International Journal of Advances in Computer Science & Its Applications – IJCSIA 2018
Copyright © Institute of Research Engineers and Doctors , SEEK Digital Library
Volume 8 : Issue 2- [ISSN : 2250-3765] - Publication Date: 28 December, 2018

PCA Based Dimension Reduction of Feature Matrix to Train SVM for Balance Disorder Diagnosis

Serhat Ikizoğlu, Saddam Heydarov

***Abstract*— This study is mainly about the research to select the discriminative features for the machine learning algorithm to figure out the reason behind the problem of people who suffer from balance disorder. The foregoing step on this way has been determining the proper algorithm where we achieved the best performance with the Support Vector Machine (SVM) with Gaussian Kernel, the so-called Radial Basis Function (RBF). In our study, we first input the complete IMU-sensor based data set collected both from the healthy people and those suffering from vestibular system disorders to SVM-RBF. Next, we reduce the feature matrix using the Principle Component Analysis (PCA). Following this procedure, the machine is trained with the new data to recognize the effect of feature transformation on the accuracy of the learning method. We observed that PCA had satisfactory influence on the elimination of redundant features that it points to high correlation between some of the members of the starting feature matrix. The study will continue to cover more input vectors to PCA. Moreover, we plan sub-classification between various problems that lead to balance disorder. The situation with the current outputs of the study encourages to go further steps to achieve a significantly high performance for the machine learning algorithm with reasonable number of features.**

***Keywords*—principle component analysis, machine laerning, support vector machines, vestibular system**

I. Introduction

The vestibular system (VS) is of high importance in human life since it is responsible from the individual's balance. Thus, a number of researches are carried out to determine the failure in this system for those who cannot keep their balance when walking. The diagnosis is mainly based on two methods. The one is using motion sensors where the most common way is the implementation of inertial measurement units (IMU). Sensors placed on appropriate locations on the body collect data about the sway [1]. Another popular method is the use of insole pressure sensors used to give information about the distribution of weight when walking [2, 3].

Serhat Ikizoğlu
Istanbul Technical University
Control and Automation Engineering Dept.
Turkey

Saddam Heydarov
Istanbul Technical University
Control and Automation Engineering Dept.
Turkey

Following the data collection from a pre-defined set, a suitable algorithm trains a machine which in turn will serve for data mining. In this context, two points are to be investigated to increase the accuracy in the conclusion about the new data. Obviously, the choice of the right machine learning algorithm is one of the critical issues; and to increase the efficiency of the algorithm, the input data should also be able to give correct information. That is, the feature matrix should not have missing vectors nor redundant ones. It follows that there is need for reduction of the feature matrix in case it is too crowded which causes complexity in processing the data.

Machine Learning is a widely used technique in studies on medical applications. It is very useful in identification of problems. So far a number of applications of the technique have been carried out on different diseases. A study on chronic kidney disease is performed by Anusorn. C. et al (2016) [4]. They apply different ML algorithms on the data related to chronic kidney disease and try to decide on the algorithm giving the best performance.

A study to implement the ML in recognition of the Parkinson Disease is carried out by Marziye K. S. et al where they offer Extreme Machine Learning to achieve good results [5].

An application of ML on Parkinson Disease is the one from Aunsia K. and Muhammad U. (2015) where they propose a new method consisting of three steps as initial preprocessing, imperative attribute selection and final classification [6].

Identification of heart rate failures is another application area of ML in medicine. As an example, the study from Patel J. et al is on heart disease prediction based on ML and data mining [7].

Though a lot of studies on ML application in medical area are presented in literature, the technique is not yet effectively used for classification of vestibular system-related problems. From this aspect, this study has significant contribution to literature.

Once data is collected from the environment, the next meaningful step is to eliminate redundant information from the data set. An important reason for this size reduction process is to by-pass the unnecessary time consumption for calculation, where the other is to simplify the hardware for data acquisition by specifying the useful data among the whole set. In this context, PCA is extensively used [8].

In consideration of above explanations, the rest of this paper is arranged as follows: In Section 1 we introduce necessary information about the data acquisition process and dataset formation. Section 2 describes the ML application steps. The experimental results obtained are submitted in Section 3. Finally, in Section 4 we put forth the conclusion and discuss about future work.

