

The Comparison of Classification Performance of Machine Learning Techniques Using Principal Components Analysis: PCA for Screening β -Thalassemia

Patcharaporn Paokanta
Department of Knowledge Management,
Collage of Arts, Media and Technology,
Chiang Mai University,
Chiang Mai, Thailand
E-mail: patcha535@gmail.com

Somdat Srichairatanakool
Department of Biochemistry,
Faculty of Medical,
Chiang Mai University,
Chiang Mai, Thailand
E-mail: ssrichai@med.cmu.ac.th

Napat Harnpornchai
Department of Knowledge Management,
Collage of Arts, Media and Technology,
Chiang Mai University,
Chiang Mai, Thailand
E-mail: napatresearch@gmail.com

Michele Ceccarelli
Department of Biology,
Faculty of Science,
Sannio University,
Benevento, Italy
E-mail: ceccarelli@unisannio.it

Abstract—Feature Selection plays an important role in many areas especially in classification tasks. It is also an important pre-treatment for every classification process. Not only decreasing the computational time and cost, but selecting an appropriate variable is also increasing the classification accuracy. In this research, the comparison of classification performance of machine learning techniques using Principal Components Analysis (PCA) for screening the genotypes of β -Thalassemia patients will be proposed. The aim of this study is to reduce the dimentions of data before classification. According to using PCA, the classification results show that the Multi-Layer Perceptron (MLP) is the best algorithm, providing that the percentage of accuracy reaches 86.61, K- Nearest Neighbors (KNN), NaiveBayes, Bayesian Networks (BNs) and Multinomial Logistic Regression with the percentage of accuracy 85.83, 85.04, 85.04 and 82.68. In the future, we will search for the other feature selection techniques in order to improve the classification performance such as the hybrid method, filtering method etc.

Keywords— β -Thalassemia, Classification Techniques, Principal Components Analysis (PCA), Feature Selection, Machine Learning Techniques.

I. INTRODUCTION

Over the past decades, machine learning has tried to adapt algorithms to a specific task. Learning algorithms are widely used for many tasks both in industry and in academia such as face recognition, text classification and credit card fraud detection etc. especially, a classification task in medical diagnosis [1]. The research on classification accuracy is aimed at building an efficient model to predict the class memberships of data, to produce a correct label on training data, and to predict the label of any unknown data

correctly [2]. One of these machine learning techniques that can be used to improve the data analysis of high dimensional data is a feature selection technique [3]. It is the technique which is able to reduce the cost of recognition and often provide the better classification accuracy by reducing the number of features [4]. Multivariate statistical methods such as a Principal Components Analysis (PCA), partial least squares and more recently independent component analysis have been developed and applied for this purpose. Comparing to those mentioned, PCA is the most popular one, which relates to its conceptual simplicity [5].

Nowadays, Principal Components Analysis (PCA), or the subspace method, has been extensively investigated in the field of computer vision and pattern recognition. One of the attractive characteristics of PCA is that a high-dimensional vector that can be represented by a small number of orthogonal basis vectors, i.e. the principal components. The conventional methods of PCA, such as singular value decomposition (SVD) and eigen-decomposition, are performed in batch mode with a computational complexity of $O(m^3)$ where m is the minimum value between the data dimension and the number of training examples. Undoubtedly these methods are computationally expensive when dealing with large-scale problems where both the dimension and the number of training examples are large [6].

From the study of Yulan Liang et.al, they proposed the review of both general feature reduction approaches for high dimensional correlated data and more specific approaches for Single Nucleotide Polymorphism (SNPs) data, which include unsupervised haplotype mapping, tag SNP selection, and supervised SNPs selection using statistical testing/scoring, statistical modeling and machine learning methods with an emphasis on how to identify interacting loci. Moreover they illustrated that the Principal Components Analysis is a suitable tool for categorical SNPs information which is arguable, since it has been more appropriate for the continuous scale data [7]. The research of Hong-Qiang Wang et. al. presented that the other two main approaches for the

feature transformation: Principal Components Analysis (PCA) and partial least squares (PLS). PCA can be taken as an unsupervised learning and PLS as a supervised learning. Having compared PCA and PLS, Nguyen et al. drew a conclusion that PLS trends to be more competitive for cancer classification owing to the supervised learning ability. The main drawback of the separate strategy of the dimensionality reduction is that the strategy might cause the loss of the useful information. Furthermore, they introduced a new probabilistic technique to extract the gene regulation information from the microarray data for the cancer classification using k-nearest neighborhood (k-NN), Fisher linear discriminant (FLD) and SVM [8]. Besides, the research of S.R. Amendolia et al. showed Thalassemia screening indicators by using the Principal Components Analysis (PCA) which the selected features are RBC, Hb, Ht and MCV. They compared the study of K-Nearest Neighbor, Support Vector Machine and Multi-Layer Perceptron for Thalassemia screening [9]. The proposes of this paper aim to focus on the classification of β -thalassemia with the new data set which is different from the other researches by using several algorithms of machine learning.

