25	0.840	0.880	0.890	0.830	0.865

Table 3.3: Overall Classification Accuracy versus the Categorization, n= 300 and strong correlation

No. of Categories	*DT*	*Neural Networks*	*k-Nearest Neighbor*	*Naïve Bayes*	*SVM*
2	0.840	0.885	0.878	0.830	0.850
3	0.835	0.855	0.860	0.825	0.855
4	0.805	0.840	0.850	0.835	0.825
10	0.740	0.755	0.760	0.750	0.730

According to the simulation results for the effect of sample size shown in table 3.1, it can be seen that the variation in sample size has similar effect on almost all the methods and as the sample size increases the classification accuracy increases.

When looking at the effect of the presence of multicollinearity on the performance of a method, we can see from Table 3.2 that all the methods have significant improvement in performance in the absence of multicollinearity. The performances of the Neural Network, k-Nearest Neighbor, and Support Vector Machines are superior compared with all other methods for any correlation level.

When looking at the effect of the number of categories on the performance of a method, it can be seen from Table 3.3 that as the number of categories increases, the classification accuracy decreases for the performance of all the methods, and the classification performance in case of a binary categorical variable for each method is superior to its performance in case of more than two classes categorical variable.

From the simulation results in the different situations (sample size, categorization and correlation), it seems that Neural Networks and k-Nearest Neighbor, gave the highest averaged rate of classification accuracies.

4. NUMERICAL EXAMPLES

4.1 Data Description

We used three different datasets as shown in table 4.1, these datasets are different among each other in their size, number and type of predictors, and also different in the number of the classes of their dependent variable, they have a binary and multi-class categorical dependent variable.

Table 4.1: A brief description of the properties of the datasets

Dataset	Size	No. of variables	Type of predictors	No. of Classes
Diabetics	1566	11	Mixed	2
Iris	150	5	Numeric	3
Fish Species	159	6	Numeric	7

The mixed type of predictors means that the independent variables or predictors are having different variable types (ordinal, numeric, and categorical). R Statistical package is mainly used for data exploration, description, and classification modeling and analysis. In the following, some brief descriptions of each dataset:

4.2 Diabetics Dataset

This data was taken from the Demographic and Health Survey done in 2004, by Palestinian Central Bureau of Statistics. As shown in Table 4.2, a brief description of diabetics dataset, shows the important properties of this dataset regarding the size of this dataset and the type and number of the predictors available. The dependent variable of this dataset has a binary (2-class) categorical variable, it will be our target here to classify and predict the class membership of persons that diabetics (labeled as Yes), and those not diabetics (labeled as No).

Table 4.2: Diabetics Data Description

	Description		
Dataset Size	1577 records	**Number of Attributes**	11
Dependent Variable	2-ClassCategorical Variable	Yes (1)	758 Cases
		No (2)	808 Cases
Number of Predictors	10 variables–6 categorical,1 ordinal, and 3 numeric attributes		

4.2 Iris Dataset

A famous multivariate dataset exists in R MASS package; we will use this data set to classify the Iris plant species according to the measure of its sepal length and width, along with its petal length and width. As we can see in Table 4.3, a brief description of Iris dataset, shows the important properties of this dataset regarding the size of this dataset, the type, and the number of the available predictors. The dependent variable of this dataset considered to be a multi-class (3-class) categorical variable, it will be our target here to classify and predict the species of the Iris plant of Setosa, Versicolor, and Virginica.

Table 4.3: Iris dataset description:

	Description		
Dataset Size	150 records	**No. of Attributes**	5
Dependent Variable	3-Class Categorical Variable	Setosa	50 Cases
		Versicolor	50 Cases
		Virginica	50 Cases
Number of Predictors	4 variables - Numeric attributes		

4.3 Fish Species Dataset

The Fish Catch dataset from the Journal of Statistical Education (Web1) contains measurements on 159 fish caught in the lake Laengelmavesi, Finland. For the 159 fishes of 7 species the weight, length, height, and width

were measured. Three different length measurements are recorded: from the nose of the fish to the beginning of its tail, from the nose to the notch of its tail and from the nose to the end of its tail. The height and width are calculated as percentages of the third length variable. This result in 6 observed variables, Weight, Length1, Length2, Length3, Height, and Width. Observation 14 has a missing value in variable Weight, therefore this observation is usually excluded from the analysis, The 7 species are 1=Bream, 2=Whitewish, 3=Roach, 4=Parkki, 5=Smelt, 6=Pike, 7=Perch. A brief description of this data set is illustrated in the following table 4.4

Table 4.4: Fish dataset description

	Description		
Dataset Size	159 records 1 is missing	**Number of Attributes**	7
Dependent Variable	7-Class Categorical Variable	Bream	35 Cases
		Parkki	11 Cases
		Perch	56 Cases
		Pike	17 Cases
		Roach	20 Cases
		Smelt	14 Cases
		Whitewish	6 Cases
No. of Predictors	6 variables- Numeric attributes		

Here we are going to use this to classify the various seven species of fish according to their body measures.