International Journal of Advances in Computer Science & Its Applications – IJCSIA 2018
Copyright © Institute of Research Engineers and Doctors , SEEK Digital Library
Volume 8 : Issue 2- [ISSN : 2250-3765] - Publication Date: 28 December, 2018

II. Dataset Formation and Data Acquisition

We collected data from 37 people of different ages out of which 21 were healthy and the rest complaining about balance disorder. Table 1 gives detailed information about the overall group.

TABLE I. DISTRIBUTION OF SUBJECTS

Subject	Healthy	VS-Disorder
Male	11	5
Female	10	11

The group having VS-Disorder suffered from different diseases such as Multiple Sclerosis (MS), Benign Paroxysmal Positional Vertigo (BPPV), Vestibular Neuritis (VN), Benign Positional Vertigo (BPV) etc. mainly diagnosed by Computerized Dynamic Posturography.

Table 2 lists the specs of the IMU sensor used. It is the MTW2 Wireless unit from Xsens that houses 3D accelerometers, 3D gyroscopes and 3D magnetometers [9].

TABLE II. SPECIFICATIONS OF THE IMU SENSOR USED

Mass	27 grams
Physical dimensions	34.5 x 57.8 x 14.5 mm
Static accuracy (Roll, pitch)	< 0.5 degree
Static accuracy	< 1 degree
Dynamic accuracy	2 deg RMS
Angular resolution	0.05 deg
MTw internal sampling rate	1800 Hz
Max acceleration	16 g
Max MTw update rate	75 Hz (for 6 MTw)

We placed sensors on several locations on the body such as the feet, ankles, knees, waist etc. The locations were chosen to give related data for the features listed in Table 3. Most of these features are widely used in literature [10-13]. Explanations for some features are given in Table 4.

TABLE III. FEATURES SELECTED

Average step length-right	Average ascent by right foot
Average step length-left	Average ascent by left foot
Average velocity	Total Dist. Traveled by right foot
Step symmetry 1	Step symmetry 2
Total Dist. Traveled by left foot	Left knee swing
Left knee flexion angle	Average swing by left knee
Avrg flexion angle by left knee	Right knee swing
Right knee flexion angle	Average swing by right knee
Avrg flexion angle by right knee	Waist anterior swing
Waist posterior swing	Slope of Waist during walking
Lateral swing to left	Lateral swing to right

TABLE IV. EXPLANATIONS FOR SOME OF THE FEATURES

Parameter	Definition
Waist posterior swing [deg]	Waist angle as a result of swing done during walking, deg
Lateral swing [deg]	Swing done to the right or left from stationary position
Ascent by foot [m]	Distance between foot and ground during toe off
Velocity [cm/s]	Ratio of length of walking path to time spent on walking

The data was acquired at Istanbul University - Cerrahpasa Medical School – Audiology Department upon Ethical Committee Approval taken from the administration. Approval was also taken from the subjects for data collection. To form the required data, the free acceleration data of the motion sensors were used. These make use of the magnetic field of the earth; thus, it is critical that almost no other magnetic field should exist nearby to prevent from interference with the earth field in order not to lead to miscalculation. To provide the described ambient we collected data on weekends when almost all the electrical units were shut down. Figure 1 presents a view from the environment. The subjects were asked to walk along a straight line of approx. 12 meters.

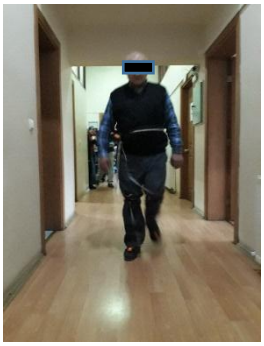

Figure 1. A view from the environment of data collection.

Figure 2 demonstrates data for some features on 3D histograms.

