β-Thalassemia Knowledge Elicitation Using Data Engineering: PCA, Pearson's Chi Square and Machine Learning

P. Paokanta

Abstract—**Data Engineering is one of the Knowledge Elicitation and Analysis methods, among serveral techniques; Feature Selection methods play an important role for these processes which are the processes in data mining technique esspecially classification tasks. The filtering process is an important pre-treatment for every classification process. Not only decreasing the computational time and cost, but selecting an appropriate variable is increasing the classification accuracy also. In this paper, the Thalassemia knowledge was elicited using Data engineering techniques (PCA, Pearson's Chi square and Machine Learning). This knowledge presented in form of the comparison of classification performance of machine learning techniques between using Principal Components Analysis (PCA) and Pearson's Chi square for screening the genotypes of β-Thalassemia patients. 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. On the other hand, these results were compared to the Pearson's Chi Square and presented that….**

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 mathod etc.

Index—**Knowledge elicitation, data engineering, feature selection, principal component analysis (PCA), pearson's chi square, machine learning, β-thalassemia.**

I. Introduction

Over the past decades, Knowledge and Data Engineering (KDE) have been applied extensively in many areas such as engineering, sciences, medical [1], [2]. The various issues of KDE include [3].

- Representation and manipulation data and knowledge
- Architecture of database, expert, Knowledge-based system
- Construction of data/knowledge-based
- Aplicaions: Data administration, knowledge engineering
- Linguistic tools
- Comunication aspects

Among related various KDE techniques, Data Mining (DM) or Knowledge Discovery in Database (KDD), which is one of the popular research issues, applied to data analysis in the last 50 years [4]. The concept of this technique is the process to extract knowledge from the huge data [5]. This technique is not only the combination of Machine Learning, and Statistics but related to database system and various disciplines also. There are 4 main types of DM tasks, including Association rules, Clustering, Classification, and Summarization. Among DM tasks, classification task is well known for solving complex problems especially, medical diagnosis.

Nowadays, β-Thalassemia, the common genetic disorder can be found around the world especially, in Asia and Thailand. As the permanent problems of the Thalassemia genetic widespread which are the complex diagnostic processes, lack of genetic counselors, lack of expert knowledge, expensive equipments the time consuming in labolatory and a few research on Thalassemia. For these problems, KDE techniques including PCA, Pearson's Chi square and Machine Learning will be applied to elicit β-Thalassemia knowledge and compared the obtained results to the results of the other techniques in the previous research.

In this paper, firstly, the introduction of the Thalassemia elicitation using DE (PCA, Pearson's Chi square and Machine Learning) will be reviewed. Next, KDE will be illustrated in terms of the theory and historical review. In the third section, Materials and Methodology of this study will be proposed and the fourth section the results of this study will be presented as the comparison of using Pearson's Chi square, PCA and Machine Learning algorithms for classifying the β-thallassemia patients to the obtained results of the classification performance form the previous research for example, Multi-Layer Perceptron (MLP), K- Nearest Neighbors (KNN), NaiveBayes, Bayesian Networks (BNs) and Multinomial Logistic Regression. Finally, we shall conclude this paper and discuss some future research issues in Section 5.

II. Feature Selection in KDE

Knowledge and Data Engineering is one of the methodologies for sloving complex problems which it can provide the excellently solutions [6]. Among several proceses of KDE, the permanent method in KDE process is Feature selection technique. This method is the filtering variable technique that the concept based on the finding correlation (relationships) between variables. There are 4 main approaches for feature selection algorithms including filter method, wrapper methods, embeded methods and

Manuscript received June 17, 2012; revised August 3, 2012.

Patcharaporn Paokanta is with the Development, System Analysis and Design, and Information Technology at the College of Arts, Media and Technology, Chiang Mai University (CMU), Thailand.

hybrid methods [7]. The advantages of this technique are not only selecting the appropriate variables but also reducing the run time consuming. Moreover, these techniques have widely popular used in DM or KDD processes. The feature selection techniques are used to select the appropriate features in the initial process of DM called the selecting process and these selected variables are provided in the transformation prcess in the next step. Then the transformed data is sent to the data mining process to find the patterns of data by using clustering, classification, association rules and summarization algorithms depend on the purposes of those studies and these patterns are interpreted to discover knowledge in the final step [8].

In the third section, theory of Pearson's Chi square and PCA are illustrated for filtering related variables.

III. Pearson's Chi Square vs. PCA

A. *Pearson's Chi Square*

In case of the data is not normal form, the non-parametric statistics is selected to parameter and Googness of fit test. Among feature selection techniques such as Runs, Binomial, Kolmogorov-Smirnov, Mann-Whitney, Wald-Wolfowiz, Sign, Wilcoxon, Median, Kruskal Wallis, Friedman, McNemar, Cochran Q etc., Pearson's Chi square is a popular non-parametric method to parameter and normal distribution test for data in norminal scale. Moreover, this technique is used for the independent test. The equation of the test-statistic shown below,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

by E_i can be calculated as

$$E_i = \frac{N}{n}$$

and x^2 = Pearson's cumulative test statistic, which asymptotically approaches a x^2 distribution.
O_i= an observed frequency;
E_i = an expected frequency, affirmed by the null hypothesis;
n = the number of cells in the table.

The chi-squared statistic can be transformed to a p-value by comparing the value of the statistic to a chi-squared distribution. The number of degrees of freedom is calculated by the number of cells n, minus the reduction in degrees of freedom p.

