Hellinger Distance Decision Tree (HDDT) Classification of Gender with Imbalance Statistical Face Features

Muhamad Hasbullah Mohd Razali[1], Muhammad Farhan Muhammad[2], Balkiah Moktar[3*]
[1,2,3]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perlis Branch, Malaysia

*Corresponding author: *balkiah@perlis.uitm.edu.my*

Received Date: 1 August 2018
Accepted Date: 12 November 2018

ABSTRACT

Face recognition is one of the technologies used for assets protection. Face recognition also presents a challenging problem in the field of image and computer vision and has been used for application such as face tracking and person identification. It also frequently used in a security system such as security camera in airport, banks and offices. Practically, there are problems on improving face recognition performance particularly for gender identification. It is very difficult to differentiate the person based on face appearance from different poses, lighting, expressions, ageing and illumination. Sometimes it is also difficult to identify the shape of human faces because different people have different structure of faces. This study used image retrieved from Student Information Management Systems (SIMS)from 10 male and 43 female students who's taking MAT530. The image was then generated 12 geometric landmarks using TI nspire software. The main goal of this research is to classify the gender through face images and to resolve for imbalance data using Hellinger Distance Decision Tree (HDDT) classifier. This classifier was proposed as an alternative to decision tree technique which used Hellinger Distance as the splitting criteria. The result from validation split shows that percentage split at 40% produced the highest value of accuracy rate at 77.2727% and has the most significant value of sensitivity and specificity.

Keywords: *Face Recognition, gender, shape, landmark, HDDT, WEKA*

INTRODUCTION

Face recognition is one of popular technology nowadays that is used in many areas, for example, face recognition system and image monitoring system (Hung et al., 2017). It is a biometric identification technology that has great potential in the future, and great theoretical and practical values. According to Zhu, (2012), face detection techniques are beginning from previous knowledge and combination of features while for others based on the skin segmentation. In Malaysia, the current system has been developed, but the challenges are when one person thinks that the variations in the visual stimulus due to viewing direction or poses, facial expressions, aging such as hair, glasses or cosmetics. According to Wilhelm, (2005), the challenging task in recognition patterns is to estimate gender based on facial images that have many application areas such as the personal identification and verification, video surveillance, and human-computer interaction.

The main problem of face recognition is difficult to identify the shapes of human faces because different people have a different structure of faces. Besides, there are a lot of environmental and personal factors that affect facial appearances such as the presence of eyeglasses, pimples or different facial expressions and the persons' hair which covers their faces.

The main objectives of this study are to classify gender using Hellinger Distance Decision Trees for imbalance data. The Sub-objectives are:

i) To perform Procrustes analysis of the Delaunay Triangulation shape.
ii) To generate Delaunay Triangulation of face images between male and female students.

LITERATURE REVIEW

Procrustes Component Analysis (PCA)

PCA is the most useful in shape correspondence, because of the orthogonal nature of the rotation matrix. Hurley and Cattell who are the first used the term Procrustes analysis in 1962 with a method that did not limit the transformation to an orthogonal matrix (Stegmann, 2002). Procrustes analysis is better to be used in face recognition because it can detect and know on the points and shape, for example, face, either that face is male or female and also it makes it able to identify if that face is smiling, being angry or feeling sad. Shapes and landmarks involved in the Procrustes analysis since both good for detecting and differentiate faces. According to Igual (2014), the Procrustes analysis is a form of statistical shape analysis used to identify the distribution of a set of shapes.

Delaunay Triangulation (DT)

DT is a unique construction that no vertex from any triangle may lie within the circumcircle of any other triangles. The face image is abstracted in terms of the DT which based on barycenters replaces triangles with small successive triangle so that the stop criteria is fulfilled. To recognize the face, the histogram of the final set of areas is used as discrimination. According to (Yi and Hong, 2001), theoretical graphical techniques are often used in clustering planar point sets. The density of the point is measured by Delaunay triangle's edge for the cluster. When the cluster is close to each other, it will form a planar area with any pair of points and if it is between two different clusters, it will have a greater distance than that in the cluster.

Hellinger Distance Decision Trees (HDDT)

A decision tree is a powerful method for classification, predicting and used as a guide for decision making that leads to different outcomes depending on chance, due to their efficiency, simplicity, and interpretability. The decision trees, added with sampling techniques, have proven to be an effective way to address the imbalanced data problems. Despite its effectiveness, however, sampling methods add complexity and parameter selection requirement. Dealing with imbalanced datasets is one of the weaknesses of decision trees. A dataset is called imbalanced if the majority class vastly outnumbers the minority class in the training data. Class imbalance is the biggest challenge as it impedes the ability of the classifier to learn the minority concept due to the nature of learning algorithms. This is a fact when learning under highly imbalanced training data, classifying all instances as negative will result in high classification accuracy. The benefits of HDDT is to improve the accuracy of predictions, the speed, and single-pass through the data. An ensemble of HDDTs is used to combat concept drift and increase the accuracy of single classifier (David et al., 2012).

