UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume 2 | Year: 2017 | ISBN: 978-967-0194-95-0

# Gender Determination Based on Femur Bones Using Principal Component Analysis and Support Vector Machine (PCA-SVM)

Ahmad Asyraf Abu Bakar[1], and Norhaizan Mohamed Radzi[2]

Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia
[1]asyrafler@gmail.com, [2]haizan@utm.my

**Abstract.** Dimension reduction is an important pre-processing for classification problem and can improve the accuracy of classification technique. Thus the need of Dimension reduction is crucial to reduce the dimension of data and extracted the features that contain only most relevant information for the classification testing purpose. The purpose of this paper is to investigate the use of Principal Component Analysis (PCA) as a variable reduction for Support Vector Machine (PCA-SVM) classification. For evaluation purpose, the other techniques used are Support Vector Machine (SVM) and Artificial Neural Network (ANN) and both result will be compared to the proposed technique, PCA-SVM. The result shows that the implementation of PCA-SVM classification can improve the performance of SVM using four variable. As a conclusion, integrated PCA- SVM give better performance of classification compared to SVM and ANN.

**Keywords:** Machine learning, Variable Reduction, Principal Component Analysis (PCA), Support Vector Machine (SVM), Artificial Neural Network

## 1 Introduction

Forensic anthropology is a branch of biological anthropology. It has become the fastest growing discipline and has objective to identify biological profile from the skeletal remains or skeleton (Iscan,1980; Iscan and Olivera, 2000; Vaz and Benfica, 2008). Generally, there are four important of biological profile which investigated namely determination of gender, age, ancestry, and stature (Love and Hamilton,2011). Determining gender of an evidentiary sample can be an important part of casework analyses (Varlaro and Reynold, 1996). Thus gender identification is then used to investigate the age, ancestry and stature estimation.

Gender determination is the classification of an individual as either male or female from skeletal part (Afrianty et al., 2015). Knowledge of the gender of an unknown set of remains is essential to make more accurate for age determination and others (continuous) (Afrianty et al, 2015). In determination of gender from skeletal part such as femur bones used must be completed and available. Skeletal part draw more intention recently is femur bone. For example femur bone are used in the work of gender identification conducted by (Steyn and Iscan, 1997; Rissech et al.,2008; G. Mall et al.,2003; Bhosale and Zambare, 2013 ).

Previous study shows that researchers have used many classifications method to identify gender from skeletal remain. Other classification technique such as Discriminant function Analysis (DFA) has been used by many researcher; (Luo,1995; Jr et al.,1998; Kemkes-Gronttenthaler, 2005; Patil and Mody,2005; Dixit et al., 2007; Zeybeck et al.,2008;

Gonzalez, 2008,Jardin, 2009 ) and is found to be most famous used technique for gender determination since it is easy to use, simple, quick and accurate without any need of specific skill (Anuthama et al., 2011). However previous study shows that many researcher consume all the attribute or variable of skeletal part for the gender identification. According to Chen et. al, 2005, dimension reduction is important pre-processing for classification and have potential to increase the performance of classification. This is due to the data behaviours which consisting of irrelevant data, noisy and human error and might affect the performance of classification. Thus the variable reduction is really needed in some case especially involving large dataset.

The aim of this research is to develop a classifier that can produce better performance in term of accuracy, sensitivity and specificity for gender determination. To achieve this aim, the performance of proposed method which is Principal Component Analysis and Support Vector Machine (PCA-SVM) will be compared with two previous technique Support Vector Machine (SVM) and Artificial Neural Network (ANN). The model that give the better accuracy, sensitivity and specificity will be selected as the best model for gender determination using femur bones. The data used in this paper are femur bones which is obtained from Goldman Osteometric Data used containing 137 Femur Bones ( http://web.utk.edu/~auerbach/GOLD.htm ). There are seven variable of femur bone used in this study which are Femur Maximum Length (LFML), Femur Bicondylar Length (LFBL), Femur Epicondylar Mediolateral Breadth (LFEB), Femur Distal Articular (Bicondylar) Mediolateral Breadth (LFAB), Femur Head Anteroposterior Diameter (LFHD), Femur 50% Diaphyseal Mediolateral Diameter (LFMLD) and Femur 50% Diaphyseal Anteroposterior Diameter (LFAPD).