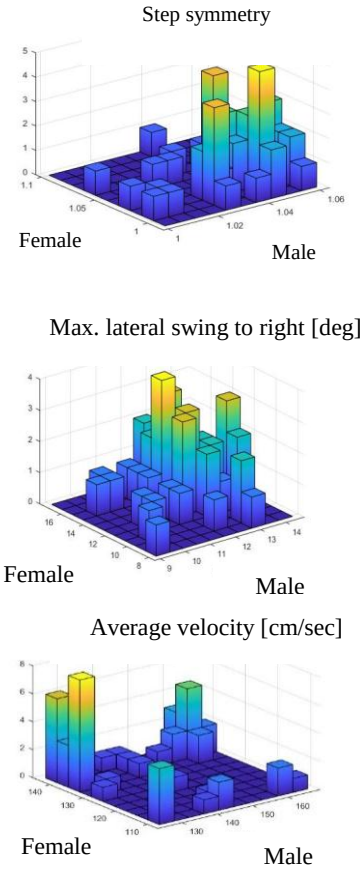

Figure 2. 3D histograms for some features.

III. Machine Learning Algorithm

In order to decide upon the ML algorithm we first checked for the performance of different algorithms using MATLAB. Table 5 presents the success of some algorithms.

TABLE V. PERFORMANCE COMPARISON OF SEVERAL CLASSIFICATION ALGORITHMS

Classification Algorithm	Kernel Function	Accuracy (%)
Complex Tree	X	50
Medium Tree	X	50
Simple Tree	X	50
Linear Discriminant	X	66.7
Logistic Regression	X	50
SVM	Linear	72.2
SVM	Quadratic	83.3
SVM	Cubic	83.3
SVM	Fine Gaussian	66.7
SVM	Medium Gaussian	83.3
SVM	Coarse Gaussian	55.6
KNN	Medium	55.6
KNN	Coarse	55.6
KNN	Cosine	55.6
KNN	Weighted	72.2
Ensemble	Boosted Tree	55.6
Ensemble	Bagged Tree	72.2
Ensemble	Subspace Discriminant	66.7
Ensemble	Subspace KNN	72.2

As it is also observed from Table 5, the best performance was achieved with Support Vector Machine. This in fact confirms the reality that for biological signals SVM is the algorithm mostly used in literature.

A. *Support Vector Machines (SVM)*

SVM is a well-known machine learning algorithm for supervised classification. The idea behind it is to find a hyper-plane between classes for optimum separation [14]. As illustrated in Figure 3a, one might be able to define a number of hyper-planes to separate two classes, but SVM searches for the optimum one. On the mathematical basis, this can be explained briefly as follows:

For the separating hyper-plane, the equation is

$$h(x) = g(\mathbf{w}^T\mathbf{x}+\mathbf{b}) \quad (1)$$

where **x** is the feature vector and **w** and **b** stand for weight and bias vectors respectively [14]. The function g(r) gives -1 for r<0 and +1 otherwise. The method searches for the geometric margin that is the shortest Euclidian distance of the feature from the hyper-plane (Figure 3b).

Hence, the maximum geometric margin is achieved with

$\max_{h,w,b}$ h s.t. $y(i)(\mathbf{w}^T\mathbf{x}(i) + \mathbf{b}) >$ □ for all training examples ||w||=1 [14]. (2)

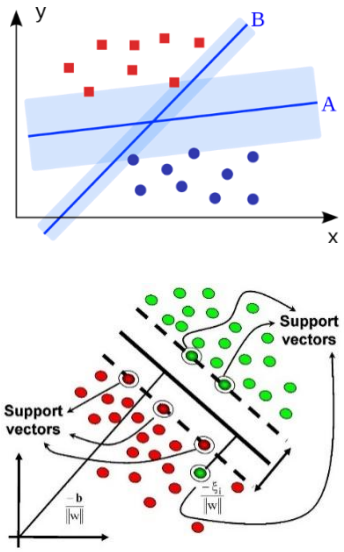

Figure 3. Support vector illustrations (a: top) [15, 16].

The corresponding (w, b) pair forms the optimal hyper-plane. The explanations so far are valid for classes that are linearly separable. If this is not the case, Kernel functions are to be used for space transformation that the features are linearly separable in the new space. In our study, Gaussian Kernel is used which is given as [17]

$$K(x,x') = exp(-\square \quad ||x-x'||^2) \quad (3)$$

B. *Reducing Dimension*

The number of features used for machine learning and data mining is an important issue. Especially if the data set is not large, large number for the feature set may cause over-fit in the learning which in turn may lead to misclassification of the unknown input vector. Thus, the feature matrix should include a reasonable number of vectors which is the issue of dimensionality reduction. Two main methods are mainly applied for this purpose. The one is the so-called 'Feature Selection' (FS) where sufficient number of features will be selected among the starting feature set. The other method is the 'Feature Transformation' (FT) where using the original feature set, a transformation to a new set is performed that contains less number of feature vectors than the original set.