METHODOLOGY

The image of 53 students of CS248 who's taking MAT 530 was retrieved from the SIMS. It comprised of 10 male and 43 female students. The data was analyzed to differentiate the shape of images of male and female students. The flow of the study was as below:

Step 1: Transfer the images into TI inspire.

Step 2: Extract 12 chosen landmark from the facial images involved in this study such as mouth, nose and others.

Step 3: Generate geometric coordinate for all student's images. Every student has different coordinates because of the varieties of the face structures.

Step 4: Conducting PCA and DT

Step 5: Performing HDDT classification

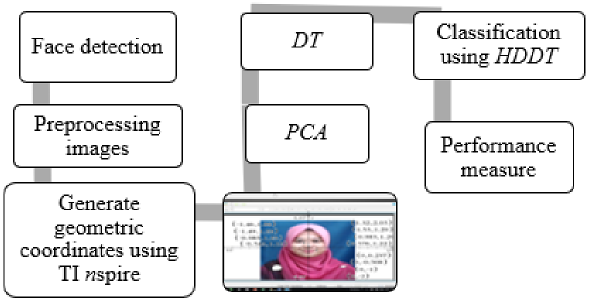
Face detection
Preprocessing images
Generate geometric coordinates using TI *n*spire
DT
PCA
Classification using HDDT
Performance measure

Figure 1: Flowchart of the methodology

FINDINGS AND DISCUSSION

PCA and DT

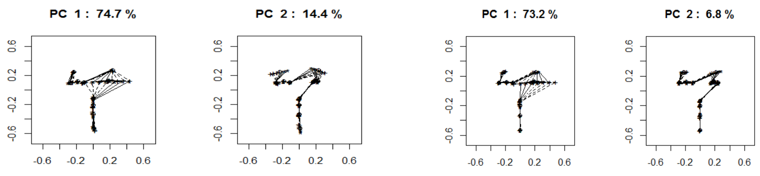

Figure 2: Plots of the first two PCA for male (left) and female (right)

Figure 2 shows the shape of principal component analysis (PCA) between male and female. This result shows that the percentage of variability for the male is greater than the percentage of variability for female.

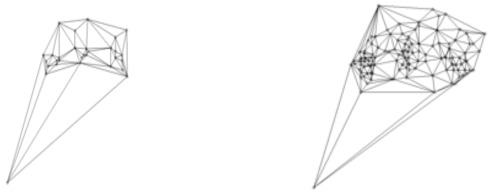

Figure 3: Delaunay Triangulation plot for male (left) and female (right)

There are 12 landmarks located in each image of the students. Each of them has different coordinates. Figure 3 plot the edges adding in male and female triangles respectively. The result shows that there are differences of edges in the triangles of male and female where female datasets produce more edges than male due to a large number of female data than the male data.

Overall Performance Measurement

Table 1: Performance comparison of HDDT for training set

Percentage split	50%		60%		70%		80%		90%	
Gender	M	F	M	F	M	F	M	F	M	F
Male	7	0	7	0	7	0	6	1	8	0
Female	1	18	0	24	3	27	2	33	3	36
Sensitivity	87.50%		100%		70%		75%		72.727%	
Specificity	100%		100%		100%		97.059%		100%	
Accuracy	96.15%		100%		91.89%		92.86%		93.62%	
Kappa statistic value	0.9065		1		0.773		0.7568		0.8083	

Based on all the percentage splits for imbalanced data, it shows that 60% of the percentage split produced 100% of accuracy which means all the gender were correctly classified according to their gender.

Table 2: Performance comparison of HDDT for validation set

Percent split	50%		40%		30%		20%		10%	
Gender	M	F	M	F	M	F	M	F	M	F
Male	1	2	1	2	1	2	0	3	1	1
Female	10	14	3	16	3	10	1	7	1	3
Sensitivity	9.09%		25%		25%		0%		50%	
Specificity	87.50%		88.89%		83.33%		70%		75%	
Accuracy	55.56%		77.27%		68.75%		63.64%		66.67%	
Kappa statistic value	-0.0385		0.1538		0.0909		-0.1579		0.25	

Based on all percentage split, it shows that 40% of the percentage split performed good because it produced the highest value of accuracy.