## 2 Related Work

SVM is considered as the most technique used for classification for small and high dimensional data set since its ability to provide better result for classification problem. Such researcher used SVM as classification method for gender determination are Yoo et. al 2005, Santos et. al 2014 and Afrianty et al.,2015. However these researcher use all the available data (attributes) for classification. Variable reduction is believe to be vital expect for pre-processing process before undergo classification. This problem can be proved from the experiment carried out by Afrianty et. al 2015 which using PCA as dimension reduction for the Back-Propagation Neural Network (BPNN) classification.

The result obtained by the experiment shows that the integrated classification technique Principal Component Analysis – Back Propagation Neural Network (PCA-BPNN) shown that the accuracy and sensitivity of BPNN increase. Before implementing the PCA, the accuracy and sensitivity of BPNN for training are 97.90% and 95.58% while for testing the accuracy obtained is 96.77% and sensitivity of 94.45 %. Meanwhile when PCA is implemented, the accuracy and sensitivity of PCA-BPNN increase to 99.02% and 98.30% indicate that the performance of classification increase when the dimension reduction is implemented. Thus this paper believe the present of PCA for dimension reduction will increase the performance of SVM classification.

# 3 Methodology

This work describe the methodology used to develop SVM classifier and PCA-SVM

## 3.1 SVM

SVM classifier is trained and tested multiple with addressed to obtain the best variable setting that able to obtain high accuracy. There are five learning kernels in the SVM classification which is Linear kernel, Sigmoid kernel, Computed kernel, Radial Basis Function kernel and polykernel. kernel function because RBF has ability to maps the data into higher dimensional space, improve the drawback if the relation between class label and attributes is nonlinear (Omar,2013).

To train SVMusing RBF kernel, it is necessary to select proper C value and y value as well as kernel parameters (Afrianty et al, 2015). Selection of C is by using LIBSVM default value while for gamma value the selection is from try an error. Flowchart of SVM is shown blow in Figure 3.4

Below shows overall process of SVM classification using LibSVM (Omar,2013).

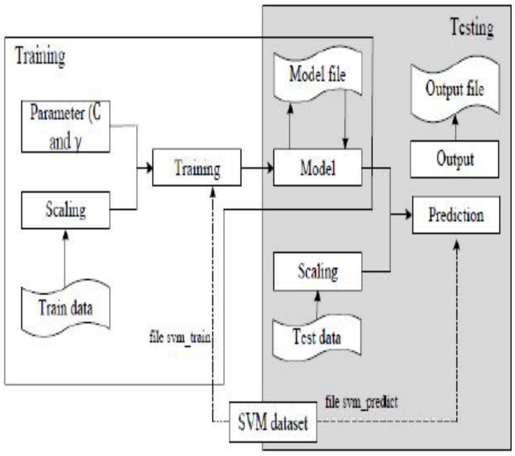


**Figure 3.1 (a)** SVM classifier

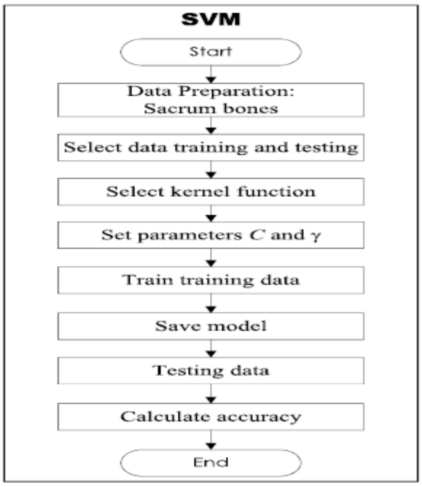


**Figure 3.1 (b)** Flowchart of SVM

The initial process of SVM classifier starts from data preparation which is femur bones. In the data preparation, the data normalization also involve to make sure the data are in the specific scale. In this case, the scale is from the range of 0 to 1. The data will be divided into testing and training which is 30% for testing and 70% for training. After that selection of kernel is made. There are four kernel in SVM which is linear, polynomial, Radial Basis Function (RBF) and sigmoid kernel. In this research, RBF is use as kernel function because RBF has ability to maps the data into higher dimensional space, improve the drawback if the relation between class label and attributes is nonlinear (Omar,2013). After set kernel is completed, proceed to set parameter C and Y. For the process, value of Y and C will be validate and find the best value for the SVM training and testing process. After that training data and model will be save. The model developed will be used for testing data as well as calculate the accuracy.

