International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 2 06 – 09

Support Vector Machine to Detect Hypertension

[1]Zainab Assaghir, [1]Ali Janbain, [1]Sara Makki, [1]Mazen Kurdi, [2]Rita Karam
[1,2]Lebanese University
[1]Faculty of Science, [2]Faculty of medicine,
Beirut, Lebanon
[1]*zassaghir@gmail.com*,[1]*alijanbain.aj@gmail.com*,[1]*smakke.ch@gmail.com*,
[2]*mazen_kurdi@hotmail.com*,[2]*ritakmouawad@hotmail.com*

Abstract—Development of tools to facilitate diagnosis of some disease such as cancer, cardiovascular, hypertension, diabetes, is of great relevance in the medical field. In this paper, we will present a method based on Support Vector Machine regression model to detect the hypertension based on some risk factors including obesity, stress, systolic and diastolic blood pressure, physical exercises, cigaret consumption and diet lifestyle. Data represents a group of students from the Lebanese universities. After the data pre-processing, two Support Vector Machine models are designed and implemented in order to estimate systolic and diastolic blood pressure. The outcomes of the methods are diastolic and systolic blood pressure. Accurate results have been obtained which proves the effectiveness of the proposed Support Vector Machine for preliminary detection of hypertension.

Keywords-*Support Vector Machine Regression,Hypertension, Medical Diagnosis, Machine Learning.*

*****

I. INTRODUCTION

The cardiovascular diseases constitute an important problem in public health. Hypertension is caused by blood pressure and it is considered as a major risk factor of cardiovascular disease. Hypertension can cause stroke, heart failure, heart attack, and vision problems. The earlier diagnosis of hypertension saves enormous lives. In some cases, the use of computer based diagnoses can be more accurate than the clinical decision such as neural network [1], Support Vector Machine, regression models. The Support Vector Machine is a very robust classifier with many applications in several fields. This is a popular tool for machine learning tasks involving classification, regression or novelty detection. Support Vector Machine has been applied to some time series modelling problems [2,3], financial time series forecasting [4], in chemistry [5] and also in medical fields [6,7].

Support Vector Machines are very specific class of algorithms, characterized by usage of kernels, absence of local minima, sparseness of the solution and capacity control obtained by acting on the margin, or on number of support vectors. They were invented by Vladimir Vapnik and introduced at the Computational Learning Theory in 1992 [8].

In medical fields, the Support Vector Machine is used for cancer classification with microarray data [6]. This method is a promising classification approach for detecting persons with common diseases such as diabetes and pre-diabetes in the population [7]. Support Vector Machine can implement the complex medical processes by software. Software systems are more effective and efficient in various medical fields including predict, diagnose, treatment and help to the clinicians and physicians and the general population. This is a powerful tool to help doctors in the medical field with several advantages such as the ability to deal with a great amount of data and reduce the time of the diagnoses.

In this paper, we use the Support Vector Machine regression model in order to predict the hypertension. Two Support Vector Machine regressions methods handled the estimation of values of systolic and diastolic blood pressure. Ten explanatory variables considered as factor risks of the hypertension form the entry of the model. We divided the database into train and validation examples set.

The rest of this paper is organized as follows. Section 2 describes study materials and methods and presents Support Vector Machine regression model architecture. Section 3 presents the results of the study. Finally a conclusion is given in Section 4.

II. MATERIALS AND METHODS

Data consist of a sample of 3000 students from several universities in Lebanon. Each student responds to a questionnaire that includes: general information, anthropometric measurements, cardiovascular history, genetic background, diet lifestyle, alcohol consumption, tobacco consumption, physical activities, stress and environment. Developing high blood pressure varies between men and women and among various groups. Several approaches are used to select or extract the most important variables called features in a study. The best known are powerful mathematical means of data mining such as genetic algorithm, artificial neural network, and principal component analysis [9,10]. In this paper, the most relevant hypertension risk factors to be used: Gender (male or female), Heart rate, BMI (Body Mass Index obtained from height and weight), BF (Body Fat), Waist, Hip, PhysAct (representing the physical activities and taking the value yes if the student exercises some physical activity or no otherwise), SMOKE (taking the value yes if the student smoke or no otherwise), SALT denoting the diet lifestyle (taking the value yes if the student adds salt before tasting and no otherwise) and STRESS (a numerical value obtained with respect to Cohen's Test indicating the stress level of a student). These variables will be considered as explanatory variables for the model detailed after. Moreover,

two variables SBP (systolic blood pressure) and DBP (diastolic blood pressure) are measured for all students; SBP and DBP will form the outcome of the model.
After the collection, the data are preprocessed. In addition, the cases for which some data are missing are removed from the database to avoid the decreasing of the performance of the network. The data were analyzed using R software. Descriptive statistics and statistical tests are implemented in [1]. Two Support Vector Machine regression models are applied to predict both parameters SBP and DBP using all others parameters as regressors or explanatory variables. Note that the sample used to perform the support vector machine is divided into two sets: train set composed of 70% of the database and 30 % as a validation set to assess the accuracy of the model. The Support Vector Machine regression method is detailed here after.

