Int. J. Mach. Learn. & Cyber. (2013) 4:13–24
DOI 10.1007/s13042-012-0068-x

ORIGINAL ARTICLE

Comparative study on classification performance between support vector machine and logistic regression

Abdallah Bashir Musa

Received: 4 September 2011 / Accepted: 2 January 2012 / Published online: 24 January 2012

Abstract Support vector machine (SVM) is a comparatively new machine learning algorithm for classification, while logistic regression (LR) is an old standard statistical classification method. Although there have been many comprehensive studies comparing SVM and LR, since they were made, there have been many new improvements applied to them such as bagging and ensemble. Recently, bagging and ensemble learning have become hot topics, widely used to improve the generalization performance of single learning algorithm. Therefore, comparing classification performance between SVM and LR using bagging and ensemble is an interesting issue. The average of estimated probabilities' strategy was used for combining classifiers in this paper. Different evaluation metrics assess different characteristics of machine learning algorithm. It is possible for a learning method to perform well on one metric, but be suboptimal on other metrics. Therefore this study includes a variety of criteria to evaluate the classification performance of the learning methods: accuracy, sensitivity, specificity, precision, F-score and the area under the receiver operating characteristic curve. This has not been included in previous studies of SVM, owing to the fact that it did not support estimated probabilities at that time. Other metrics used in medical diagnosis, such as, Youden's index (γ), positive and negative likelihoods ($\rho+$, $\rho-$) and diagnostic odds ratio were evaluated to convey and compare the qualities of the two algorithms. This study is distinct by its inclusion of a comprehensive statistical analysis for the results of the SVM and LR algorithms on various data sets.

A. B. Musa (✉)
Faculty of Mathematical Sciences and Computer,
University of Gezira, Wad Madani 20, Sudan
e-mail: abdubashir20@yahoo.com

Keywords Support vector machine (SVM) · Logistic regression (LR) · Machine learning algorithm · Bagging · Ensemble · Statistical analysis

1 Introduction

Logistic regression (LR) [1, 2] is a multivariable method devised for dichotomous outcomes. It is a standard statistical classification method which is particularly appropriate for models involving disease state (healthy/diseased), decision making (yes/no), or mortality (dead, living). It is widely used in binary classification problems in applied sciences such as medicine, biology and epidemiology. It has been widely applied due to its simplicity and great interpretability. Logistic regression needs special requirements regarding the data under consideration, such as, little or no collinearly among the independent variables and linearity of the independent variables with the logit. In contrast, SVM [3, 4, 5] recently, has become a very popular machine learning tool for classification. It is easy and uncomplicated as compared to LR. Nowadays, SVM is used intensively in data mining, which is a general term for the science of extracting useful information from large databases or data sets.

There are many empirical studies for comparing machine learning algorithms; these studies also include the comparison of LR and SVM. For example, Perlich et al. [6] constructed a curve analysis comparison between the decision trees and logistic regression using bagging, STATLOG [7] presented a study that included several machine learning algorithms and LR but it did not include SVM, [8] presented a study that compared logistic regression (LR), probabilistic neural network (PNN) and support vector machine (SVM) classifiers for discriminating

between normal and Parkinson disease (PD). There are various other pair comparison studies including LR and/or SVM such as: LeXu and Gao [9] who presented a study that compares logistic regression with artificial neural network (ANN) on power distribution system, Chen et al. [10] constructed a comparison study between SVM and back propagation neural networks in forecasting the six major Asian stock markets. There are numerous studies in medical fields too, such as Song et al. [11] who performed Comparative Analysis of Logistic Regression and ANN for Computer-aided Diagnosis of Breast Masses. All these studies provide comparisons pair classifiers using only one dataset for a single problem. To the best of our knowledge, the only comparison between SVM and LR has been done for the prediction of hospital mortality in critically ill patients with hematological malignancies [12]. This comparison focuses only on the mortality prediction model. In this study the authors divided the data set into training and testing sets, the dataset they used has only 350 instances. None of the previous studies used statistical analysis for evaluating the performance of the classifiers under comparison. Moreover, there are plenty of new improvements that have been applied to the classification methods to improve their performance, such as, bagging and ensemble. The classifiers' performance needs to be compared after incorporating the improvements.

The aim of this paper is to construct a standard, comprehensive comparative study between SVM and LR on multiple data sets. Recently, combining multiple classifiers has been a very active research technique. It is widely accepted that combining multiple classifiers can achieve better classification performance than a single classifier [13, 14], therefore, the bagging [15] predictors method has been used. Bagging is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The main idea of bagging is to make various samples of the training set. A classifier is generated for each of these training set samples by a selected learning algorithm. So, for k samples of the training data set we get k particular classifiers. There are many strategies for aggregating these classifiers. In this paper, the average of the estimated probabilities' strategy was used for aggregation. A variety of performance metrics have been used: accuracy, sensitivity, specificity, precision and F-score, to asses the algorithms' performances. These standard metrics were combined with other metrics that measure other properties, such as, failure avoidance and class discrimination. These metrics are Youden's index, positive and negative likelihoods and diagnostic odds ratio (DOR), which may be useful in discovering unseen characteristics of the algorithm's performance. Furthermore, The receiver operating characteristic (ROC) analysis has been used in this study, which is a more powerful evaluation tool and has not been included in previous studies. The statistical significant difference between each pair of ROC curves was tested using the Mann–Whitney nonparametric test. Statistical evaluation of experimental results is an essential part of the comparison validation, in this study, a detailed concept of using statistical analysis in comparing two algorithms has been given.

2 Classification methods

The most widely studied and well understood learning protocol is supervised learning, where a learning algorithm uses labeled instances to formulate a predictive model [16]. Logistic regression and support vector machine are two supervised classification methods which are broadly used. Logistic regression is a parametric method to analyze dichotomous response variable and finds the relationship between the response variable and the independent variables. It has been widely applied in medicine fields, but seldom used in machine learning studies. Support vector machine is also a parametric method which has been broadly used in machine learning studies. Recently, it has been extensively used in classification problems and successfully applied to many real fields [17, 18].