In this research, we, firstly, introduce the use of PCA to classify the β -thalassemia patients and compare the classification performance of several machine learning techniques for example, Multi-Layer Perceptron (MLP), K-Nearest Neighbors (KNN), NaiveBayes, Bayesian Networks (BNs) and Multinomial Logistic Regression. The remainder of the paper is organized as follows. In Section 2, the relevant machine learning background will be introduced to present PCA and the performance of classification. For the Section 3, we describe the materials and methodology of using PCA and machine learning techniques which is applied to the β -thalassemia data set. In Section 4, presents the results of using PCA before training and classifying by applying many classification algorithms are presented in term of the comparison between their accuracy percentages of them. Finally, we shall conclude this paper and discuss some future research issues in Section 5.

II. MATERIALS AND METHODOLOGY

A. Materials

Thalassemia syndromes can be one of the most common genetic disorders in the world, and an estimation of 1.5% of the worldwide population diagnosed with β -Thalassemia. The disorder is recognized in the areas where malaria used to be widespread, such as Africa, the Mediterranean region, the Middle East, Southeast Asia (India, Thailand and Indonesia), and the Far East. It is found so often in the Southeast of Asia, where there are approximately 55 millions are carriers. The gene frequencies of alpha-Thalassemia reach 30-40% in Northern Thailand and Lao PDR, of β -Thalassemia varies between 1- 9%, and HbE has a frequency of 50-60% at the junction of Thailand, Lao PDR. Approximately 40% of Thai people are heterozygous carriers of these genes [10]. From this statistics, the data of β -Thalassemia patients 127 records were collected and used for classifying the genotypes of them which identify the types of β -Thalassemia. These data

were obtained from hospitals in the Northern Part of Thailand. The data set for this experiment is as follows,

TABLE I. THE DATA SET OF β -THALASSEMIA

Variables	Diecton
Genotype of children	Output
F-cell of children	Input
HbA2 of children	Input
HbA2 of father	Input
HbA2 of mother	Input
Genotype of father	Input
Genotype of mother	Input

There are 7 variables which were elicited from some experts (biochemist and medical practitioner in the hospital) and some documents (Out Patien Department Record). Some variables in this table were reduced the dimation by PCA which is a popular reduction technique. It can reduce the dimationality of data set whilst retaining as much information as possible. Fourthermore, it computes a concise and an optimal description of the data set also. The results of this technique will be demonstrated in the next section.

B. Methodology

In an effort to focus on the comparison of classification performance of machine learning techniques aspects of β -Thalassemia screening study, a data set of 127 β -Thalassemia patients were used for this study.

Figure 1 represents the methodology of feature selection and classification processes. In order to the processes of the comparison of classification performance, in the first steps, the variables were defined and collected by Diagnosis template in CommonKADs suite. They were elicited through the experts and related documents. Next steps, these variables were transformed by the feature selection technique which is PCA to reduce the dimation of data set. Afterwords, in the third steps, Machine Learning Techniques for example Multilayer Perceptron (MLP), K- Nearest Neighbors (KNN), NaiveBayes, Bayesian Networks (BNs), Multinomial Logistic Regression etc.were used to classify the genotype of β -thalassemia. Finally, the results of classification, accuracy percentage that were obtained from these algorithms, were compared that are shown in the next section.

The result of the first steps is β -thalassemia variables and in the second steps, the variables of the previous steps were reduced the data dimensions by PCA. Moreover, the results of using this technique were represented in the component matrix, eigenvalues and eigenvectors. The accuracy percentage of using several classification algorithms were obtained at the third steps and the results of these steps were compared for classification performance of β -thalassemia. Moreover, the training and testing data set was separated for classification through a ten fold cross validation techniques. The results of each step will be demonstrated in the next section.

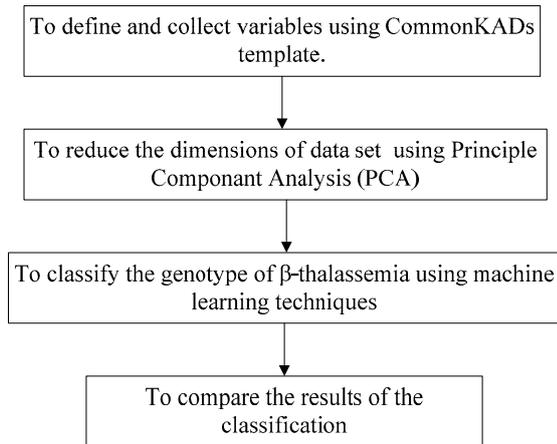


Figure 1. The methodology for classifying β -Thalassemia

III. RESULTS

According to the methodology for classifying β -thalassemia in Figure 1, the result of the first steps is the data set of β -thalassemia which was obtained in elicitation process using the diagnosis template in CommonKADs suite of Knowledge Engineering Process through the expert interview and document review.