Support Vector Machine

Let $x_i = x_i^1, \dots, x_i^{10}$ the set of inputs mentioned in the previous section, used as regressors and y_i^{sbp} and y_i^{dbp} the corresponding target values, our goal is to find a function f(x) that estimates the relation between the inputs and the target values.

Regression uses a loss function L(y, f(x)) that shows how the estimated function f deviates from the true values y. There are many forms of loss functions: linear, quadric loss function, exponential, etc. Vapnick's loss function is used with SVM, also known as ε-insentitive loss function defined as:

$$L(y, f(x)) = \begin{cases} 0 & if\ |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon & otherwise \end{cases}$$

Where $\epsilon > 0$ is a constant that controls the error. So the aim is to find a function f(x) that has at most ε difference from the actual values y, and to be as flat as possible.
If the function is linear, $f(x) = \langle w, x\rangle + b$, flatness means $\|w\|$ is small.

Where $\langle .,. \rangle$ denotes the dot product $\mathbb{R}^{10}$
$\|.\|$ is the Euclidean norm
$w \in \mathbb{R}^{10}$ are the weights
$x \in \mathbb{R}^{10}$ is the inputs vector
$b \in \mathbb{R}$ is bias

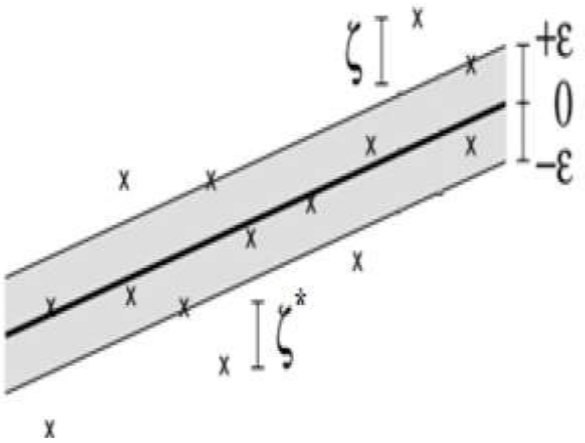
The optimization problem is summarized as:

$$\text{Minimize} \frac{1}{2}\|w\|^2$$

$$\text{Subject to} \begin{cases} y_i - \langle w, x_i\rangle + b \le \epsilon \\ \langle w, x_i\rangle + b - y_i \le \epsilon \end{cases}$$

This assumes the existence of such function f that estimates the relation between x and y with ϵ accuracy. This optimization problem may not always be feasible, and such function may not always exist.
To deal with this problem, slack we allow for some errors by adding slack variables ζ and ζ^* and a cost parameter C to indicate the trade-off between the flatness of f and the amount up to which deviations larger than ϵ are acceptable as shown in Figure 1.