2.1 Support vector machine

Support vector machine (SVM) [3–5] is a comparatively new classification method. It has drawn much attention in recent years [19]. The concept of SVM is as follows: input vectors x are mapped to a very high dimension feature space z through some nonlinear mapping $\Phi(x), z = \Phi(x)$. In this space, an optimal separating hyperplane is constructed. For a given training dataset with n samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is a feature vector in a d-dimensional feature space R^d and $y_i \in \{1, +1\}$ is the corresponding class label. The task is to find a classifier with a decision function f (x, w, b) $= w^T\Phi(x) + b$, SVM finds an optimal hyperplane with the maximal margin that separates the data points into two classes. To find the optimal separating hyperplane having maximal margin, a learning machine should

$$\min \frac{1}{2} w^T w \quad i = 1, \ldots, n \tag{1}$$

Subject to

$$y_i[w^T\Phi(x_i) + b] \geq 1 \quad i = 1, \ldots, n \tag{2}$$

where w is the normal vector for the "separating" hyperplane, $(W, \Phi(x_i)) + b = 0$, this can be transferred into its dual form by minimizing the following primal lagrangian

$$L_d(w,b,\alpha) = \frac{1}{2}w^T w - \sum_{i=1}^{n} \alpha_i \{y_i[w^T \Phi(x_i) + b] - 1\} \quad (3)$$

In respect to w and b by using $\partial L_d/\partial w = 0$ and $\partial L_d/\partial b = 0$, i.e., by exploiting

$$\frac{\partial L_d}{\partial w} = 0, \quad w = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i) \quad (4)$$

$$\frac{\partial L_d}{\partial b} = 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (5)$$

Substituting w from (4) and using (5), this lead to the following dual lagrangian problem

$$L_d(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \quad (6)$$

where, $k(x_i, x_j) = \Phi^T(x_i)\Phi(x_j)$ is a Mercer's kernel that allows us to calculate the dot product in high-dimensional space without explicitly knowing the nonlinear mapping. The $L_d(\alpha)$ in (6) should be solved subject to the following constraints:

$$\begin{aligned} &\alpha_i \geq 0 \quad i = 1, \ldots, n \\ &\sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \quad (7)$$

In a more general case, when the problem is not separable or is judged too costly to separate due to an overlapping of training data points, the constraints in solving dual lagrangian problem in (6) change to the following constraints:

$$\begin{aligned} &0 \leq \alpha_i \leq c \quad i = 1, \ldots, n \\ &\sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \quad (8)$$

where $(\alpha_1, \ldots, \alpha_n)$ are the weights assigned to the training sample x_i. If $\alpha_i > 0$, x_i is called a support vector. c is a "regulation parameter" used to trade-off the training accuracy and the model complexity so that a superior generalization capability can be obtain. There are different forms of kernel function, however, support vector machine (SVM) with the Gaussian (RBF) kernel has been popular for practical purposes, since it can handle the case when the relation between classes and features is nonlinear, and it also has less parameter than other nonlinear kernels such as the polynomial kernel [20–22], therefore, RBF kernel function which is given in Eq. (9) is used in this paper.

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (9)$$

After the lagrangian variables $(\alpha_1, \ldots, \alpha_n)$ calculated by solving (6) subject to (8) and using (4), the decision function can be formulated as follows:

$$f(x) = w^T \Phi(x) + b = \left(\sum_{i=1}^{n} \alpha_i y_i k(x, x_i) + b\right) \quad (10)$$

where x is the d-dimensional vector of the test examples and b is the SVM bias term which depends upon the applied kernel, it can be implicitly part of the kernel function. It will be found by fulfilling the requirements that the values of a decision function at the support vectors should be the given y_i, $(y_i = \pm 1)$. i.e. $f(x_s) = y_s = \pm 1$. For the given Pattern x_p, if $f(x_p) > 0$, Pattern x_p belongs to class $(y = +1)$, otherwise, it belongs to class $(y = -1)$.

2.2 Logistic regression

Logistic regression (LR) [1, 2] is a well known statistical approach to model dichotomous (binary) data; logistic regression is a member of generalized linear models. In logistic regression, a single outcome variable y_i, where $i = 1, \ldots, n$, each y_i takes only two values 0 or 1 (but not both), so it follows a Bernoulli Probability density function $p(y_i) = (\pi_i)^{y_i}(1 - \pi_i)^{1-y_i}$. that takes the value 1 with probability π_i and 0 with probability $(1 - \pi_i)$. Our interest is in $y_i = 1$ with the interest probability π_i, which varies over the observations as an inverse logistic function of a vector $\mathbf{x}_i$, which includes a constant (x_0) and k explanatory variables (x_1,…, x_k). Its function can be given as follows:

$$\begin{aligned} &y_i \sim \text{Bernoulli}(\pi_i) \\ &p(y_i = 1) = \pi_i = (1 + e^{-x\beta})^{-1} \end{aligned} \quad (11)$$

where $\beta = (\beta_0, \beta_1')'$ is a $(k+1) \times 1$ vector that contains the parameters that need to be estimated, β_0 is an intercept term corresponding to x_0 and β is $(k \times 1)$ vector with elements corresponding to the explanatory variables. The odd ratio of $y = 1$ is $p(y = 1)/(1 - p(y = 1) = \pi_i/(1 - \pi_i)$. By using this odd ratio; the following transformation can be obtained.

$$\text{log}it(y_i = 1) = \ln[odd] = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta x_i \quad (12)$$