In this study we chose the FT method, since the nature of data acquisition gives the opinion that significant number of features might be close correlated. To apply the mentioned method, we made use of the Principle Component Analysis procedure.

C. *Principle Component Analysis (PCA)*

As explained above, we search for the optimal hyperplane which is the one on which sum of squares of projection of features are minimum. Here, PCA helps to

reduce the number of features to decrease the calculation complexity without significant loss of information.

PCA is a statistical method used to reduce the dimension of the original matrix [18]. The steps to go for the aim can be summarized as follows:

Let any raw feature have the mean value

$$\mu=\frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

and the variance

$$o_j^2 = \frac{1}{m-1}\sum_{1}^{m}(x_j^{(i)})^2$$

- Raw features are first normalized to have zero-mean and unit-variance
- The covariance matrix is evaluated as

$$\sum = \frac{1}{m-1}\sum_{i=1}^{m}(x^{(i)})(x^{(i)})^T$$

- The eigenvalues and eigenvectors of the covariance matrix is found.
- The eigenvectors are perpendicular to each other and give information about the characteristics of the data. The principle component is the eigenvector having the highest eigenvalue. So, the eigenvectors are put in order from having the highest eigenvalue to lowest.
- The feature-vector is constituted with selected number of eigenvectors according to predefined criteria (Usually acceptable accuracy).
- The final step to form the new data set is that the transpose of the mean-adjusted original data set is multiplied on the left with the transpose of the feature-vector.

The reduced matrix is in form of

$$Z = U_{red}^{T} * \mathbf{x} \quad (4)$$

where U_{red} is the matrix with reduced features.

Our criteria to determine the number r of new vectors has been that the loss of information will be less than 1%. This is, that we select r vectors in replacement of m to satisfy the inequality

$$\frac{\frac{1}{r}\sum_{i=1}^{r}||x^{(i)} - x_{proj}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} < \%1$$

where $\mathbf{x}_{proj}$ is the projected vector of the new matrix.

IV. Experimental Results

The ML algorithm was decided to be SVM with RBF after using 5-fold cross validation to prevent over-fitting [19]. The accuracy of the selected algorithm was also verified with Receiver Operating Characteristics (ROC) graphs of different algorithms using MATLAB [20]. Next, PCA was applied on the raw feature matrix for size reduction using MATLAB toolbox. The criteria as the information loss of maximum 1% was applied in this phase of the study. The final feature matrix contained 13 vectors, significantly less than the raw matrix with 22 vectors. This obviously points to the fact that many of the original features are correlated to each other as expected. The accuracy with the new matrix was calculated as 82.6%, somewhat less than the starting value with 83.3%. The drop in accuracy is considered to be acceptable.

V. Conclusion and Future Work

This study is a valuable part of a project on defining the problems of people suffering from balance disorder. In this context, data is collected from both the healthy people and those having complaints concerning the vestibular system. The overall aimed procedure can briefly be listed in two steps, namely the machine learning and the data mining. Regarding the first step, SVM with RBF is determined as the most powerful method for our data set. At the beginning, we have determined 22 features to be helpful for accurate decision. Next, to reduce the calculation complexity, PCA is applied on the original feature matrix. Under the criteria that the information loss will be less than 1%, we have obtained a reduced matrix of 13 vectors. The final success in learning is found to be 82.6%.

Though we have gone a substantial way so far, the study has not come to an end. Within the frame of this study we are planning to add new feature vectors, thus, enlarging the feature matrix. The data acquisition process is also going on. These factors will obviously contribute in increasing the accuracy of the whole system. Another significant part of the future study will be to develop the system not to classify between healthy people and sufferers only, but also sub-classify between specific reasons of balance disorder.

Acknowledgment

This research is a part of the project ‘Development of a dynamic vestibular system analysis algorithm & Design of a balance monitoring instrument’ (ID:115E258) supported by the Scientific & Technological Research Council of Turkey (TUBITAK).