## 3.2 PCA-SVM

Process of PCA start with data reparation of femur bones and will then undergo five process. From the data of six variable or attribute of femur bone, the first process is to calculate the Mean and Standard deviation. The second step is transforming the data into matrix. From the matrix, the covariance matrix will calculated. The value of covariance matrix then will be used to calculate eigenvalue and eigenvector. The last step is calculate the PC score based on the eigenvalue and eigenvector. Finally the variable represent the best value based on PC-score will be selected. The overall process of PCA is shown in Figure 3.3.

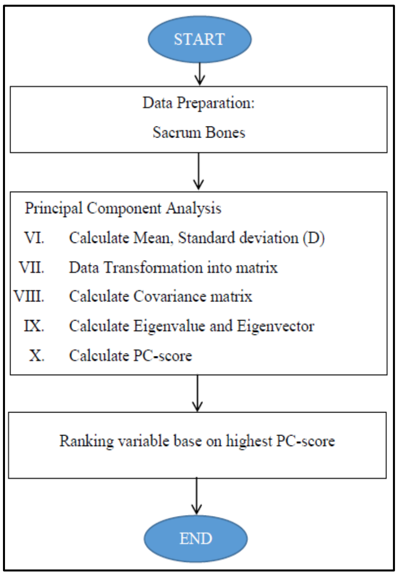

START
Data Preparation:
Sacrum Bones
Principal Component Analysis
VI. Calculate Mean, Standard deviation (D)
VII. Data Transformation into matrix
VIII. Calculate Covariance matrix
IX. Calculate Eigenvalue and Eigenvector
X. Calculate PC-score
Ranking variable base on highest PC-score
END

**Figure 3.2** Process of PCA

## 3.2 PCA step

i. Calculate Mean, Standard Deviation and Variance of Dataset

$$\text{Mean} = \text{x} = \frac{\sum_{i=1}^{n} x_i}{\text{n}}$$

Where $\bar{\text{x}}$ = mean ,n = number of data, Xi = data i.

Standard Deviation

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{\text{x}})}{(\text{n}-1)}}$$

$$\text{Variance} = \text{S}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{\text{x}})}{(\text{n}-1)}$$

ii. Transform the data into standardized matrix Z.

$$Z_i = \frac{(x_i - \bar{\text{x}})}{S_{xi}}$$

iii. Calculate covariance (correlation) Matrix.

$$\text{cov(X, Y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{\text{x}})(y_i - \bar{y})}{(\text{n}-1)}$$

iv. Calculate Eigen Value and Eigen vector

The eigenvalue and eigenvector re calculated from covariance matrix. Covariance matrix was decompose to obtain a matrix of eigenvectors in an n-dimensional space, which consist of set of eigenvector in n dimension and their corresponding eigenvalue with the following equation that obtained all of the eigenvalue

$$|R - \lambda I| = 0$$

$$\lambda 1 \geq \lambda 2 \geq \ldots \geq \lambda n$$

and the corresponding eigenvector tj = ( t1j, t2j, ..., tmj ). Eigenvector is calculated in a way iteration of matrix correlation, and normalized

$$Y_j = \sum_{k=1}^{m} t_{kj} X_k$$

Where ;

R is determined by the principle, K is kth measured values of ith and jth factor, K = 1, 2, …, r

v. Calculation of Principal Component (PCs) Score of each variable
PC Score = eigenvalue each variable / (total eigen value)

# 4 Experimental Result

SVM Classification from Femur dataset, provide the accuracy of 92.68% , Sensitivity of 100% and Specificity of 85%. For the male class, SVM show the predicted class that belong to male is 21 out of 24 class while for female class, SVM predict all the 17 class of female belong to female class. Figure 5.1 below is the bar chart representing overall performance of SVM for training and testing.