The above logit function can be expressed in matrix form as follows:

$$\text{log}it[p(y = 1)] = x\beta \quad (13)$$

The importance of the transformation in (13) is that it has many of the desirable properties of the linear regression model. The logit is linear in the parameters vector β. These parameters will be estimated using the maximum likelihood function. The maximum likelihood function of Bernoulli density function is $L(\pi_i/y_i) = (\pi_i)^{y_i}(1 - \pi_i)^{1-y_i}$. By assuming independence over the observations, the maximum likelihood function for $y = y_1, \ldots, y_n$ can be written as follows:

$$L(\beta/y) = \prod_{i=1}^{n} (\pi_i)^{y_i} (1-\pi_i)^{1-y_i} \tag{14}$$

By taking the logarithm, the log-likelihood will be

$$L(\beta/y) = \sum_{i=1}^{n} [y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)] \tag{15}$$

After estimating the parameters, the significance of each of these parameters will be assessed by comparing the observed values of the response variable to the predicted values obtained from the model with and without the variable in the model. In logistic regression this comparison is based on the log likelihood function defined in (15). This can be obtained by using the following statistic:

$$G = -2\left[\frac{\textit{likelihood with out the variable}}{\textit{likelihood with the variable}}\right] \tag{16}$$

This statistic will be compared with $\chi^2(\alpha, 1)$ to test the hypothesis whether the parameter is equal to zero or not, if $G > \chi^2(\alpha, 1)$, then the parameter is not significant and should be deleted from the model. There are several selection procedures used to construct the best fitting model such as forward selection which looks at each explanatory variable individually and selects the single explanatory variable that fits the data the best on its own as the first variable included in the model, among the remaining variables the one that adds the most is included. This is repeated until none of the remaining variables will add significantly. Backward selection starts with a model that contains all of the explanatory variables, and then a variable that, if removed, would cause the smallest change in the overall fit of the model is removed. This continues until all variables in the model are significant. For assessing the goodness-of-fit for the model, there are several goodness-of-fit tests that can be obtained by comparing the overall difference between the observed and fitted values. Among these tests Pearson Chi-Square χ^2 and Deviance D test are used the most. Suppose the number of the covariate patterns is j, let $j < n$, let m_i denote the number of $(y_i = 1)$ among these patterns. The Pearson statistic is defined as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \tag{17}$$

And the residual deviance statistic is defined as follows:

$$D = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)\ln\left(\frac{(m_i - y_i)}{m_i(1-\hat{\pi}_i)}\right)\right] \tag{18}$$

It is clear that the above two statistics rely on the principle of comparing observed y_i to predicted $m_i\hat{\pi}_i$ values and they should be small if the model fits the data well. These two statistics are compared to the value of $\chi^2(\alpha, n-k-1)$ to judge their statistical significance. These statistics are used when j < n. Their results are invalid when $j \sim n$ [1, 23]. In this case there are other alternative statistics that can be used, such as Osius and Rojek statistic, Farrington statistic and Hosmer–Lemeshow statistic.

The predicted label for the logistic regression model will equal to 1 if $\hat{\pi}_i$ is greater than or equal to some threshold (the default is 0.5), as shown below:

$$\begin{aligned} &\textit{if } (p(y=1)) \geq 0.5 \quad \textit{the instance} \in \textit{class}\,(y=1) \\ &\textit{if } (p(y=1)) < 0.5 \quad \textit{the instance} \in \textit{class}\,(y=0) \end{aligned} \tag{19}$$

3 Materials and methods

3.1 The data sets

The data sets used in this study were composed of 13 data sets with binary class attributes, 11 from the UCI repository (ftp to ics.uci.edu/pub/machine learning-databases) and 2 from the LIB-SVM data: classification (binary class) at (http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets). These data sets are of different sizes, six of them are almost balanced and the remaining seven are unbalanced. Table 1 gives a numerical summary of the data sets.

3.2 Bagging and aggregating classifier decisions

Ensembles of classifiers represent one of the main research directions in machine learning [24]. Empirical studies showed that both in classification and regression problems ensembles are often much more accurate than the individual base learner that make them up [25], recently there have been different ensemble methods. Bagging [15] is one of the most important recent developments in classification methodology. This method was proposed by Leo Breiman in 1996. Using bagging in many classification algorithms results in high improvement in performance and gives substantial gains in accuracy. Breiman shows that bagging works well for unstable procedures where a small change in the training data set can result in large changes in predictions (e.g., neural networks, decision trees). Although SVM is a stable classification method, its performance can generally be improved by bagging [26, 25]. Bagging has been applied widely to machine learning techniques, but it has rarely been applied to statistical tools such as logistic regression [6].

Bagging works by sequentially applying a selected classification algorithm in respect to modifications of the training data set. So for each sub sample of the training data a classifier should be created.

Table 1 Summary of the data sets

Data set	Data size	Number of variables	Nominal	Total	Data type
Page block	5,473	10	0	10	Unbalanced
Cod-rna (sample)	5,136	8	0	8	Balanced
Spam	4,601	57	0	57	Unbalanced
Chest	3,196	36	36	73	Balanced
Car evaluation	1,594	6	6	21	Unbalanced
Contraceptive	1,474	9	7	24	Balanced
German number	1,000	24	0	24	Unbalanced
Pima diabetes	768	8	0	8	Unbalanced
Breast cancer	683	10	0	10	Unbalanced
Credit approval	650	14	7	34	Balanced
Ionosphere	351	34	0	34	Unbalanced
Liver disorder	345	6	0	6	Balanced
Heard Scale	270	13	0	13	Balanced

Our experiment was done according to the following bagging algorithm.

(i) Initialization of the training data set T.
(ii) Divide the training data set T into two sets T1 and T2, based on the data classes.
(iii) Draw two random samples (bootstraps) with replacement from T1 and T2 with the same proportions (some of the examples can be selected repeatedly and some may not be selected at all).
(iv) Mix the two samples (bootstraps) together to represent the new training data set, in this way the proportions of the classes will be the same as in the original data set, and all the training data sets will be of the same size.
(v) Train a particular classifier using this sub training data set by a selected Learning algorithm.
(vi) By repeating the previous steps K times, K classifiers will be obtained.

