Scientific Reports in Medicine, 2024; 1(3): 120-125

Scientific Reports in Medicine

Evaluation of machine learning methods in medicine: real data application

Running title: machine learning methods in medicine

Hülya Binokay[1], Yaşar Sertdemir[2]

DOI: 10.37609/srinmed.25

Abstract: **Objective:** One of the aims of a health study is to identify risk factors associated with the disease or to obtain predictive models for classification such as healthy / diseased. When the aim of a health study is classification, machine learning methods are widely used. Some of these methods; Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and Naive Bayes. The aim of this study was to evaluate the performance of the machine learning such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and Naive Bayes, for different sample size, prevalence and determination coefficient in real data sets.

Method: The data were randomly split into 70% training and 30% test set, and Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and Naive Bayes were applied to the training set. The performance measure (Accuracy, Area Under Curve and Adjusted F Measure) of the methods evaluated on the test set were saved. This procedure was repeated 1000 times.These procedures were performed in the R 3.5 1.

Results: When all variables in the data are categorical, and determination coefficient is low with a moderate sample size, the Naive Bayes method exhibited higher performance. When all variables in the data are continuous, and determination coefficient is moderate with a low sample size, support vector machines method demonstrated superior performance. In cases where the dataset has a high number of categorical variables and a high determination coefficient, the Naive Bayes method outperformed others. The Random Forest method showed higher performance when determination coefficient is high, and the sample size is moderate.

Conclusion: This study provides valuable insights for researchers dealing with classification problems, guiding them to choose the most effective machine learning based on the characteristics of the datasets.

Keywords: Binary Logistic Regression, Random Forest, Support Vector Machine, Naive Bayes, Decision Tree, Real Data Sets

[1]Cukurova University Faculty of Medicine, Department of Biostatistics, Adana, Türkiye
email: hulyabinokay@gmail.com
ORCID iD: 0000-0002-0162-4574

[2]Yaşar Sertdemir: Cukurova University Faculty of Medicine, Department of Biostatistics, Adana, Türkiye
email: yasarser@cu.edu.tr
ORCID iD: 0000-0003-4455-3590

Recieved: xxxxxxxxxx
Accepted: xxxxxx

3023-8226 / Copyright © 2024 by Akademisyen Publishing. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

INTRODUCTION

Classification is a type of problem in machine learning (ML) that is commonly addressed using methods such as Random forest (RF) and Support vector machines (SVM) in areas like marketing, telecommunications, and medicine.[1]

Among the ML models mentioned above, Logistic regression (LR) is one of the fundamental methods in classifying binary (alive/dead, patient/control) groups. Although LR is widely used, the use of other ML models has become widespread recently. Some of these methods are Decision Tree (DT), Artificial Neural Networks, K-nearest neighbor, Ensemble Methods (Bagging, Boosting and RF), Naive Bayes, SVM[2].

As in many other areas, decisions play an important role in medicine, especially in medical diagnostic processes. Since conceptual simple decision-making models that are capable of ML models should be considered for performing such tasks, DT is a very proper candidate.[3] The DT is potent ML model that has been used successfully in many medical studies as it provides easily understandable graphical classification rules.[3] However, in the RF, which is one of the commonly used ensemble learning methods, each tree is built based on recursive partitioning, and the prediction is made on the average of an ensemble of trees rather than of a single tree.[4]

The NB is simple probabilistic ML model based on Bayes' theorem with the assumption of independence between variables.[5]

The SVM is a ML model based on the statistical learning theory developed by Vapnik.[6] SVM and LR use both linear and non-linear data to separate the two groups, but SVM classifies non-linear data better than logistic regression because it uses kernel functions. LR generates the linear decision boundary through logit transformation. SVM finds the linear hyperplane that provides the maximum margin. Therefore, SVM is more optimal than logistic regression as the margin is maximized.

The most commonly used performance criteria for evaluation of ML models in the literature are Accuracy (ACC), Area Under Curve (AUC) and Adjusted F Measure (AGF).