References

[1] D. Basta, M. Rossi-Izquierdo, A. Soto-Varela, & A. Ernst, “Mobile posturography: posturographic analysis of daily-life mobility” Otology & Neurotology, 34(2), 2013, 288-297.

[2] S. Crea, S. M. M. De Rossi, M. Donati, P. Reberšek, D. Novak, N. Vitiello, M. C. Carrozza, M. C. “Development of gait segmentation methods for wearable foot pressure sensors”, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2012, pp. 5018–5021. http://doi.org/10.1109/EMBC.2012.6347120

[3] A. K. Chong & P. Milburn, “Human Plantar Pressure Image and Foot Shape Matching”, Journal of Biosciences and Medicines, 3(June), 2015, 36–41.

[4] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, & N. Ninchawee, “Predictive analytics for chronic kidney disease using machine learning techniques”, Management and Innovation Technology International Conference (MITicon), IEEE, 2016 (pp. MIT-80).

[5] M. K. Shahsavari, H. Rashidi, & H. R. Bakhsh, "Efficient classification of Parkinson's disease using extreme learning machine and hybrid particle swarm optimization", Control, Instrumentation, and Automation (ICCIA), IEEE, 2016 4th International Conference, pp. 148-154.

[6] A. Khan & M. Usman, "Early diagnosis of Alzheimer's disease using machine learning techniques: A review paper", Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), IEEE, 7th International Joint Conference, 2015, Vol. 1, pp. 380-387.

[7] J. Patel, D. TejalUpadhyay & S. Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique" Heart Disease, 2015, 7(1).

[8] I. T. Jolliffe, Principal component analysis, Springer, 2002.

[9] Xsens website: https://www.xsens.com/products/mtwawinda/

[10] S. Heydarov, S. İkizoğlu, K. Şahin, E. Kara, T. Çakar, A. Ataş, "Performance comparison of ML methods applied to motion sensory information for identification of vestibular system disorders", IEEE, 2017.

[11] S. Ikizoglu, K. Şahin, A. Ataş, E. Kara, T. Çakar, "IMU Acceleration Drift Compensation for Position Tracking in Ambulatory Gait Analysis", ICINCO 2017-2, Madrid 2017.

[12] S. Tadano, R. Takeda, K. Sasaki, T. Fujisawa & H. Tohyama, "Gait characterization for osteoarthritis patients using wearable gait sensors (H-Gait systems)", Journal of biomechanics, 2016, 49(5), 684-690.

[13] S. Qiu, Z. Wang, H. Zhao & H. Hu, "Using distributed wearable sensors to measure and evaluate human lower limb motions", IEEE Transactions on Instrumentation and Measurement, 2016, 65(4), 939-950.

[14] A. Ng, Support vector machines. Machine Learning, 2008.

[15] https://www.bing.com/images/search?view=detailV2&ccid=VBgsq%2b%2bf&id=4AC4F AD3D6B510574B0DB60D2A3E49CB43FD46CD&thid=OIP.VBgsq--fJqJUiXbYYVpE1gHaFS&q=support+vector+machine&simid=608015857886822999&selectedIndex=15&ajaxhist=0

[16] https://www.bing.com/search?q=support+vector+machine&form=EDGTCT&qs=AS&cvid=c67ae0cf0bd9441390d866b6591034d7&refig=07b42fd09d914e3fb41ad59f2e852367&cc=TR&setlang=en-US

[17] C. W. Hsu, C. C. Chang & C. J. Lin, "A practical guide to support vector classification", 2003.

[18] http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[19] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", Ijcai, 1995, Vol. 14, No. 2, pp. 1137-1145.

[20] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers", Machine learning, 2004, 31(1), 1-38

Serhat Ikizoglu graduated from Istanbul Technical University (ITU)-Control and Computer Engineering. He completed his doctorate at the same institute in 1992. He is currently Associate Professor at ITU-Control and Automation Engineering Dept. His area of interest mainly covers: Measurement, Instrumentation, Control and Mechatronics.

Saddam Heydarov graduated from Electrical and Electronics Engineering Department with Control Engineering concentration from Middle East Technical University-Turkey. He is currently MSc student in Control and Automation Eng. Dept. at Istanbul Technical University. His research area of interest includes data analysis, modeling and simulation of biomedical systems.