Any instance in the training data set T has the probability $[1 - (1 - 1/\mathrm{K})^{\mathrm{K}}]$ of being selected, at least once in the K times randomly selected instances from the training data set. For a large k, as in this experiment where $k = 100$, this probability will approximately equal to 0.634 which means that each sub training sample contains about 63.4% unique instances from the original training data set. In this way we can build classifiers with samples that are not identical.

After, each classifier is trained independently for each algorithm. We have to aggregate their results in an appropriate combination approach. Some combination strategies are suggested by previous studies. First, the simplest one is a majority vote which can be used where only class's labels are considered. Second, for the case of continuous-valued outputs like posteriori probabilities are available, the average of the estimated probabilities can be an appropriated strategy. In this case the decision is made according to the mean of posteriori probabilities of the combined classifiers. Third is the average of estimated parameters, where the final classifier is obtained by averaging the coefficients of the combined classifiers. Since both SVM and LR support estimated probabilities, the average of estimated probabilities' strategy has been used. Each training set (bootstrap) generates estimated probabilities $\hat{p}(j/x)$, which is an object with prediction vector x belonging to class j. Then the class corresponding to $\boldsymbol{x}$ is estimated as arg max$j$ $\hat{p}(j/x)$. The bagging ensemble is obtained by averaging the $\hat{p}(j/x)$ over all bootstrap replications to obtain $\hat{p}_{Be}(j/x)$, and then uses the estimated class arg max$j$ $\hat{p}_{Be}(j/x)$ as a final prediction. This estimate was computed in all the classification examples in this paper. The resulting misclassification rate was always virtually identical to the voting misclassification rate [15].

3.3 Performance measures

Central to constructing, deploying, and using classification method is the question of method performance assessment. The support vector machine and logistic regression are now used in many domains, and different performance measures are appropriate for each domain. The different performance metrics measure different tradeoffs in the predictions made by algorithm and it is possible for learning algorithm to perform well on one metric, but be suboptimal on other metrics, because of this it is important to evaluate algorithms on a broad set of performance metrics. Therefore, in this comparative study a variety of performance metrics has been used. The performance metrics were divided in to three. The first one is the common metrics that are well known and have been widely used in machine learning

comparisons which are threshold metrics. The default is 0.5. These metrics only consider the prediction above or below the threshold (0.5). These metrics are: accuracy (ACC), the number of correct predictions on the test data is divided by the number of test data instances, sensitivity (SN), specificity (SP): assesses the effectiveness of the algorithm on positive and negative classes respectively, F-score is a composite measure benefits algorithm with higher sensitivity and challenges algorithm with higher specificity, precision is the assessment of the predictive power of the algorithm for positive or negative classes. Secondly, new suggested metrics other than those common metrics have been used to assess the performance of the algorithm. The goal of using these metrics is to evaluate the performance of the algorithm in other ways. These new suggested measures are used in the medical area, and they are: Youden's index (γ), likelihoods ratio (LR) and diagnostic odds ratio (DOR).

Youden's index (γ) (1950) measures the ability of an algorithm to avoid failure. It equally weights the algorithm's performance in negative and positive examples, it can be expressed as:

$$\gamma = senstivity - (1 - specitificity) \tag{20}$$

A high value of γ indicates better ability to avoid failure [27].

Positive and negative likelihoods (LR_S) [28] are familiar epidemiologic measures, used to select appropriate diagnostic test and are useful and helpful for comparing two algorithms. Their advantages over sensitivity and specificity are to evaluate the algorithm's performance with respect to both classes. The values of the positive ($\rho+$) and negative ($\rho-$) can be expressed as:

$$\rho+ = \frac{sensitivity}{(1 - specificity)}, \quad \rho- = \frac{(1 - sensitivity)}{specificity} \tag{21}$$

A higher positive likelihood and a lower negative likelihood indicate better performance on positive and negative classes respectively [28]. Here it should be mentioned that if $\rho+ < 1$ likelihood metrics should not be used. The relationship between the likelihoods and the performance of the two algorithms A and B is as follows [28]:

$$\begin{aligned}
&\text{if } \rho_+^A > \rho_+^B \text{ and } \rho_-^A < \rho_-^B \quad \text{implies A is superior over all.}\\
&\text{if } \rho_+^A < \rho_+^B \text{ and } \rho_-^A < \rho_-^B \quad \text{implies A is superior for confirmation of negative examples.}\\
&\text{if } \rho_+^A > \rho_+^B \text{ and } \rho_-^A > \rho_-^B \quad \text{implies A is superior for confirmation of postive examples.}\\
&\text{if } \rho_+^A < \rho_+^B \text{ and } \rho_-^A > \rho_-^B \quad \text{implies A is inferior overall.}
\end{aligned}$$

Diagnostic odds ratio (DOR) [29] is also a global performance measure. It has been suggested as a superior measure of diagnostic discrimination and it is used in medicine for the comparison of diagnostic accuracies between two or more diagnostic tests. Similarly this measure can be used in machine learning to measure the algorithm's performance and compare them. It evaluates how the algorithm distinguishes between positive and negative examples. It is calculated using the following equation:

$$\text{DOR} = \frac{sensitivity/(1 - sensitivity)}{(1 - specificity)/specificity} \tag{22}$$

Combining these three metrics with the common metrics helps to obtain balanced evaluation of the algorithm's performance. Thirdly to assess the algorithm's performance with respect to their estimated probabilities, the area under the ROC (receiver operating characteristic) curve (AUC) is used [30, 31] which compares visually the algorithm's performance averaged across all possible probability thresholds. The ROC curve plots observed sensitivity versus (1-specificity) for all possible classification thresholds. It also measures the ability of the algorithms to separate the instances of the different classes. The power of the ROC curve comes from the fact that it characterizes the performance of a classification model as a curve rather than a single point. The important statistical property of (AUC) is that it is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance which is equivalent to the Mann-Whitney statistic [32]. A high value of the statistic test indicates that the probability ranking is generally better. Thus we used the area under the ROC curve for comparing class probability estimators of the two algorithms. To test the statistical difference between each pair of ROC curves of the two algorithms, the Wilcoxon test as an appropriate nonparametric test was used [33, 31].

