Comparison of different classification techniques for the discrimination of patients with pathologies of Parkinson

D Devia Narváez1, D Devia Narváez1, and R Ospina2
1 Universidad Tecnológica de Pereira, Pereira, Colombia
2 Universidad Industrial de Santander, Bucaramanga, Colombia

E-mail: dmdevian@utp.edu.co

Abstract. This paper presents a brief description of the application of the technique of support-vector machines in order to classify patients with pathologies of Parkinson's disease from voice samples. The problem of linear classification and the case of non-linear classification are addressed, for which the Kernel functions will be used in order to transform the input space into a space of greater dimensionality where it can be classified by means of a hyperplane.

1. Introduction

The problems of classification in recognition of patterns, have taken much consideration in recent years given the computational advance that is had for the execution of different techniques that allow to give a solution to this task. Its areas of application range from communications: voice recognition, image classification, biometric identification; analysis of biological samples for the detection of diseases; junk mail recognition. The basic idea of classification is to determine to which set of categories a new observation belongs from the analysis of a set of data or observations of which its membership is already known, that is, there is a training based on past data [1].

In this paper, the classification problem is addressed by the support-vector machines (SVMs) algorithm, which is a highly-used algorithm developed by Vapnik [2]. The operation of the SVMs is based on the determination of a hyperplane that separates the classes by the greatest possible distance. The data set used is \( \{x, y\} \) (where \( x \in \mathbb{R}^p \) where \( p \) is the number of input variables or characteristics and the output \( y \in \{-1, 1\} \) denoting the class label [3,4] (Equation (1)).

\[
\begin{align*}
\max_{\beta_0, \beta} & \frac{1}{\|\beta\|^2} \\
\text{s.a} \quad & y_i (\beta_0 + x_i^T \beta) \geq 1 - \xi_i \quad (1) \\
& \xi_i \geq 0 \\
& \sum_{i=0}^{n} \xi_i \leq B
\end{align*}
\]

Here, \( \xi_i \) is a slack variable and is a control variable. One of the main differences of the SVMs with respect to the majority of learning methods is the minimization of the structural risk instead of the empirical error minimization, where the selection of the hyperplane is made keeping in mind that the separation between the closest samples of each class to the hyperplane are equidistant, with which a maximum margin is obtained on each side of the hyperplane as seen in Figure 1 [5,6].
The problem developed herein is the classification of patients with diseases of Parkinson's disease, using a set of data generated by Oxford University for detecting the disease by measurements of biomedical voice 31 persons, 23 with disease Parkinson. The database consists of 195 different samples, 22 features (inputs) and an output state of the person, where 0 and 1 state for a healthy person with pathological disorder.

2. Materials and methods

2.1. Support-vector machines: Bi-class classification linearly separable

The SVMs search among all the possible hyperplanes that are capable of complying with the restrictions given that which is optimal, that is, the distance between samples of each class closest to the hyperplane is maximum, therefore, the concept of margin must be introduced [7,8]. Defined as the distance of the closest sample of either of the two classes to the hyperplane called $\tau$, as shown in Figure 1.

Starting from the geometric distance of a random sample $x'$ to the hyperplane is given by Equation (2):

$$\frac{|D(x')|}{\|\omega\|}$$

Now, knowing that each of the samples outside the given range complies with Equation (3):

$$\frac{y_i D(x)}{\|\omega\|} \geq \tau$$

Finding the optimal hyperplane implies finding $\omega$ that maximizes this distance, however, the solution of this problem are infinite given that there are values of $\omega$ that only differ in scale, that is, there are $\lambda \in \mathbb{R}$ such that $\lambda(<\omega, x > + b)$, they represent the same hyperplane, therefore, taking $\tau \|\omega\| = 1$ we get Equation (4).

$$y_1 D(x) \geq 1$$

Thus, Equation (5):

$$y_1 (<\omega, x > + b) \geq 1$$

The solution of finding the optimal hyperplane is determined with the solution of obtaining the appropriate values of $\omega$ and $b$ that minimize the functional $\|\omega\|$ subject to the constraints of Equation (5), that is Equations (6) and Equation (7):

$$\min f(\omega) = \frac{1}{2} \|\omega\|^2 = \frac{1}{2} <\omega, \omega >$$
\[ s. \ a \ y_i(\omega, x > +b) - 1 \geq 0 \quad (7) \]