Figure 1. ϵ-regression using support vector machines



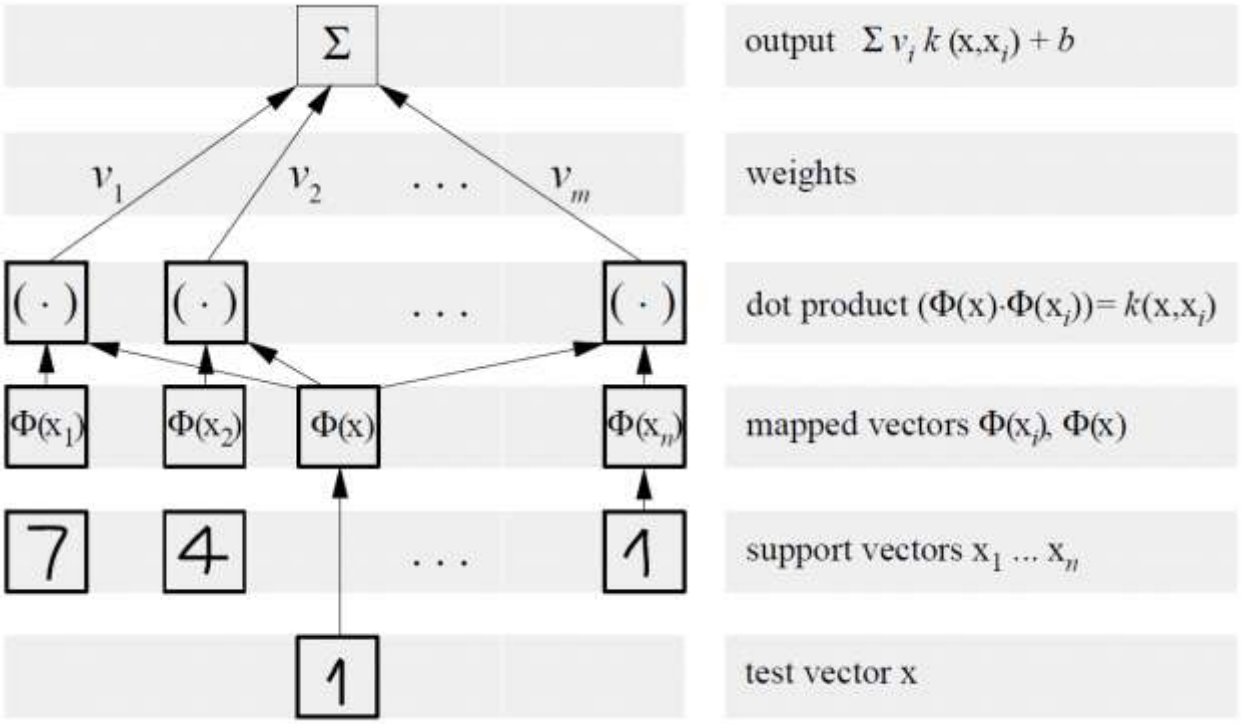
The optimization problem is then defined as follows:

$$\text{Minimize} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n}(\zeta_i + \zeta_i^*)$$

$$\text{Subject to} \begin{cases} y_i - \langle w, x_i\rangle + b \le \epsilon + \zeta_i \\ \langle w, x_i\rangle + b - y_i \le \epsilon + \zeta_i^* \\ \zeta_i \ge 0\ , \zeta_i^* \ge 0 \end{cases}$$

This problem is solved using a dual problem transformation and Lagrange multipliers detailed in [11]. The basic idea is to introduce dual variables $\alpha_i, \alpha_i^*, \lambda_i$ and λ_i^* and to use them to solve the Lagrange function.
The final solution obtained is $f(x) = \sum(\alpha_i, -\alpha_i^*)\langle x_i, x\rangle + b$.
A final note, not all data samples are used to describe the weights, only the ones that has non-zero value for α_i or α_i^*, therefore we have a sparse expansion of w in terms of x_i, these samples are called "Support Vectors".
Most of the cases, it is difficult to find a linear function that fits the model, so it is necessary to find a non-linear SVM algorithm. This is done by mapping the inputs in another feature space $\mathcal{F}$ of higher dimension where they are linearly separable using a mapping function ϕ. Since the SVM algorithm only depends on the dot products between data points [11], it is sufficient to define a function called "Kernel function" defined as: $k(x, y) = \langle \phi(x), \phi(y)\rangle$, without the need to explicitly find $\phi(x)$ because it may be too complicated. This is known as "the Kernel trick".
The expansion is therefore written as: $f(x) = \sum v_i\, k(x_i, x) + b$, where $v_i = (\alpha_i, -\alpha_i^*)$ when α_i and α_i^* are not simultaneously zero (the sparse expansion). Figure 2 shows an overview of the multiple phases of non-linear support vector machines regression.

Figure 2. Architecture of SVM regression



III. RESULTS

More than 3000 students participated to the study. After data preprocessing, some cases were removed for the performance of the Support Vector Machine regression model. A total of 2954 cases are used for this work: 48 % are males and 52 % are females. Results of continuous variables (means/standard deviations) and categorical variables (frequencies/percentages) are evaluated [1]. Note that the difference is significant between male and female groups for all continuous variables. Moreover the cigaret consumption is most present in the male group students. Moreover, a significant difference was observed between males and females in SMOKE and Physical exercises and no significant difference is observed for the lifestyle diet represented here by the variable SALT.

