Using support vector machines in predicting and classifying factors affecting preterm delivery

Batoul Ahadi[1], Hamid Alavi Majd[1], Soheila Khodakarim[2], Forough Rahimi[3], Nourossadat Kariman[4],Mahieh Khalili[1], Nastaran Safavi[5]

[1]Department of Biostatistics, Para-Medical Faculty, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Department of Epidemiology,Health Faculty, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3]Department of English Language, Paramedical Faculty, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[4]Department of Midwifery, School of Nursing and Midwifery, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[5]Department of Midwifery, Ardabil Branch. Islamic Azad University, Ardabil, Iran.

*Corresponding Author: email address: alavimajd@gmail.com (H. Alavi Majd)

ABSTRACT

Various statistical methods have been proposed in terms of predicting the outcomes of facing special factors. In the classical approaches, making the probability distribution or known probability density functions is ordinarily necessary to predict the desired outcome. However, most of the times enough information about the probability distribution of studied variables is not available to the researcher in practice. In such circumstances, we need that the predictors function well without knowing the probability distribution or probability density. It means that with the minimum assumptions, we obtain predictors with high precision. Support vector machine (SVM) is a good statistical method of prediction. The aim of this study is to compare two statistical methods, SVM and logistic regression. To that end, the data on premature infants born at Tehran Milad Hospital is collected and used.

Keywords: support vector machines; logistic regression; premature birth

INTRODUCTION

Classical statistical methods based on restrictive assumptions are the same as probability distribution or probability density function. However, in many cases, sufficient information is not available about the probability distribution variables for the researcher. In such circumstances, we need some methods functioning well without knowing the probability distribution. On the one hand, with the advances of science and the explosion of information in many studies, we face data with very high dimensional spaces. To use classical statistical methods in such conditions, we require large samples that are sometimes impossible to provide. Today, with advances in science and technology, medical centers collect a wide range of data for a variety of purposes in which discovering useful and hidden patterns in this collection of information could lead to new medical knowledge. In recent decades, machine learning algorithm as a data mining tool is used to discover hidden patterns in medical data. Data mining was formed in the late 1980s, and in the 1990s, great steps were taken in this branch of science and it is expected to continue growing and developing in the present century [1].

In machine learning algorithm, we seek to provide a predictor with maximum accuracy with the least assumptions. A suitable method to identify classification pattern machine is the support vector machine which was presented in 1989 by Vepnick and Cheronkis. This classification method is binary and has multi-classes and in addition to classification, it is used in the regression analysis. Logistic regression is a

common statistical method used to predict and classify the dependent multimode variables which is the most useful method in medical studies.Preterm birth is one of the public health risks in the society causing death and complications during pregnancy. A delivery that occurs twenty weeks after the start of labor and before 37 weeks of pregnancy is called preterm birth[2]. An approximate incidence of preterm birth is 11-10% [3].70% of deaths in neonatal period are due to preterm delivery[4] .The incidence of preterm birth in developed countries is 5 to 7 percent, but it is estimated that this rate is higher in developing countries[5] Neurological problems for babies born prematurely are estimated to be up to 50 percent. The cause of preterm birth is complex, multi-factorial and not fully understood[6] .Premature birth can cause the placenta previa, placental abruption pregnancy, multiple pregnancy, cardiovascular disease, high blood pressure, and preeclampsia. Also, the bacterial infection may also play a role in the onset of preterm labor [7].The prediction or early detection of diseases, including predicting preterm birth, martial disorders during pregnancy, various cancers, diabetes, etc., and studying the factors affecting these is one of the issues in medicine. In this case, the researchers face a host of variables and characteristics of individual patients.

MATERIALS AND METHODS

Data quality is a factor that can be an effective tool for machine learning success. If there is duplication and ambiguous data, pattern discovery during the process would be difficult. Choosing a subset of variables is a process in which the variables having ambiguous or duplicated information are excluded from the model as much as possible.

In the classification method of SVM methods, different algorithms exist for selecting a subset of variables. Reducing the size of data which leads to rapid and efficient performance of the algorithm is among the benefits of choosing a subset of variables. It is even possible to increase the prediction of accuracy in the classification.