For the results of the second steps, the filtering techniques, PCA was used to reduce the data dimension which obtained from the first steps by combining the related variables to be the new variables called factors. In this data set, it can reduce the dimensions of data from 6 to 3 variables as the result shows in Table II. This techniques use the diagonal value of correlation matrix which is 1 to be the initial value for calculating the communities. In PCA, the factor loading is play important role because it is refer to the relational value between variables and factors which they should be grater than [0.5]. Because if any variables have the high loading in a factor, it means that this variable should be in this factor. The factor loading can be obtained from the diagonal of Eigen matrix or component matrix that shown in Table III.,

TABLE II. THE REDUCED DATA DIMENSION SET USING PCA FOR β -THALASSEMIA CLASSIFICATION

Variables	Direction
Genotype of children	Output
Factor 1	Input
Factor 2	Input
Factor 3	Input

TABLE III. THE COMPONENT MATRIX

Variables	Components		
	1	2	3
Genotype of mother	0.934	0.116	
Genotype of father	-0.916	0.158	
HbA2 of father	-0.715	0.412	
HbA2 of mother	0.693	0.405	

HbA2 of children		0.826	-0.316
F-cell of Children		0.758	0.948

Table III. shows that 4 variables should be in component 1 with the factor loading 0.934, 0.916, 0.715 and 0.693 respectively. On the other hand, there is a variable should be in component 2 with the factor loading 0.826 and there is only a variable in component 3 with the factor loading 0.948.

TABLE IV. THE ROTATED COMPONENT MATRIX

Variables	Components		
	1	2	3
Genotype of mother	0.926		
Genotype of father	-0.920	0.202	
HbA2 of father	0.750	0.318	0.133
HbA2 of mother	-0.654	0.447	0.142
HbA2 of children		0.886	
F-cell of Children			0.983

Form the rotated component matrix, Table IV. which can be rotated by Varimax with Kaiser Normalization, it confirms that for this data set should be 3 components. The component plot was represented in Figure 2.

TABLE V. THE EIGEN VALUES

Component	Eigen values
Component 1	2.715
Component 2	1.121
Component 3	1.002
Component 4	0.741
Component 5	0.287
Component 6	0.134

Table IV. shows that the Eigen values of component 1-6 are 2.715, 1.121, 1.002, 0.741, 0.287 and 0.134 respectively.

From Figure 3. It shows that the Eigen values are the summary of variance of all variables in each factor. In the factor analysis, the 1st order factor is able to separate the variance away from other factors therefore it has the heighted Eigen values more than the other order of factors. Generally, the Eigen values should be grater than 1. From the Eigen values of this data, there are only component 1, 2 and 3 that they are a suitable factor because the Eigen values are more than 1.

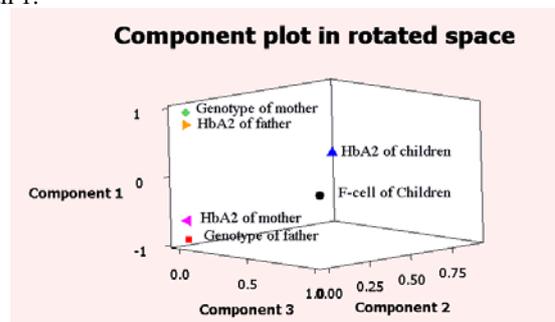


Figure 2. Component plot in rotated space

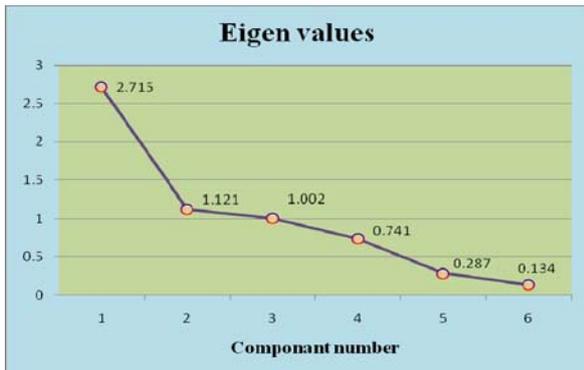


Figure 3. The Eigen values

The results of the third steps are the comparison of classification performance (accuracy percentage) that shows in Table V.

The accuracy percentage can be defined as follows,

$$\text{Accuracy Percentage} = (\text{TC}/\text{N}) * 100$$

Where N is a total number of test cases, TC is a total number of subjects correctly classified.