The aim of this study was to evaluate the performance of the ML models such as LR, DT, RF, SVM and NB, for different sample size (n), prevalence (prev) and determination coefficient (R^2) in real data sets.

METHOD

Binary Logistic Regression

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. Generally, logistic regression model is the case where the outcome variable is discrete by taking two or more possible values. The difference between an LR model and a linear regression model is that the outcome variable in LR is binary or dichotomous.[7] LR can be used for classification as well as for determining significant risk factors.

2.2. Decision Tree

DT is a non-parametric used for classification.[8] It consists of four parts, which are the decision node, the root node, leaf node, and branches.[9] In this structure, decision nodes represent the splitting measure on explanatory variables, leaf nodes represent a class label, and the root node represents the starting variable of the tree. Branches connect the nodes.

2.3. Random Forest

Breiman (1999) proposed RF, which combines the Random Subspace algorithm with the Bootstrap method.[11] Each DT was constructed from a set obtained from the starting training set using a bootstrap.[12] Ho (1998) has written many papers on "the random subspace" method, which does a random selection of a subset of features to use to grow each tree[13] .

2.4. Naive Bayes

NB is based on the assumption that the variables are conditionally independent[14]. This assumption is called class conditional independence. This assumption is made to simplify the computations involved, hence is called "naive". Despite this unrealistic assumption, the resulting classifier known as naive Bayes is remarkably successful in practice, often competing with much more sophisticated techniques.[15]

2.5. Support Vector Machine

SVM is an ML model based on the statistical learning theory developed by Vapnik (1998). SVM aims to find a maximal margin hyperplane to separate classes. The kernel function is used to map data to a higher dimensional space for learning non-linearly separable functions. The accuracy of the SVM largely depends on the properly chosen kernel and its parameters.[16]

The kernel function can be linear, radial, and polynomial functions. The Radial basis function is affected by the kernel width (γ) and the regularization (C) parameters; therefore, determination of the best pairs of parameters for the study was carried out.[17] The tune parameters for RF and SVM were automatically selected using the Caret package. Analyses were performed using R 3.5.1.

Real Data Study

ML models are tested on data sets from the UCI machine learning repository, including Breast Cancer[18], Breast (Breast Cancer coimbra)[19], Indian diabet pima [20], diabet[21], heart[22], Chronic kidney disease (CKD)[23]. The data were randomly split into 70% training and 30% test set, and the performance criteria of the methods in the test set were recorded. This procedure was repeated 1000 times. These procedures were performed in the R 3.5 1.

Performance Measures

In literature, performance evaluation of ML models is usually based on one performance measure. However, using these criteria, the performance of the methods is evaluated separately. in this evaluation, different evaluations can be made according to each performance criterion. for example, the method with the best performance for accuracy may have the worst performance according to the sensitivity value. In this case, it becomes difficult to determine which method performs better. To overcome this situation, ACC, AUC and AGF are evaluated together in this study.

The standard F measure has some limitations, especially in classification problems with class imbalance or significant differences between classes. The F-measure is defined as the harmonic mean of precision and recall and is often used to evaluate classification models. However, in some cases this metric may not provide sufficiently meaningful results. These tend to over-emphasize the majority class in imbalanced datasets. For example, in a dataset with 95% negative instances and 5% positive instances, a model that correctly classifies only the negative class may still have a high F-measure value, which may misrepresent the performance of the model. Therefore, the adjusted F-measure is used.

This evaluation is the mean performance measures were calculated for each ML model and ordered from largest to smallest and scored from 5 to 1. By summing the scores on each performance measure a final score was obtained. Table 1 shows how the ACC, AUC and AGF performance measures are calculated.