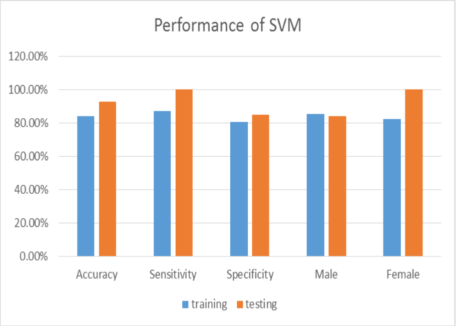
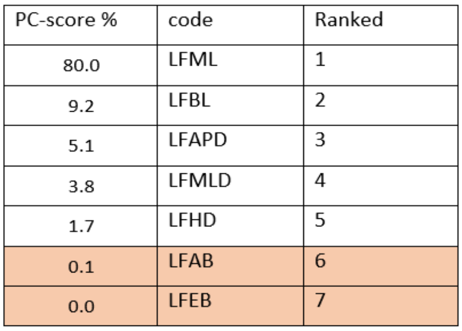


Figure 4.1 Performance Of SVM

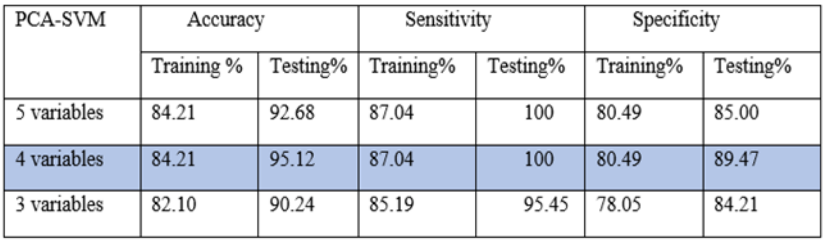
## 4.1 PCA-SVM

From the table PC-Score the highest PC-score obtain are LFML with percentage variation explained value of principal component 80% follow by LFBL, LFAPD, LFMLD, LFHD and LFAB. Since the PC score obtained by LFAB and LFEB is below than 1%, thus this two variables is reduced. The remains variables that will be used as an input to SVM classification is LFML, LFBL, LFAPD, LFMLD and LFHD.The performance of PCA-SVM is measured using data with 5 variables, 4 variables and 3 variables. From the result obtained, the best performance for PCA-SVM classification are using four variable. The variables that are mentioned are (LFML, LFBL, LFEB and LFAB). The best performance for classification accuracy obtained are 84.21% for training and 95.12% for testing, sensitivity of 87.04% for training and 100% for testing. Meanwhile, specificity obtained are 80.49% for training and 89.47% for testing. Thus, the variable selected for PCA-SVM is four variables.



| PC-score % | code | Ranked |
|---|---|---|
| 80.0 | LFML | 1 |
| 9.2 | LFBL | 2 |
| 5.1 | LFAPD | 3 |
| 3.8 | LFMLD | 4 |
| 1.7 | LFHD | 5 |
| 0.1 | LFAB | 6 |
| 0.0 | LFEB | 7 |

**Figure 4.1** PC score on each variable



| PCA-SVM | Accuracy | | Sensitivity | | Specificity | |
|---|---|---|---|---|---|---|
| | Training % | Testing% | Training% | Testing% | Training% | Testing% |
| 5 variables | 84.21 | 92.68 | 87.04 | 100 | 80.49 | 85.00 |
| 4 variables | 84.21 | 95.12 | 87.04 | 100 | 80.49 | 89.47 |
| 3 variables | 82.10 | 90.24 | 85.19 | 95.45 | 78.05 | 84.21 |

. **Table 4.1** Performance Of PCA-SVM using five, four and three variable

## 5 Discussion

The Table 4.1 and Figure 4.1 show the overall performance of classification experimented in this paper. For the accuracy training, ANN provide better result 94.75% however for accuracy testing PCA-SVM give better accuracy which is 95.12 compared to other classifiers. For sensitivity training, ANN give better result with percentage of 95.12% while for testing all the three classifiers, ANN, SVM and PCA-SVM provide the same result which is 100%. For performance specificity of classification, ANN provide better result for training with 94.87% while PCA-SVM give better result for specificity testing which is 89.47%. From overall performance classifier, ANN show to give better performance for training while PCA-SVM is better in testing purpose.