TABLE VI. THE ACCURACY PERCENTAGE OF CLASSIFICATION USING PCA AND MACHINE LEARNING TECHNIQUES

Classification techniques	Percentage of accuracy
1. Multi-Layer Perceptron (MLP)	86.6142
2. K- Nearest Neighbors (KNN)	85.8268
3. NaiveBayes	85.0394
4. Bayesian Networks (BNs)	85.0394
5. Multinomial Logistic Regression	82.6772

Finally, Table V. represents the result of the fourth steps which is the comparison of classification performance of machine learning techniques using PCA for screening β -thalassemia. According to Table V, the best algorithm using PCA on this data set is MLP, KNN, NaiveBayes, BNs and Multinomial Logistic Regression with the accuracy percentage of 86.6142, 85.8268, 85.0394, 85.0394 and 82.6772 respectively.

Even though the obtained results of using PCA is proper to filter features on β -thalassemia data if compared to the using Chi square for the feature selection on the same data set [11]. The better accuracy percentage than PCA can be obtained because some variables in this data set can be transformed to nominal and interval scale and Chi square can be used as well with nominal scale. Therefore, from this result, the appropriate feature selection technique for screening β -thalassemia on this data set is Chi square.

IV. CONCLUSION

Using PCA for the feature selection task can be obtained the satisfied accuracy percentage. As the results show that the best algorithms on this data set is MLP with the accuracy percentage 86.6142. Moreover, the other algorithms obtained

the results as KNN, NaiveBays, BNs and Multinomial Logistic Regression with the accuracy percentage of 85.8268, 85.0394, 85.0394, 82.6772. From these results, they implies that PCA can be used as well but if comparing to Chi square which is a feature selection, Chi square can produce the better accuracy percentage. On the otherhand, for medical data, it needs the highest accuracy percentage therefore in the future work, the other algorithms which is feature selection techniques will be studied for example the hybrid feature selection methods: fuzzy, genetic algorithm etc. for β -thalassemia classification.

As the involvement of this study, there is a few researchs on the application of data mining techniques for thalassemia data. Form this reason, the combination of feature selection techniques for classification on this data set which it is difference from the other thalassemia data set will be studied for next works.

ACKNOWLEDGEMENT

The authors would like to thank SOMMALADA LERTROMYANANDA for improving English language in this paper and I feel appreciate with her kindly support.

REFERENCES

- [1] Amir Navot. "On the Role of Feature Selection in Machine Learning." Ph.D. thesis, Hebrew University of Jerusalem, Israel, 2006.
- [2] Cheng-San Yang, Li-Yeh Chuang, Chao-Hsuan Ke, and Cheng-Hong Yang, Member and LAENG. "A Hybrid Feature Selection Method for Microarray Classification." IAENG The International Journal of Computer Science, vol. 35, Issue 3, Aug. 2008.
- [3] Supoj Hengpraprom and Prabhas Chongstitvatana. "Feature Selection By Weighted-SNR For Cancer Microarray Data Classification." International Journal of Innovative Computing, Information and Control, vol. 5, pp. 4627-4635, Dec. 2009.
- [4] Caio Soares. "A Class-Specific Ensemble Feature Selection Approach For Classification Problems." M.S. thesis, Auburn University, USA, 2009.
- [5] Xueqin Liu, Uwe Kruger, Tim Littler a, Lei Xie and Shuqing Wang. "Moving Window Kernel PCA For Adaptive Monitoring of Nonlinear Processes." Chemometrics and Intelligent Laboratory Systems, vol. 96, pp.132-143, Apr. 2009.
- [6] Yongmin Li. "On incremental and robust subspace learning." Pattern Recognition, vol. 37, pp. 1509-1518, Jul. 2004.
- [7] Yulan Liang. "Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex Diseases." Statistics Surveys, vol. 2, pp. 43-60, 2008.
- [8] Hong-Qiang Wang, Hau-San Wong, De-Shuang Huang and Jun Shuc. "Extracting gene regulation information for cancer classification." Pattern Recognition, vol. 40, pp. 3379-3392, Dec. 2007.
- [9] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala and G.M. Murac. "A comparative study of K-Nearest Neighbour, Support Vector Machine and Multi-Layer Perceptron for Thalassemia screening." Chemometrics and Intelligence Laboratory Systems, vol. 69, pp. 13 – 20, Nov. 2003.
- [10] Valairat Dhamcharee, Orasri Romyanan and Tanimporn Ninlagarn. "Genetic Counseling for Thalassemia in Thalind : Problems and Solutions." Southeast Asian Journal Trop Med Public Health, vol. 32, pp. 413 – 418, Jun. 2001.
- [11] Patcharaporn Paokanta, Michele Ceccarelli and Somdat Srichairatanakool. "The Efficiency of Data Types for Classification Performance of Machine Learning Techniques for Screening β -Thalassemia," in *Proc. ISABEL*, 2010, pp. 1-4.