3.4 Statistical comparison methods

Although there is no specific procedure for comparing algorithms over multiple data sets, there are different sta-

tistical tests and common-sense techniques to test whether the two algorithms are significantly different or not. The key question of using the statistical test is the suitability and the assumptions that should be satisfied. In this paper, the paired sample T test as a parametric test and the wilcoxon signed-ranks test as a nonparametric test have been used.

3.4.1 Paired sample T test

The paired T test is used to compare two population means where there are two samples in which observations in one sample can be paired with observations in the other sample. It is used in this paper to test the statistical difference between the two algorithms over the various evaluation measures. The hypothesis is whether the average difference in their Performance over the data sets is significantly different or not. Let $c1j$ and $c2j$ be the metric scores of the two algorithms on the j-th data set and let dj be the difference $c2j - c1j$. The T statistics is computed as $(\bar{d}/s\bar{d})$ and is distributed according to T distribution with $N-1$ degrees of freedom, where N is the number of the data sets. Paired T test is true and can be safely used even if the variances of the two random variables under comparison are not homogeneous. However it may be less effective if the two random variables under comparison are not distributed normally [34]. In addition, the Paired T test requires the minimum sample size (number of the data sets) to be ~ 30. To ensure that the variable is normally distributed, there are many tests that can be used such as the Kolmogorov–Smirnov test. All these tests that are used for checking the normality assumption are affected by the sample size; therefore it would be useless to check the normality of the samples that are less than 30.

3.4.2 Wilcoxon signed-ranks test

The alternative test to the paired T test is the Wilcoxon signed-ranks test which is a nonparametric test. It does not need the assumptions of homogeneity and normality and it will not be affected by the sample size [35]. Therefore, it can be appropriated when the paired T test's assumptions are violated. The Wilcoxon signed-ranks test ranks the differences in performance measurement of the two algorithms for each data set, ignoring the signs, and compares the ranks for the positive and the negative differences. The differences are ranked according to their absolute values; average ranks are assigned in case of ties. Let $R+$ be the sum of ranks for the data set, in which the second algorithm outperformed the first. Let $R-$ be the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the two sums. Let $T = \min(R+, R-)$, then the test statistics is computed as follows:

$$Z = \frac{T - \frac{1}{4}(N(N+1))}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \tag{23}$$

The statistics in (23) is distributed approximately normally. The null hypothesis is rejected if $Z > Z_{(\alpha/2)}$.

4 Experimental set up

The bagging method is used to construct the experiment. As shown previously, a random sample (bootstrap) was drawn with replacement from the original data set to form a training set. Each training set contains approximately 66% of the data from the original data set. The support vector machine (SVM) with the Gaussian (RBF) kernel and LR were used in this experiment for classification. With the methods like cross-validation, the relevant parameters of SVM can be chosen more scientifically, so they are widely used to choose the optimal parameters for SVM [36], hence, with each bootstrap 10 fold cross validation (CV) was used to determine the best values of γ and C [37]. Normally Cross validation (CV) is used to estimate the generalization capability on new samples that are not in the training dataset. A k-fold cross validation randomly splits the training dataset into k approximately equal-sized subsets, leaving out one subset, builds a classifier on the remaining samples, and then evaluates classification performance on the unused subset [38–40]. This process was repeated k times for each subset to obtain the CV performance over the training dataset. The best values were determined for each training set. The training set with its associated best values was used to construct the support vector machine model. This model generated estimated probabilities $\hat{p}(j/x)$. This procedure repeated 100 times. Finally we have 100 SVM classifiers. They were combined by taking the average of the $\hat{p}(j/x)$ to obtain $\hat{p}_{Be}(j/x)$. The same procedure was used for LR to construct 100 models. For the categorical variables we have deleted any training data set that does not include all the categorical variables, so all the testing and training data sets include all the categorical variables. The 2-way interactions between the independent variables were added to the model, the correlation and collinearly were checked before the analysis. The quartile method was used to assess the relationship between the continuous variables and the outcome to check whether the categorization for continuous variables was needed. The backward selection procedure was used with 0.05 as the default significance level. After the estimated class arg max$j$ $\hat{p}_{Be}(j/x)$ was calculated for support vector machine and logistic regression, the various evaluation

metrics were computed for both of them. The ROC curves were constructed using $\hat{p}_{Be}(j/x)$. The paired T test and the Wilcoxon signed-ranks test were applied to the all performance measures. The probabilities multiplication rule was used to combine the results of the paired T test and the Wilcoxon signed-ranks test on those performance measures, to obtain the final decision.

5 Results and discussions

The results of the support vector machine were carried out by using LIBSVM (3.0–1) [41] software package, available at http://www.csi.ntu.edu.tw/~cjlin/libSVM under matlab (7.8.0347- R2009a) interface, while the results of multiple logistic regression were carried out using spss.16.0 (SPSS Inc, Chicago, IL, USA). The statistical tests were also calculated by using spss.16.0.