The algorithm used in this study to select a subset of variables is Wrapper. Wrappers use statistical methods for repeated sampling (e.g. mutual authentication) and by the desired learning algorithm, it chooses a subset of the variables that has the best accuracy in the prediction [8] .SVM is a relatively new method of learning by monitoring the topic of machine learning and a nonparametric method which was invented and developed by Vipnik. SVM is based on the statistical learning theory. Models made by this algorithm due to applying the principle of minimum structural error have a better performance than the neural networks and new similar methods theoretically. Furthermore, this algorithm has a high ability to model complex nonlinear patterns due to applying kernel. Then, the kernel function, by mapping samples to a space, has more dimensions than the samples input space. It also converts complex and non-linear patterns into the linear and easier patterns [9] .SVM is a binary classifier, separating two classes by the linear border and is dependent on generalized linear classifications family. The main purpose of these algorithms is finding the largest gap between the two classes and thus increasing the accuracy of classification, yet the generalization error is also reduced as much as possible [10] .In this method, it is assumed that the samples are labeled as $y_i = \{-1,1\}$. Each sample is shown as a vector. The maximum margin method is used to find the optimal decision boundary. Thus, the decision boundary should both divide examples of the two classes into two categories properly and find a hyper-plane (boundary decision) that has the most distance from all support vectors. The mathematical expression of decision boundary can be stated in the vector space as follows:

$$f(x) = sgn(w.x + b)$$

W is normal vector on the hyper-plane and b is the intercept[11]. As mentioned earlier, the samples should be properly categorized by boundary decision. The mathematical expression is as follows:

$$y_i = (w.x + b) \geq 1$$

On the other hand, the decision boundary should have the most distance with the examples of each class according to figure (1) meaning to maximize $\frac{2}{\|w\|}$ (11).

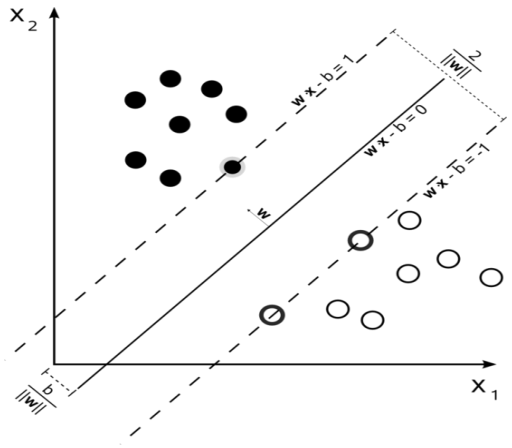


Figure 1. Classification by SVM in two classes by linear kernel in a two-dimensional space.

We can define optimization problems when the data is linearly separable as follows:

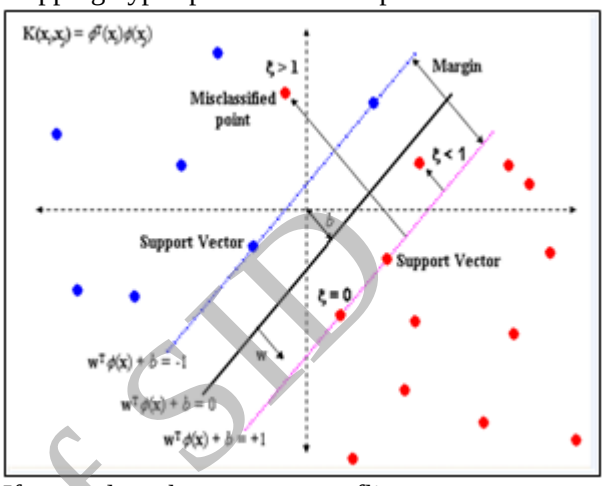
$$min\frac{1}{2}\|w\|^2 s.t\, y_i(w.x_i + b) \geq 1$$

Sometimes, in linear classification the classes may overlap and in this case, we can separate the samples linearly by excluding the consideration of some points. In this case, the variable named Slack (ζ_i) is defined as an amount equal to the distance between the exception point from the decision boundary. In this case, the objective function is as follows:

$$min\frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}\zeta_i$$

$$s.t\, y_i(w.x_{i+b}) \geq 1 - \zeta_i$$

C is a parameter that must be specified by the user and controls the fine imposed on the objective function for each exception [9].In many cases, not only may the classes overlap, but also they may separate functions in two classes having non-linear functions. This is the closest case to real cases. In this case, the input vectors into more dimensional space (feature space) are written as the input samples space. By increasing dimensions, it is generally possible to increase linear rating. SVM finds optimal decision boundary in the feature space. It is determined by mapping hyper-plane into the input.



If metadata has many conflicts, we can use polynomial kernels with polynomial degree and different gamma from RBF kernels and space of equation for decision boundary. The relationship among three kernels is used in Table 1 below.