For the development of Equation (6) the theory of quadratic programming must be applied, with which a solution can be determined through a dual problem because the initial problem known as primal has objective function and convex restrictions [9]. The dual problem approach is performed by calculating the Lagrangian function of Equation (6), as indicated Equation (8):

\[
L(\omega, b, \alpha) = \frac{1}{2} \|\omega\|^2 - \sum_{i=1}^{n} \alpha_i[y_i(\omega, x > +b) - 1] \quad (8)
\]

Where \( \alpha_i \) are known as Lagrange multipliers. Now, from the conditions of Karush-Kuhn-Tucker (KKT), we have Equation (9) to Equation (11):

\[
\frac{\partial L(\omega^*, b^*, \alpha)}{\partial \omega} \equiv \omega^* - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad (9)
\]

\[
\frac{\partial L(\omega^*, b^*, \alpha)}{\partial b} \equiv \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (10)
\]

\[
\alpha_i[1 - y_i(\omega^*, x_i > +b^*)] = 0 \quad (11)
\]

Therefore, the primal problem can be rewritten by its dual problem, which will only depend on \( \alpha_i \) as shown Equations (12) to Equation (13):

\[
\max \ L(\alpha) = \sum_{i=0}^{n} \alpha_i - \frac{1}{2} \sum_{i=0}^{n} \alpha_i \alpha_j y_i y_j < x_i, x_j > \quad (12)
\]

\[
s. \ a \ \sum_{i=0}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0 \quad (13)
\]

The advantage of solving the problem of optimization by means of the dual problem, is because this scale with the number of samples \( n \) while the dual problem does with the number of inputs \( p \), which will have a much lower computational cost even for databases of a high dimensionality [10,11].

2.2. Classification non-separable bi-class linearly

Given the existence of data sets in most practical cases that do not have separation by hyperplanes in the input space, it is necessary to use a set of bases, non-linear functions that allow transforming the input space to a new space (Figure 2), known as the largest (\( \mathcal{F} \)) space of characteristics, in which the classification problem can be solved linearly by the procedure described in the previous section [12].

The transformation is defined by means of the function \( \Phi: \mathcal{X} \to \mathcal{F} \) which makes correspond to each input vector a point in the space of characteristics, in such a way that \( \Phi(x) = [\phi_1(x), ..., \phi_m(x)] \) where \( \phi_i(x) \) is a non-linear function. The decision function of Equation (2) will be defined as follows Equation (14):

\[
D(x) = \omega_1 \phi_1(x) + \cdots + \omega_m \phi_m(x) \quad (14)
\]
3. Kernel functions

Kernel functions allow to transform a space of dimension $p$ to a space of greater dimensionality, said functions must comply with the following properties: be continuous, possess symmetry and positive. Now, some Kernels of greater use are: polynomial Kernel of degree, Gaussian Kernel and Sigmoidal Kernel.

3.1. Kernel principal component analysis and canonical correlation analysis

There are two techniques that can be used in conjunction with the Kernels functions, which allow to analyze in an optimal way a set of data with a high dimensionality.

3.2. Principal component analysis

The analysis of main components is a classic technique that allows analyzing data sets whose dimension is very high, which is why a subset of characteristics that best describe the information contained in the data is extracted from its hidden structure. Use is made of the $k$ principal eigenvectors with which the quadratic distance between the original data and the transformed data is minimized. Said eigenvectors are called principal axes of the data set, where the new coordinates are obtained from the projection of each point to these axes. The eigenvectors are defined as $v = \sum \alpha_i x_i$ where the value of $\alpha_i$ must be determined by using Kernel functions solving the following Equation (15):

$$m\lambda \alpha = \mathbf{K}\alpha$$

4. Description of the problem

Parkinson's disease is a progressive neurodegenerative disease whose cause is not known, its affectation occurs primarily in the muscles of the body, causing those who suffer from it to experience less fluid movements and affecting the reach of these. On the other hand, this disease can affect the muscles of the jaw and the tongue in the same way, therefore there will be an affectation in the speech, this affectation is known as dysphonia or speech disorder, and it has been demonstrated that around 90% who suffer from Parkinson's disease [13], are affected in this area. Taking the voice disorder as an early indicator for the detection of Parkinson's disease, it is necessary to determine means that allow the classification of patients with pathologies or not, through voice records. For this, we work with the database of UCI machine learning of 2008, called Parkinson’s Disease data set whose general information is as follows:

- Number of samples or records: 195
- Number of patients: 31
- Records per patient: approximately 6
- Number of features or entries: 23

5. Results

For the analysis of the data set used, the analysis of principal components (PCA) mentioned in section 2 was taken into account, with which it is possible to reduce the dimensionality of the input space by taking into account those characteristics that have greater relevance, Figure 3 shows the relevance of the characteristics, which means that there are attributes that do not provide more information to the data set, such as the characteristics:

- X1: MDVP: Fhi(Hz)
- X2: PPE
- X18: DFA
- X3: MDVP: Flo(Hz)
- X16: HNR
- X1: MDVP: Fo(Hz)
- X22: PPE
- X3: MDVP: Flo(Hz)

These characteristics may be discarded to reduce the dimensionality of the data set and have a better performance.
On the other hand, we have the correlation matrix, which indicates that equivalence between characteristics is as much, being -1 the indicator of inverse proportionality and 1 direct proportionality. Figure 4 shows that there are two groups of characteristics whose relationship is direct, these are the characteristics from X4 to X8 because they are responsible for measuring the frequency variations in consecutive cycles, while the characteristics of X9 to X14 deal with the variability of amplitude in consecutive periods. The percentage of correct answers in the training can be seen in Figure 5, in which the SVMs have a very high performance of 95.8% for 15 characteristics.

Finally, when carrying out the classification of sick people and healthy people, taking into account the projection for the middle of the Kernel’s functions, we obtain Figure 6:
6. Conclusion

In this document an application of the binary classification of patients with pathologies of Parkinson's disease was presented, based on the SVMs technique using Gaussian Kernel. In order to obtain the best performance of said Kernel, cross validation was performed with 10 iterations which allowed the choice of Kernel parameters. The application of Kernel functions does not depend on the number of characteristics, which presents a great advantage over data sets whose dimensionality is very high.

For the analysed data set the SVMs presented a percentage of assertiveness that was increasing until reaching 15 characteristics, from this the performance deteriorated, however, if you work with all the characteristics of the data set you will have a percentage of accuracy greater than 94%.

References

[1] Cortes C and Vapnik V 1995 Support-vector networks *Machine Learning* 20 273
[2] Little M A, McSharry P E, Roberts S J, Costello D A and Moroz I M 2007 Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection *Biomedical Engineering Online* 6 23
[3] McCulloch W S and Pitts W 1943 A logical calculus of the ideas immanent in nervous activity *The Bulletin of Mathematical Biophysics* 5 115
[4] M Sayan, O Edgar and G Federico 1997 Nonlinear prediction of chaotic time series using support vector machines *Neural Networks for Signal Processing VII Proceedings of the 1997* (Amelia Island: IEEE) p 511
[5] Platt J 1999 Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods *Advances in Large Margin Classifiers* 10 61
[6] O Sophocles J 1988 *Optimum signal processing: An introduction* (New York: Macmillan publishing company)
[7] Wang L, Zhu J and Zou H 2006 The doubly regularized support vector machine *Statistic Sinica* 16 589
[8] B Bernhard, G Isabelle and V Vladimir 1992 A training algorithm for optimal margin classifiers *Proceedings of the Fifth Annual Workshop on Computational learning theory* (New York: ACM) p 142
[9] V Vladimir 1998 *The support vector method of function estimation* (Boston: Springer)
[10] T Theodore and I Huseyin 2000 Support vector machine for regression and applications to financial forecasting *International Joint Conference on Neural Networks, IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium* (Italy: IEEE) p 348
[11] Lichman M 2013 *UCI Machine learning* (California: University of California)
[12] S Stefano 2010 Influence of raised floor on zone design cooling load in commercial buildings *Energy and Buildings* 42 1182
[13] Y Gala 2013 *Algoritmos SVM para problemas sobre big data* (Madrid: Univiersidad Autónoma de Madrid)