5.1 Performances by measures

The results of the support vector machine and logistic regression on the data sets for each common and new suggested performance measures are shown in Tables 2 and 3 respectively. For each table the first six rows represent the results for the balanced data sets. Each metric value in these tables represents the average of the 100 classifiers for the corresponding data set. The area under

Table 2 The results of the performance measures for SVM

Data set	The performance measures									
	Accuracy	Sensitivity	Specificity	F-score	Precision	AUC	γ	$\rho+$	$\rho-$	DRP
Cod (sample)	0.861	0.944	0.777	0.872	0.811	0.921	0.72	4.221	0.073	58.061
Chest	0.997	0.997	0.995	0.997	0.996	1.000	0.993	217.62	0.002	108,810
Contra	0.731	0.844	0.579	0.782	0.729	0.783	0.422	2.000	0. 270	7.4074
Credit	0.875	0.888	0.865	0.866	0.845	0.946	0.753	6.569	0.129	50.9226
Liver	0.768	0.662	0.845	0.706	0.755	0.828	0.507	4.270	0.399	10.7018
Heard	0.900	0.946	0.842	0.913	0.882	0.943	0.788	5.979	0.063	94.9049
Page	0.960	0.977	0.810	0.978	0.978	0.983	0.788	5.163	0.028	184.393
Spam	0.927	0.882	0.957	0.905	0.930	0.978	0.839	20.322	0.123	165.220
Car	0.990	0.996	0.984	0.986	0.989	0.999	0.980	297.77	0.016	18,610.6
Breast	0.980	0.980	0.978	0.971	0.959	0.995	0.960	43.656	0.017	2,568.00
Diabetes	0.839	0.668	0.932	0.744	0.840	0.910	0.599	9.822	0.356	27.5899
Number	0.871	0.670	0.957	0.757	0.870	0.914	0.627	15.630	0.345	45.3044
Ionosphere	0.966	0.964	0.968	0.973	0.982	0.994	0.932	30.380	0.037	821.081

Table 3 The results of the performance measures for LR

Data set	The performance measures									
	Accuracy	Sensitivity	Specificity	F-score	Precision	AUC	γ	$\rho+$	$\rho-$	DRP
Cod (sample)	0.953	0.962	0.944	0.954	0.945	0.991	0.906	17.037	0.040	425.925
Chest	0.981	0.981	0.981	0.982	0.983	0.997	0.962	51.646	0.020	2,582.3
Contra	0.726	0.826	0.590	0.776	0.731	0.771	0.419	2.029	0.294	6.901
Credit	0.883	0.922	0.851	0.877	0.837	0.945	0.772	6.176	0.092	67.130
Liver	0.760	0.676	0.825	0.710	0.737	0.822	0.501	3.860	0.393	9.822
Heard	0.900	0.933	0.858	0.912	0.891	0.951	0.791	6.588	0.077	85.558
Page	0.962	0.992	0.696	0.979	0.966	0.975	0.689	3.269	0.011	297.182
Spam	0.932	0.894	0.957	0.912	0.931	0.979	0.851	20.772	0.111	187.135
Car	0.960	0.969	0.935	0.920	0.910	0.991	0.900	30.573	0.067	456.313
Breast	0.974	0.966	0.978	0.963	0.958	0.997	0.944	42.913	0.034	1,262.147
Diabetes	0.789	0.610	0.886	0.668	0.741	0.858	0.494	5. 335	0.442	11.312
Number	0.793	0.543	0.903	0.610	0.703	0.827	0.439	5.520	0.513	10.760
Ionosphere	0.952	0.973	0.913	0.963	0.952	0.991	0.886	11.149	0.029	384.448

the curves (AUC), shown in these tables, were calculated using the average of the estimated probabilities among all of the 100 classifiers.

5.2 The ROC curve analysis

As described above, the average of the estimated probabilities has been used to construct the ROC curves for SVM and LR. Figures 1 and 2 show the ROC curves for credit approval and Pima Indian diabetes as an example of balanced and unbalanced data sets respectively. From Fig. 1, the two ROC curves for the credit approval for the SVM and LR are almost the same, which is the situation in most of the balanced data sets. From Fig. 2, the ROC curve for the Pima Indian diabetes is higher for the SVM; however, some of the other unbalanced data sets have almost similar curves for SVM and LR. The relationship between the area under these ROC curves of SVM and LR is depicted in

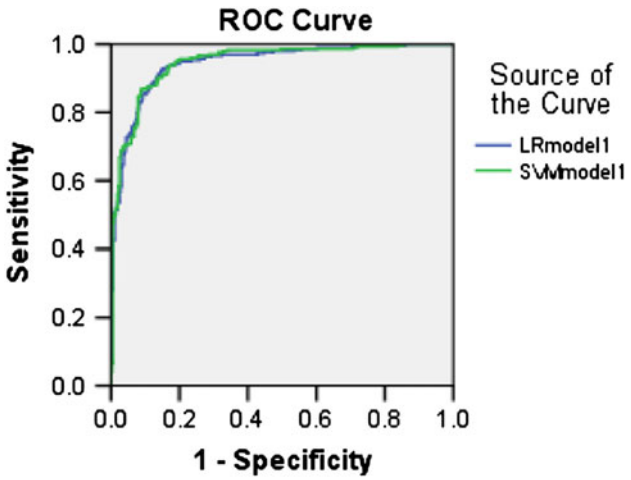

Fig. 1 ROC curve for credit approval

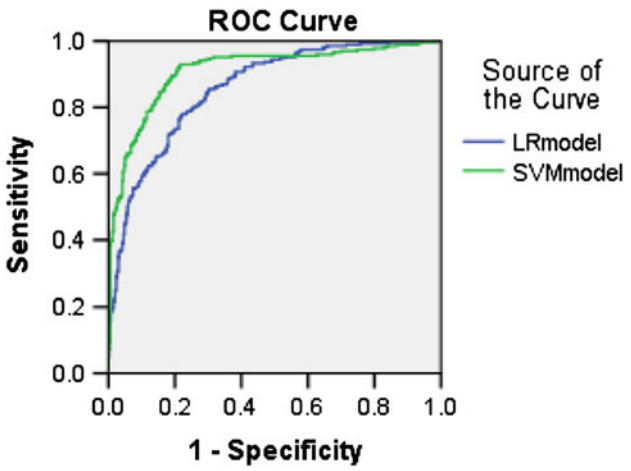

Fig. 2 ROC curve for Pima Indian

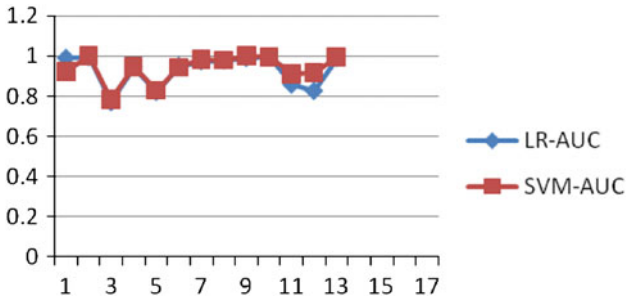

Fig. 3 The relationship between RUCs of SVM and LR for the data sets

Fig. 3. Figure 3 represents the AUCs of each data set, using the same order in Table (1) for SVM and LR; it shows that the most pairs of AUCs of SVM and LR for each data set lie closely.