Table 1. kernel function

linear	$K(x_i, x_j) = x_i^T x_j$
Polynomial	K(xi‹xj) = (gxiTxj + r)d‹g>0
RBF	K(xi‹xj) = exp(-g‖xi - xj‖2)‹g>0

The relationships shown in Table (1) includes “T” as matrix transpose, “g” as gamma, “d” as degree of polynomial, “xi and xj” as the i-th and j-th components of the vector. The non-linear kernels of SVM, gamma parameter controls the form of frontier in which small amounts causes the decision border to be close to the linear state and as the amount increases, the flexibility of the

decision border becomes higher and the decision border will be closer to the form of metadata for each class. Changing d parameter also makes the hyper-plane separator flexible. The current sample includes 600 pregnant women of 1-13

weeks who referred to prenatal care clinic of Milad super-specialist Hospital, Tehran. These women have characteristics of the studied subjects and provide written consent to the officials of the research.

During the nine months of pregnancy, the survey information was collected every three months. In this study, a questionnaire was used to collect the data. The first part of this form includes demographic and pregnancy characteristics, such as the period of time taking folic acid tablets and conducted tests in the first, second and third quarters of pregnancy; the second part includes the pregnancy characteristics, conducted tests and existing problems; the third part includes pregnancy characteristics, types of delivery, the weight of infant, and the common drugs used during pregnancy collected through interviewing and observing the samples.

Using a support vector machine algorithm, we could predict and classify factors affecting preterm birth and we could compare the results with the obtained results from our logistic regression in the present study.

RESULTS

In order to model the SMO algorithm, linear kernel is used. The accuracy of the model on the training data is 63.30% and on experimental data is 67.66%. The closeness of these accuracies shows the model stability and the absence of over-fitness in the training data. The sensitivity is 65% in the training data and it is 63% in the experimental data. Furthermore, the characteristics of the training data and experimental data are 67% and 61% respectively.

The area under the ROC curve in the training data and experimental data is 0.63 and 0.65 respectively which is shown in the curve below:

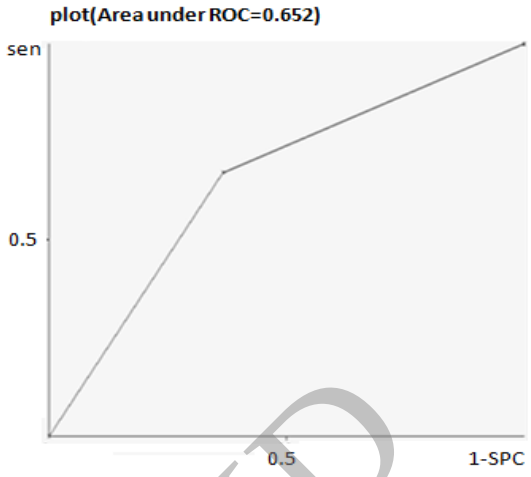


Figure 2. ROC curve in the SVM model on the experimental data

The accuracy of the model on the training data is 65.63% and on the experimental data is 56.54%.

The sensitivity is 72% in the training data and it is 50% in the experimental data. In addition, the characteristics amount of the training data and experimental data are 58% and 57% respectively.The area under the ROC curve in the training data and experimental data are 0.70and 0.60 respectively which is shown in the curve below:

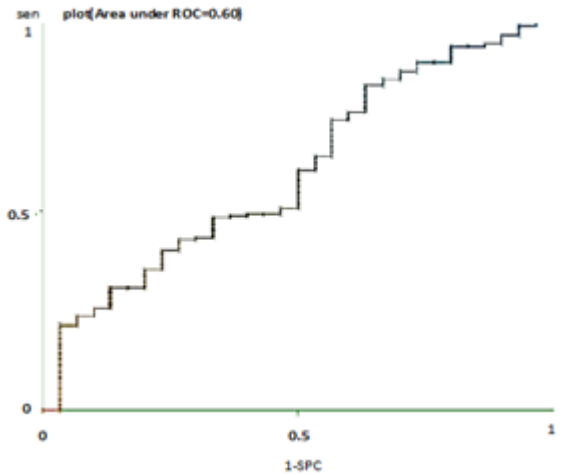


Figure 3. ROC curve in the logistic regression model on the experimental data

Finally, comparing the performance of SVM and logistic models in the training data are as follows:

Table 2. The comparison of the SVM and logistic regression models on the experimental data