Table 1. Basic 2x2 Count Table			
Disease	**Test results**		
	Positive (*T=1*)	Negative (*T=0*)	Total
Present (D=1)	(True Positive)	(False Negative)	
Absent (D=0)	(False Positive)	(True Negative)	
Total			N

$$\text{Sn(Recall)} = P(T=1|D=1) = s_1 / n_1 \qquad \text{Sp} = P(T=0|D=0) = r_o / n_0$$

$$\text{PPV(Precision)} = P(D=1|T=1) = s_1 / m_1 \qquad \text{NPV} = P(D=1|T=1) = s_1 / m_1$$

$$\text{ACC} = (s_1 + r_0)/N$$

$$F_2 = 5 * \frac{Sn * precision}{(4*Sn) + precision} \qquad \text{Inv } F_{0.5} = \frac{5}{4} * \frac{Sn * Precision}{(0.25*Sn) + Precision} \qquad AGF = \sqrt{F_2 * \text{Inv}F_{0.5}}$$

RESULTS

The performance criteria of the ML models were evaluated using real data sets. The performance scores and properties of the real data sets are given in Table 2.

Table 2 Properties and performance scores of ML models											
	Properties of data sets						**Performance scores**				
Datasets	Prev	R^2	n	NV	#Cat	#Cont	LR	DT	RF	SVM	NB
Breast cancer	0.3	0.3	277	9	9	0	3	8	12	7	**15**
Breast cancer coimbra	0.6	0.4	116	9	0	9	5	6	12	**13**	9
Chronic kidney disease	0.3	0.8	158	24	13	11	3	6	13	9	**15**
Heart	0.3	0.6	299	12	5	7	3	11	**15**	8	8

NV: Number of variables, Cat: Number of Categorical variables, Cont: Number of Continuous variables

In scenarios where Prev=0.3, R2= (0.3, 0.8) and n= (158, 277), NB method has higher performance than other methods. In scenarios where the number of categorical variables in the data is high, the NB method has higher performance.In the scenario where prev=0.3, R2= 0.6 and n=299, RF method has higher performance than other methods, while in the scenario where prev=0.6, R2= 0.4 and n=116, RF and SVM methods have similar and higher performance than other methods. In scenarios where R2 is medium and high and the number of continuous variables in the data is high, RF method has higher performance.

DISCUSSION

Machine learning methods are used to classify diseased and healthy individuals in health studies. Correctly classifying diseased and healthy individuals is of great importance for early diagnosis of diseases and determining treatments for these diagnoses. There are many papers in literature investigating the performance of classification methods, but it is not clear which method performs better under which conditions. Given this situation, our aim in this paper is to evaluate the performance of classification methods on real data sets with n, prev and (R^2). Performance evaluation of ML models is based on one real data set, mostly two- or three-ML models

were compaired based on one or two and rarely three performance criteria. In this study, the performance of five ML models was evaluated based on ACC, AUC and AGF under real data sets. In this context, when all variables in the data were categorical, R^2 was low, and the sample size was moderate, the NB method demonstrated superior performance. When all variables in the data were continuous, and R^2 was moderate, and the sample size was low SVM method exhibited higher performance. When the number of categorical variables in the data was high, and R^2 was high, the NB method outperformed others. The RF method showed higher performance when R^2 was high, and the sample size was moderate to high.

Arasakumar et al. compared LR, DT, and RF on the breast cancer dataset and they observed that RF method shows better performance, which is consistent with our data[24].

Gokiladevi et al. compared SVM, RF, LR and DT on the chronic kidney disease dataset and observed that the performance of RF method shows better performance. This result is compatible with our real data[25].

Yu et al. compared DT, NB, RF and SVM according to the accuracy criteria, on breast cancer dataset and did not observe any significant difference[26].

Limitations of the study

More datasets can be used for comparisons, and different ML models can also be applied.

CONCLUSION

In conclusion, the performances of the data sets differ according to the structure of the data sets (n, r^2 and prev, continuous and categorical). Therefore, evaluating the data sets according to the characteristics of the data sets will enable us to make more accurate comments. We hope that this study helps any researcher confronted with classification problems to select the best performing two- or three-ML models based on the characteristics of the data set.