The Wilcoxon signed-ranks test with $\alpha = 0.05$ for the balanced data sets shows no significant difference between the ROC curves of the SVM and LR since all the p values are in the range (0.078, 0.475). However, for the unbalanced data sets, it shows significant difference for the German number and page block data sets, where their p values are less than 0.025.

5.3 The statistical tests analysis

The Paired T test and the Wilcoxon signed-ranks test are used to see whether the two algorithms perform equally well or not. Because we have no guarantee for normality assumption and also because of the relatively small samples number, we applied both the paired T test and the Wilcoxon signed-ranks test. The results are shown in Tables 4 and 5. Each value in these tables represents the p value of the corresponding measure among all the data sets.

Multiplication rule [42] is used to combine theses results, since all these tests are independent, this rule can be used to obtain one p value to make the final statistical decision. The formula of this rule, in case of independency is as follows:

$$p\left(\bigcap_{i=1}^{n}\right) = p\left(\prod_{i=1}^{n}\right) \tag{24}$$

This means that the probability of no statistically significant difference between the algorithms is equivalent to the probability of no statistically significant difference between all their performance measures. The p value is defined as the probability of H_0 is true, thus according to the above rule the p value for no significant difference between the two algorithms is equal to the product of all the p values. The p value for the Paired T test is (5.387×10^{-6}), while the p value for the Wilcoxon signed-ranks test is

Table 4 Paired T test's results

Measure	Accuracy	Sensitivity	Specificity	F-score	Precision	RUC	γ	$\rho+$	$\rho-$	DRP
p value	0.486	0.280	0.475	0.274	0.282	0.450	0.283	0.155	0.211	0.259

Table 5 Wilcoxon signed-ranks test's results

Measure	Accuracy	Sensitivity	Specificity	F-score	Precision	RUC	γ	$\rho+$	$\rho-$	DRP
p value	0.158	0.382	0.130	0.263	0.093	0.141	0.124	0.075	0.294	0.196

(1.4×10^{-8}). Similarly, α is the probability of rejecting H_0 when it is actually true. So the α value of testing no significant difference between the two algorithms is equal to $(0.025)^{10} = (9.536743 \times 10^{-17})$.

5.4 Discussion

The results show that using several performance measures with different data sets can help in understanding and comparing the performance of the algorithms. Moreover, the results show that, it is not always reliable to compare algorithms using their performance measures scores only. Regarding, the comparison that has been done in this paper, for the balanced data sets, it has been found that support vector machine and logistic regression have much close overall performance measures in most of the data sets. For the Hared–Scale data set, which is the smallest one, the two classifiers performed equally in accuracy, sensitivity, specificity, precision, F-score and AUC. This indicates that the two algorithms can perform equally well in the small data sets. Also the results obtained, using the unbalanced data sets, show that overall the common performance measures is almost the same for the support vector machine and logistic regression. However, support vector machine achieves higher values in some unbalanced data sets. For the semi unbalanced spam data set, the two classifiers performed equally well. In the highly imbalanced data sets (German number and page block), logistic regression is found to be biased towards the majority class. However, this would not have a major effect on the general algorithms' performance. When comparing the ROC curves, for the results of the balanced data sets, we found the minimum p value is (0.078) for the credit approval dataset. This indicates that there is no significant difference between the two classifiers on these data sets. While for the comparison of the ROC curves, the results of the unbalanced data sets show the only significant difference between the two classifiers is found when we used German number and page block data sets, because their p values are less than 0.025. Generally, according to the results of ROC curves, the two classifiers have equal performance. However, support vector machine outperforms LR in the highly unbalanced data sets. Because there is no evidence that our sample satisfies the normality assumption, both Paired T test and the Wilcoxon signed-ranks test are used. The results of the Paired T test with $\alpha = 0.05$ show that there is no significant difference in the overall performance measures. Also the results of the Wilcoxon signed-ranks test with $\alpha = 0.05$ show that there is no significant difference in the overall performance measures. Moreover, the general p values of the Paired T test and the Wilcoxon signed-ranks test are higher than the level of significance (α) for rejecting the null hypothesis. This indicates that there is no statistical significant difference between the SVM and LR, and both of them perform equally well.

6 Conclusion

This study has empirically compared two familiar classifiers; support vector machine and multiple logistic regression using bagging and ensemble over various different sizes of balanced and unbalanced data sets. The comparison was done in a different manner than the manner of most machine learning comparisons. This study represents a standard comparison. It includes numerous statistical analyses for several algorithm performance measures which enable us to make a warranted and verified conclusion. This study shows that, generally, the SVM and LR over all the performance measures have equal performance for balanced and unbalanced data. However, support vector machine may work better for the highly unbalanced data sets. The study also views that there are some measures higher in one classifier than in the other in some data sets, consequently, it is not appropriate to draw a conclusion from studies with one data set, that one classifier is better than the other. There is no golden standard for making such comparisons and the tests that are performed often have no statistical foundations. Logistic regression has higher interpretability while support vector machine is considered to be a black box predictor. It neither makes its prediction implicit nor gives

incite in the rules governing its prediction, which is not the case in LR. Therefore, in the case of considering classification only, each of them can be used while when the interpretation is necessary such as in many medical studies, logistic regression should be used.