Model assessment index	Logistic regression model	SVM model
Accuracy	%56/54	%67/66
Characteristic	%57	%67
Sensitivity	%50	%63
The area under the ROC curve	%60	%65

DISCUSSION AND CONCLUSION

Abadi et al., at the Student Conference on Electrical Engineering of Iran in the Technical University of Kermanshah, presented an article entitled diabetes diagnosis using SVM algorithm. The obtained accuracy for neural network model is 76.89 percent and it is 79.69 percent for the support vector machine model. It is concluded that the support Vector Machine performance is much better than the neural network[13] .In 1391, Abbas Tolouei Ishlaghi et al. in a study for predicting cancer recurrence by using data mining techniques, decision tree, artificial neural networks, and the support vector machine reported the accuracy of these three models as, 93.6%, 94.7% and 95.7% respectively showing a better performance of the support vector machine [14].Sadeghzadeh et al., at the second National Conference of computing and information technology, presented a paper comparing the intelligent algorithms to identify diabetes in which the reported accuracy in this paper for algorithms of Random forest, Naive Bayse, C4.5 and SVM was, 73.17, 76.3, 73.82 and 77.34 percent respectively. According to the results, SVM has a better performance than the other three algorithms [15].

In this study, two models including SVM and logistic regression were fitted to the data. According to the model assessment index of the obtained model from both models including accuracy, specificity, sensitivity and the area under the ROC, we concluded that the SVM model has a better performance to predict the factors affecting premature delivery than the logistic regression model.

ACKNOWLEDGEMENTS

We would like to thank God the Almighty for His endless blessings.

We would also like to appreciate all the participants of this study who helped us to achieve the goal.

"The authors declare no conflict of interest"

REFERENCES

1.Moghaddasi H, Hoseini AS, Asadi F, M J. Data mining and its application in health. Health Information Management. 2012;9(2):297-304.

2.Cunningham F, Leveno K, Bloom S, Hauth J, Rouse D. Spong CY Williams Obstetrics. USA: The McGraw-Hill Companies, Inc. Medical Publishing Division; 2010.

3.Beigi A, Saeedi L, Samiei H, Zarrinkoub F, Zarrinkoub H. Elevated CRP levels during first trimester of pregnancy and subsequent preeclampsia: a prospective study. Tehran University Medical Journal. 2008;66(1):25-8.

4.Wen SW, Smith G, Yang Q, Walker M, editors. Epidemiology of preterm birth and neonatal outcome. Seminars in Fetal and Neonatal Medicine; 2004: Elsevier.

5.Lawn JE, Cousens SN, Darmstadt GL, Bhutta ZA, Martines J, Paul V, et al. 1 year after The Lancet Neonatal Survival Series—was the call for action heard? The Lancet. 2006;367(9521):1541-7.

6.Varney H, Kriebs JM, Gegor CL. Varney's midwifery: Jones & Bartlett Learning; 2004.

7.M D. Textbook midwifery and intensive care. Tehran: Shahrab- Ayande sazan publications; 2006.

8.G R. Comparing SVM and logistic regression modeling to estimate the risk of death in patients

hospitalized in intensive care unit. Tehran: Para-Medical
Faculty ,Shahid Beheshti University of Medical Sciences; 2010.
9.Hamel LH. Knowledge discovery with support vector machines: John Wiley & Sons; 2011.
10.Zhang Y, Xu Q, Li J, Wang T. A Robust Biased Estimator for Exterior Orientation of Linear Array Pushbroom Satellite Imagery. Geomatica. 2008;62(1):33-44.
11.Saggaf M, Nebrija L. A fuzzy logic approach for the estimation of facies from wire-line logs. AAPG bulletin. 2003;87(7):1223-40.
12.Burges CJ. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. 1998;2(2):121-67.
13.Naem Abadi M, Amir Ahamdi Chamachar N, Tahami E, H R. diabetes diagnosis using SVM algorithm Student Conference of Electrical Engineering of Iran Kermanshah
Technical University of Kermanshah; 2011.
14.Tolouei Ishlaghi A, Pour Ebrahimi A, Ebrahimi M, L GA. predicting cancer recurrence using data mining techniques, decision tree, artificial neural networks, support vector machine. Iran breast disease. 2012;5(4).
15.Sadegh Zadeh M, Eshir A, Froutan F, M S. intelligent algorithms to identify diabetes 2th National Conference of computing and information technology: Mahshar branch,Islamic Azad university; 2012.