Receiver Operating Characteristic (ROC) Curve

Table 3 shows the result of ROC value for variety percentage of splitting from 50% to 90%. We can see from Figure 4, the blue line refers the training split which ROC values increase from 50% to 60% of splitting but at 70% of splitting, the ROC value decreased. For the percentage split at 60% which is equal to 1, this indicates that the classification of male and female are perfect, and all image expression has been classified according to their gender.

For validation split which in the red line, as the percentage of splitting were increasing, the ROC value also increases but at 30% of splitting, the ROC value decreased, and these results show that there are some errors involves in classifying the gender.

Table 3: ROC values for various percentage split

Percentage Split	50%	60%	70%	80%	90%
ROC training	0.996	1	0.976	0.978	0.979
ROC validation	0.444	0.544	0.50	0.542	0.625

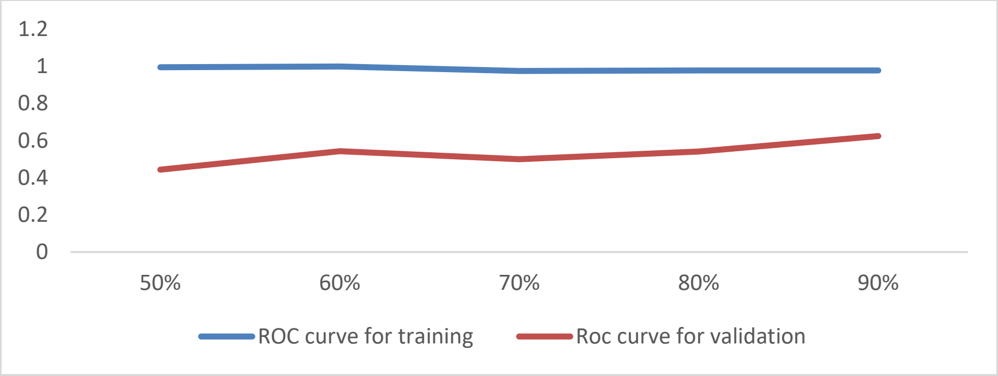

Figure 4: ROC value for various percentage split

CONCLUSION AND RECOMMENDATION

Delaunay triangulation is applied to the face images by adding edges to the landmark to obtain triangulation. From the extracted landmark the points detected from facial images portray scattered data points. The Delaunay triangulation of each face images can be created by connecting the coordinates of each landmark point to its nearest neighbor such that the circumcircle associated with each triangle. From the result, it shows that there are differences of edges in the triangles of male and female where female datasets produce more edges than male due to a large number of female data than the male data.

The classification method used is Hellinger Distance Decision Tree (HDDT). HDDT is a technique in handling the unbalanced problem. Generally, decision trees are popular classification method due to their simplicity, efficiency and interpretability and most of these methods only focus on balanced datasets. Dealing with imbalanced datasets is one of the weaknesses of decision trees. So, the HDDT was introduced to solve this problem. After applying the HDDT classifier, the highest value of the classification rate is 77.2727% and sensitivity is 25%, which refers to 60% of the percentage split.

For future work, to differentiate between genders and to solve an unbalanced problem, the researcher can try to include more attributes that will affect the precision when detecting face images such as facial expression, fingerprints and so on to view how the Decision Tree works and to gain more excellent results. The future researcher may also use, for example, Synthetic Minority Over-sampling (SMOTE) for the imbalanced problem.

ACKNOWLEDGEMENTS

The authors would like to thank Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perlis branch, for permission to conduct this study.

REFERENCES

Cieslak, D. A., Hoens, T. R., Chawla, N. V., & Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, *24*(1), 136–158. https://doi.org/10.1007/s10618-011-0222-1

Hung Yuam Chung, C. C. (2017). Face Detection and Posture Recognition in a Real Time Tracking System . IEEEXplore.

Igual, L., Perez-Sala, X., Escalera, S., Angulo, C., & De la Torre, F. (2014). Continuous Generalized Procrustes analysis. *Pattern Recognition, 47*(2), 659-671.

Stegmsnn, M. J. (2000). *An Investigation into the use of 3D Computer Graphics for Forensic Facial Reconstruction.* Glasgow University.

T. Wilhelm, H.-J. B.-M. (2005). Classification of Face Images for Gender, Age, Facial Expression and Indentity. *Proc. Int. Conf. on Artificial Neural Networks*, (pp. 560-574).

Xiao, Y., & Yan, H. (2001). Facial Feature Location with Delaunay Triangulation/Voronoi Diagram Calculation. *Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing - Volume 11*, *11*, 103–108.

Y. Zhu, C. H. (2012). "Face Detection Method Based on Multi-feature Fusion in YCbCr Color Space". *International Congress on Image and Signal Processing,*, 1249-1252.