Classification and Validation Results

As a summarization of classification analysis identifying the overall accuracies obtained by validating that set of competing classification using both of Hold-out and 10- fold cross validation methods. Table 4.5 shows the measures of DM classifiers overall accuracies and ability to classify and predict the class membership of binary (2–class) and multi-class categorical dependent variables presented in the three datasets using 10-fold cross-validation method.

Table 4.5: Overall Classification Accuracy (using *10- fold Cross-Validation*)

	DT	Neural Networks	SVM	Naïve Bayes	k-Nearest Neighbor
Diabetics	0.776	0.761	0.778	0.75	0.754
Iris	0.96	0.973	0.96	0.96	0.97
Fish	0.804	0.899	1	0.898	0.816

A similar work has been done for a similar approach, but this time another method of validation is used to measure classifiers accuracies. Table 4.6 shows the results of classification accuracies obtained by using hold out validation method.

Table 4.6: Overall Classification Accuracy (*using Hold-Out Validation*)

	DT	Neural Networks	SVM	Naïve Bayes	k-Nearest Neighbor
Diabetics	*0.787*	*0.767*	*0.787*	*0.737*	*0.772*
Iris	*0.96*	*0.98*	*0.98*	*0.96*	*0.98*
Fish	*0.827*	*0.865*	*0.865*	*0.692*	*0.728*

5. Conclusion

After viewing the results and comparing them, we may conclude the following:

There are slight differences between the classifier accuracies, validated by using 10-fold cross validation method assigned to the Diabetics dataset. However, we may consider SVM to be the most accurate classifier, since it gave the highest rate among the competing classifiers. But here we can mention that the accuracy for DT = .776 which is very close to the accuracy for SVM = .778. The same thing happened for Iris dataset. Neural networks and K- Nearest Neighbor have the most accurate classifier. In the case of assigning these classifiers to Fish dataset, we can see some differences of the rate of accuracies, and in this case SVM was the perfect classifier that gave 100% overall accuracy. We may due this to the characteristics and properties of datasets used. Where Fish dataset has 7 class categorical and unbalanced dependent variable. Therefore, we may conclude that SVM and Neural Networks are suitable classifiers to be assigned for such a case. For the same approach with other scenario, where Hold-out validation method used to measure the classifiers performance, both of Decision tree and SVM gave the highest rate of classification accuracies when assigned to Diabetics dataset. In the case of Iris dataset, all of neural networks, SVM, and k-

Nearest Neighbor gave the same highest rate of classification accuracies. And, in case of Fish dataset, both neural networks and SVM gave the same highest rate of classification accuracies.

On the other hand, we may also conclude that SVM gave the highest averaged rate of classification accuracies in the both methods of validation. By reviewing the simulation results, we can conclude that our real data results are approximately matching the simulation results and 10-fold cross validation increased the classifiers accuracies. It is matching the recommended results in such literature.

We suggest for future work to do more numerical examples using more recent data in order to achieve more reliable results.

REFERENCES

1. Sahu, B and Mishra, D. (2011), Performance of feed forward neural network for a Novel Feature Selection Approach, *IJCSIT*, **2** (4), 1414-1419.

2. Breiman L., Friedman, J. H., Olshen R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth.

3. Corinna, C. and Vladimir, V. (1995), *Support-Vector Networks*, Machine Learning, 20.

4. Efron, B. (1983), *Estimating the error rate of a prediction rule: improvement on cross-validation*, Journal of the American Statistical Association, 316–331.

5. Efron, B., and Tibshirani, R. J. (1993), *An introduction to the bootstrap*, Chapman & Hall, New York.

6. Jain, A. K., Dubes, R. C. and Chen, C. (1987), *Bootstrap techniques for error estimation*, IEEE transactions on pattern analysis and machine intelligence, Volume: PAMI-9, Issue:5, 628-633.

7. Aftarczuk, K. (2007), *Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems*, Master thesis.

8. Kiang, M. (2003), A comparative assessment of classification methods, *Decision Support Systems*, 3**5**, 441- 454.

9. Kohavi R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *In Proceedings of International Joint Conference on AI*.

10. Larson S. (1931), *The shrinkage of the coefficient of multiple correlation*, Educat. Psychol, 22:45–55

11. Lim, N. (2007), *Classification by Ensembles from Random Partitions using Logistic Regression Models*, Ph.D. thesis.

12. Zadeh, P. R., Tang, L. and Liu, H. (2009), *Cross Validation. In Encyclopedia of Database Systems*, Springer.
13. Quinlan, J. R (1986), Induction of decision trees, *Machine Learning*, **1**, 81-106.
14. Thirumuruganathan, S. (2010), *A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm*, WordPress.com weblog.

15. Evgeniou, T. Pontil, M. (1999), *Workshop on Support Vector Machines: theory and applications*.
16. Mitchell, T. M. (1997), Decision tree learning, *in T. Mitchell, Machine Learning*. 52-78.
17. Vapnik, V. (1998), *Statistical Learning Theory*, Wiley, New York.
18. Weiss, S. M. (1991), Small sample error rate estimation for k-nearest neighbor classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**, 285-289.

Web References

- Web1: Journal of Statistical Education, Fish Catch Data Set, [http://www.amstat.org/publications/jse/datasets/fishcatch.txt] accessed August, 2006.