As for the Support Vector Machine regression models, we trained two models, the first one is considered for the variable SBP and the other for DBP using for both the same regressors which are the following: Gender, HR, BMI, BF, Hip, Waist, SALT, SMOKE, PhysAct and STRESS. Data are scaled internally both inputs and target variables (SBP and DBP) to zero mean and unit variance. The center and scale values are returned and used for later predictions.
The support vector machine results were computed and evaluated using the "e1071" package in R. The kernel function used is the Gaussian Radial Basis Function (RBF), with γ = 0.1 and ν = 0.5, defined as:

$$GRF(x) = e^{-\gamma\|x-\nu\|^2}$$

The cost parameter C was set to 0.1 chosen in a way that minimizes the training set's error.

The results achieve more than 85% prediction accuracy acceptable in the diagnosis of systolic and diastolic blood pressure.
Percentage error measures are often used, they are easy to interpret and have the advantage of being scale-independent. The commonly used percentage performance metric is known as Mean Absolute Percentage Error (MAPE) defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \times 100$$

Where F_i is the predicted value (output of SVM) of either SBP or DBP, A_i is the actual value, and n is the dimension of the test set. The SVM was able to predict the value of SBP with 7.61%, and DBP with 9.36% error.
The complexity of the support vector machine allowed the model to catch the relation between the risk factors and their effect on the blood pressure through sophisticated algorithms rather than the abstract character of traditional mathematical approaches. It relies on the detailed values of the risk factors of each student alone, rather than generalizing them with unconvincing mathematical assumptions. A second advantage of the Support Vector Machine is the kernel implicitly, which contains a non-linear transformation, no assumptions about the mapping function that allows the data to be linearly separable, is necessary. The mapping occurs implicitly on a robust theoretical basis.

IV. CONCLUSION

In this paper, we present a Support Vector Machine regression based method to detect hypertension. After a preprocessing of data, two models are implemented and the results were remarkable. Support Vector Machine deliver a unique solution, since the optimality problem is convex. This is an advantage compared to Neural Networks, which have multiple solutions associated with local minima and for this reason may not be robust over different samples. In this work, we present a suitable and powerful tool to help doctors in the medical field with several advantages such as the ability to deal with a great amount of data and reduce the diagnoses time. Moreover, we outline the importance of this method to predict the hypertension using the features in the hypertension diagnoses. The Support Vector Machine was able to predict the value of SBP with 7.61%, and DBP with 9.36% error. Its use makes the diagnoses more reliable and gives more satisfaction for the patient. This model can implement the complex medical processes by software. Software systems are more effective and efficient in various medical fields including predict, diagnose, treatment and help to the clinicians and physicians and the general population.

REFERENCES

[1] Assaghir Z, Janbain A, Makki S, Kurdi M, Karam R. Using Neural Network to predict Hypertension. International Journal of Science & Engineering Development Research. Accepted January 2017.

[2] S. Mukherje, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using a support vector machine. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, Neural Networks for Signal Processing VII — Proceedings of the 1997 IEEE Workshop, New York, 1997. IEEE.

[3] K.-R. Muller, A. Smola, G. R¨atsch, B. Sch¨olkopf, J. Kohlmorgen, and V. Vapnik. Pre-dicting time series with support vector machines . In B. Sch¨olkopf, C.J.C. Burges, and A.J. Smola, editors, Advances in Kernel Methods — Support Vector Learning, pages 243–254, Cambridge, MA, 1999. MIT Press. Short version appeared in ICANN'97, Springer Lecture Notes in Computer Science.

[4] Francis E. H. TAY, and Lijuan CAO, 2001. Application of support vector machines in financial time series forecasting, Omega: The International Journal of Management Science, Volume 29, Issue 4, August 2001, Pages 309-317.

[5] O. Ivanciuc, Support Vector Machines Prediction of the Mechanism of Toxic Action from Hydrophobicity and Experimental Toxicity Against Pimephales promelas and Tetrahymena pyriformis, Internet Electron. J. Mol. Des. 2004, 3, 802–821.

[6] Chu F., Wang L. Applications of support vector machines to cancer classification with microarray data. Int J Neural Syst. 2005 Dec;15(6):475-84.

[7] Yu W., Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010 Mar 22;10:16.

[8] Cortes C. and Vapnik V. 1995. Support vector networks. Machine Learning 20: 273–297.

[9] Yan H, Zheng J, Jiang Y, Peng C, Xiao S. Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. Appl Soft Comput. 8: 1105‑1111, 2008.

[10] Verikas A, Bacauskiene M. Feature selection with neural networks. Pattern Recognition Lett. 23: 1323‑1335, 2002.

[11] A. J. Smola and B. Scholkopf. "A tutorial on support vector regression." Technical Report - ESPRIT Working Group in Neural and Computational Learning II, 1998.