## 5 Conclusion

From the result and analysis of the experiment carried out in this study, PCA able to improve the performance of SVM classification. This experiment believe that the implementation of variable reduction are important aspect to improve the classification performance. For the data with high dimension, this step is helpful since PCA able to distinguish between variable that shows high variant or significantly related to each other. From the result of PCA obtained, the seven variable from the dataset had been reduce to 5 variables. These five variable shows significantly related to each other. Overall classification performance used in this study, Artificial Neural Network (ANN) is the best classification since the result obtained for accuracy prediction. Sensitivity to distinguish between male and female class and also the specificity performance are better compared to the SVM and PCA-SVM.

**Acknowledgements.** I would like to thank to my supervisor, Dr.Norhaizan binti Mohamed Radzi for her guidance and support me to finish this research. Also, I would like to thank to previous researcher who had done the research regarding on this proposed topic.

## References

7. Yang, Ming-Hsuan, and B. Moghaddam. "Gender Classification Using Support Vector Machines." Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101) (2000). Web
8. "Feature Selection Using Support Vector Machine." Support Vector Machine in Chemistry (2004): 60-73. Web.
9. Luo, Yuan-Cai. "Sex determination from the pubis by discriminant function analysis." Forensic science international 74.1 (1995): 89-98
10. Kemkes-Grottenthaler, Ariane. "Sex determination by discriminant analysis: an evaluation of the reliability of patella measurements." Forensic science international 147.2 (2005): 129-133.
11. Akhlaghi, Mitra, et al. "The value of mandible measurements in gender prediction for the Iranian adult population." Australian Journal of Forensic Sciences 46.2 (2014): 127-135.
12. Saulsman, Bree, Charles E. Oxnard, and Daniel Franklin. "Long bone morphometrics for human from non-human discrimination." Forensic science international 202.1 (2010): 110-e1
13. Phatsara, Manussabhorn, Korakot Nganvongpanit, and Pasuk Mahakkanukrauh. "Comparative morphometric study for distinguishing between human and non-human mammalian (cow, dog, horse, monkey and pig) long bones."
14. Eshak, Ghada A., Hala M. Ahmed, and Enas AM Abdel Gawad. "Gender determination from hand bones length and volume using multidetector computed tomography: a study in Egyptian people." Journal of forensic and legal medicine 18.6 (2011): 246-252.
15. Ishak, Nur-Intaniah. Sex and stature estimation using hand and handprint measurements in a Western Australian population. Diss. University of Western Australia, 2011
16. Mastrangelo, Paola, et al. "Sex assessment from the carpals bones: discriminant function analysis in a 20th century Spanish sample." Forensic Science International 206.1 (2011): 216-e1
17. Gómez□Valdés, Jorge Alfredo, et al. "Discriminant function analysis for sex assessment in pelvic girdle bones: sample from the contemporary Mexican population." Journal of forensic sciences 56.2 (2011): 297-301
18. Ogawa, Yoshinori, et al. "Discriminant functions for sex estimation of modern Japanese skulls." Journal of forensic and legal medicine 20.4 (2013): 234-238
19. Zech, Wolf-Dieter, et al. "Sex determination from os sacrum by postmortem CT." Forensic science international 221.1 (2012): 39-43
20. Šlaus, Mario, et al. "Sex determination by discriminant function analysis of the tibia for contemporary Croats." Forensic science international 226.1 (2013): 302-e1.
21. Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1-16
22. Afrianty, Iis, et al. "Back-Propagation Neural Network for Gender Determination in Forensic Anthropology." Computational Intelligence Applications in Modeling and Control. Springer
23. The Goldman Osteometric Data Set - Dr. Benjamin M. Auerbach, Associate Professor of Anthropology at The University of Tennessee – Knoxville http://web.utk.edu/~auerbach/GOLD.htm
24. Shlens, Jonathon. "A tutorial on principal component analysis." arXiv preprint arXiv:1404.1100 (2014).

Reris, Robert, and J. Paul Brooks. "Principal Component Analysis and Optimization: A Tutorial." (2015): 21