REFERENCES

Sharma S, Agrawal J, Sharma S. Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies. IJCA. 2013; 82: 20-27.

Ashari A, Paryudi I, Tjoa AM. Performance Comparison between Naive Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. IJACSA. 2013; 4: 33-39.

Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision Trees: an overview and their use in medicine. J. Med. System. 2002; 26:445–463.

Yoo W, Ference BA, Cote ML, Schwartz A. A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions. Int J Appl Sci Technol. 2012; 2: 268.

Zhang Z. Naive Bayes classification in R. Annals of Translational Medicine. 2016; 4: 241.

Vapnik VN. An overview statistical learning theory. IEEE transactions on neural networks. 1999; 10: 988-999.

Hosmer DW, Lemeshow S. Introduction to the logistic regression model. 2th ed. New York; 2000

Wang Y, Xia ST, Wu JA. Less-greedy Two-term Tsallis Entropy Information Metric Approach for Decision Tree Classification. Knowledge-Based Systems. 2016; 20: 2-28.

Nachiappan MR, Sugumaran V, Elangovan M. Performance of Logistic Model Tree Classifier using Statistical Features for Fault Diagnosis of Single Point Cutting Tool. INDJST. 2016; 9: 1-8.

Zhang Q, Sun J, Zhong G Dong J. Random multi-graphs: a semi-supervised learning framework for classification of high dimensional data. Image and Vision Computing. 2017; 60: 30–37.

Breiman L. Random forests. Machine Learning. 2001; 45: 5–32.

Polianchik DE, Grigor'ev VY, Sandakov GI, Yarkov AV, BachurinSO Raevskii. Binary Classification of Cns and Pns Drugs. Pharmaceutical Chemistry. 2017; 50: 800-804.

Pashaei E, Ozen M, Aydın N. Splice site identification in human genome using random forest. Health Technol. 2017; 7: 141-152.

Shelestov A, Lavreniuk M, Kussul N, Novikov A, Skakun S. Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping. Front. Earth Sci . 2017; 5: 1-10.

Rish I. An emprical study of the Naive Bayes classifier. Work Empir Methods Artif Intell. 2001; 3: 41-46.

Liua M, Wang M, Wang J, Li D. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage

and Chinese vinegar. Sensors and Actuators B. 2013: 970-980.

Tien Bui D, Anh TuanT, Klempe H, Pradhan B, Revhaug I. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides. 2016; 13: 361-378.

Schlimmer JC. Concept acquisition through representational adjustment. Doctoral dissertation, Department of Information and Computer Science, 1987. University of California, Irvine, CA.

Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci USA. 1990; 87: 9193-9196.

Smith, JW, Everhart JE, Dickson WC, Knowler WC, Johannes, R.S. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care. 1988. (pp. 261--265). IEEE Computer Society Press.

Kahn M. Diabetes [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5T59G.

Janosi A, Steinbrunn W, Pfisterer M, Detrano, R. Heart Disease [Dataset]. UCI Machine Learning Repository. 1989. https://doi.org/10.24432/C52P4X.

Rubini L, Soundarapandian P, Eswaran P. Chronic Kidney Disease [Dataset]. UCI Machine Learning Repository. 2015. https://doi.org/10.24432/C5G020.

Arasakumar M, Sudhakar P. An Effective Dynamic Weight Based Grey Wolf Optimization Algorithm with Support Vector Machine for Classification in Healthcare Industry. Science, Technology and Development. 2020; 9: 125-146

Gokiladevi M, Santhoshkumar SH. Gas Optimization Algorithm with Deep Learning based Chronic Kidney Disease Detection and Classification Model. International Journal of Intelligent Engineering & Systems; 2024:17(2).

Yu S, Li X, Wang H, Zhang X, Chen S. BIDI: A classification algorithm with instance difficulty invariance. Expert Systems With Applications. 2021; 165.