Acknowledgments This work was supported by a grant from Hebei University, Baoding, Hebei, P.R. China. I wish to thank the PhD students of the departments of computer Sciences and mathematics for their encouragement, useful discussions, and interest.

References

1. Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley series in probability and statistics, Wiley, Inc, New York
2. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models, 4th edn. Irwin, Chicago
3. Wang L (ed) (2005) Support vector machines theory and applications. Springer, Berlin
4. Kecman V (2001) Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. MIT, Cambridge
5. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
6. Perlich C, Provost F, Simonoff JS (2003) Tree induction vs. logistic regression: a learning-curve analysis. J Mach Learn Res. doi:10.1162/153244304322972694
7. King RD, Feng C, Sutherland A (1995) Statlog: comparison of classification algorithms on large real-world problems. Applied Artif Intell 9(3):289–333
8. Muniz AMS, Nadal J, Liu H, Liu W, Lyons KE, Pahwa R (2010) Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait. J Biomech 43(4):720–726
9. Xu L, Chow M-C, Gao X-Z (2005) Comparisons of logistic regression and artificial neural network on power distribution systems fault cause identification. Proceedings of 2005 IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications (SMCia/05)
10. Chen W-H, Shih J-Y, Wu S (2006) Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets. Int J Electron Fin 1(1):49–67. doi:10.1504/IJEF.2006.008837
11. Song JH, Venkatesh SS, Conan EA (2005) Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. Acad Radiol 12(4):487–495
12. Verplancke T, Van Looy S, Benoit D, Vansteelandt S, Depuydt P, De Turck F, Decruyenaere J (2008) Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with hematological malignancies. BMC Med Inform Decis Mak 8:56. doi:10.1186/1472-6947-8-56
13. Kuncheva LI (2004) Combining pattern classifiers methods and algorithms. Wiley, Hoboken
14. Zhang L (2011) Sparse ensembles using weighted combination methods based on linear programming. Pattern Recogn 44(1):97–106
15. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
16. Small K, Roth D (2010) Margin-based active learning for structured predictions. Int J Mach Learn Cybernet 1(1–4):3–25
17. He Q, Wang X, Chen J, Yan L (2006) A parallel genetic algorithm for solving the inverse problem of support vector machines. ICMLC 2005 LNAI 3930:871–879
18. Wang X-Z, He Q, Chen D-G, Yeung D (2005) A genetic algorithm for solving the inverse problem of support vector machines. Neurocomputing 68:225–238
19. He Q, Congxin Wu (2011) Separating theorem of samples in Banach space for support vector machine learning. Int J Mach Learn Cybernet (IJMLC) 2(1):49–54
20. Sathiya Keerthi S, Lin C-J (2003) Asymptotic behaviors of support vector machines with Gaussian Kernel. Neural Comput 15(7):1667–1689
21. Zhang S, McCullagh P, Nugent C, Zheng H, Baumgarten M (2011) Optimal model selection for posture recognition in home-based healthcare. Int J Mach Learn Cybernet (IJMLC) 2(1):1–14
22. Wang X-Z, Shu-Xia Lu, Zhai J-H (2008) Fast fuzzy multi-category SVM based on support vector domain description. Int J Pattern Recognit Artif Intell 22(1):109–120
23. Kuss O (2002) Global goodness-of-t tests in logistic regression with sparse data. Statist Med 21:380–3789
24. Dietterich TG (2000) Ensemble methods in machine learning. Lecture Notes in Computer Science, vol. 1857, pp. 1–15. doi:10.1007/3-540-45014-9_1
25. Valentini G, Dietterich TG (2004) Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. J Mach Learn Res 5:725–775
26. Valentini G, Dietterich TG (2003) Low Bias Bagged Support Vector Machines. Machine Learning, Proceedings of the Twentieth International Conference (ICML) Washington, DC, USA, pp 752–759
27. Japkowicz N, Szpakowicz S (2006) Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. AI 2006: advances in artificial intelligence. LNCS 4304:1015–1021. doi:10.1007/11941439_114
28. Pereira BdeB, Pereira CAdeB (2005) A likelihood approach to diagnostic tests in clinical medicine. Revstat Stat J, Lisboa 3(1):77–98
29. Glasa AS, Lijmer JG, Bossuyta PMM (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56:1129–1135
30. Bradley AP (1997) The use of the area under the roc curves in the evaluation of machine learning algorithms. Pattern Recogn 30(7):1145–1159
31. Avergara I, Norambuena T, Ferrada E, Slater AW, Melo F (2008) A simple tool for the statistical comparison of ROC curves. BMC Bioinform 9:265
32. Bamber D (1975) The area above the ordinal dominance graph and the area below the receiver operating graph. J Math Psychol 12(4):387–415
33. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
34. Montgomery DC (2001) Design and analysis of experiments, 5th edn. Wiley Inc, New York, pp 21–54
35. Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
36. Liu Z, Wu Q, Zhang Y, Philip Chen CL (2011) Adaptive least squares support vector machines filter for hand tremor canceling in microsurgery. Int J Mach Learn Cyber 2(1):37–47
37. Hsu C-W, Chang C-C, Lin C-J (2010) A practical guide to support vector classification. Citeseer 1(1):1–16
38. He Q, Congxin Wu (2011) Membership evaluation and feature selection for fuzzy support vector machine based on fuzzy rough sets. Soft Comput 15(6):1105–1114

39. Stone M (1974) Cross-validatory choice and assessment of statistical prediction. J Royal Stat Soc B 36:111–147
40. Mahmood Z (2009) On the use of K-fold cross-validation to choose cutoff values and assess the performance of predictive models in stepwise regression. Int J Biostat 5(1), Article 25
41. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27
42. Mood G (1974) Introduction to the theory of statistics, 3rd edn. McGraw Hill, New York, pp 2–32