B. *Principal Component Analysis (PCA)[9], [10].*

Factor Analysis is the technique that groups the related variables to the same group. The variables that have the high relationship will be in the same factor and the directions of the relationship may possibly be positive (the same direction) or negative (not the same direction) direction. On the other hand, variables that are in the defferent factor, it means that these variables are not related.

The objectives of Factor Analysis are not only to reduce the number of variables by grouping related variables to the same factor (The number of factor must be less than the number of varibles) but also to confirm the accuracy of weighted values of variables in some research.

The linear equation of Factor Analysis shown below,

$$F_j=W_{j1}X_1+W_{j2}X_2+\ldots\ldots+W_{jn}X_n+e$$

and
X_j = j^{th} variable
W_j= j^{th} coefficent

Moreover, the linear combination of X_i in each factor can be derived as

$$Z_1=L_{11}F_1+L_{12}F_2+\ldots\ldots+L_{1m}F_m+e_1$$

$$Z_2=L_{21}F_1+L_{22}F_2+\ldots\ldots+L_{2m}F_m+e_2$$

$$Z_n=L_{n1}F_1+L_{n2}F_2+\ldots\ldots L_{nm}F_m+e_n$$

and
Z_i = the standardized X_j variable; j=1, 2, …., n
n = the number of variables
m = the number of factors; $m < n$
F_1,….., F_m = common factors
e = error

In the fourth section, the concerned materials for screening β-Thalassemia is presented.

IV. Materials

β-Thalassemia, a common genetic disorder, which can be found around the world in particular in the area as same as Malalia where the main areas locate in Asia pacific including Myanmar, Indonesia, Malesia, Cambodia, Vietnam, Loas and Singapore. Especially, Thailand, there are many types of thlassemia as the reason that the mutation of its genetics and statistic shows that 12125 new babies born with this disease. For this reason, 127 β-Thalassemia data sets form the hospital in Northern Thailand were collected to classify types of β-Thalassemia using KDE techniques (Pearson's Chi square, PCA and Machine Learning) [11]. The 7 indicators for diagnosis this disease shown in Table I below,

TABLE I: The Data sets of B-Thalassemia

Variables	Directions
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

In the next section, the methodology for this study is reviewed.

V. Methodology

The methodology for discovering Thalassemia knowledge using KDE (Pearson's Chi square, PCA and Machine Learning) presented in Fig. 1. belows,

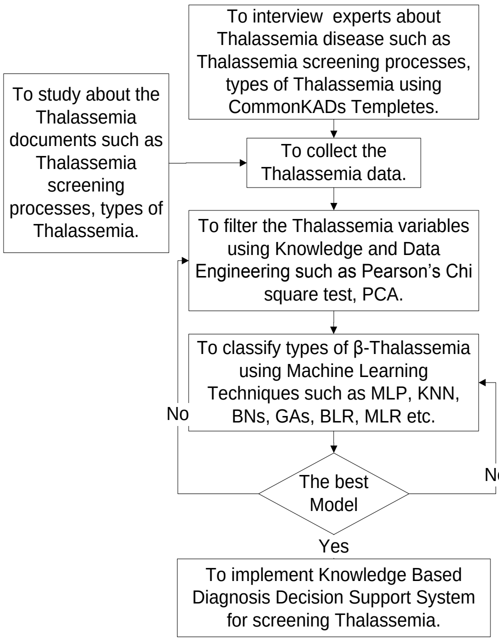

Fig. 1. The methodology for discoverying thalassemia.

According to the procedure of this study, the knowledge elicitation processes start at the first step, the related Thalassemia knowledge were captured from expert such as biochemistry, nurse, and medical practitioner and documents by using CommonKads templete (Diagnostic templete). These were used to define the Tlassemia indicators that shown in Table I. Then the 127 β-Thalassemia data sets were collected from the Northern hospital in Thailand. In the Third step, colleted data were filtered to select the appropriate variables for screening β-Thalassemia through the classification algorithms such as MLP, KNN, BNs, BRL etc. The obtained results of these algorithms were compared to select the best algorithm to develop the Thalassemia Knowledge-Based Diagnosis Decision Support System.

In the next section, the results of each step were illustrated as the comparison of classification performance of using feature selection methods and Machine learning Techniques.

VI. Results

The obtained results of each process are reviewed in this section. The comparison of classification performance (Acuracy percentage) of using MachineLearning algorithms between Pearson's Chi square and PCA is demonstrated as the different data types including interval and norminal scale that present in Table II. In addition, the run time comparison of these algoritms is presented in Table II also.

The classification performance (Accuracy Percentage) of using Pearson Chi-square and Machine Learning techniques shows that the best algorithm is KNN with 88.97% of accuracy. On the other hand, MLP, NaiveBayes, BNs and MLR obtained 87.40%, 84.25%, 83.46% and 81.89% respectively for Interval scale. In case of Norminal scale, there are tree best algorithms including, KNN, MLP, and BNs with 85.83%. For MLR and NaiveBayes reached 84.25% and 84.47% respectively (Fig. 2).

The comparison of execution time of using Pearson Chi-square and Machine Learning techniques shows that fastest algorithm is KNN and BNs for interval scale with the run time 0.00 Minute as same as Norminal scale (Fig. 3).

These comparisons reveal that KNN is the best algorithm in both interval and norminal scale in case of using Pearson Chi-square and Machine Learning techniques.

TABLE II: Accuracy Percentage of Using Pearson's Chi Square and Machine Learning Techniques

Method	HbA2 (Interval scale)		HbA2 (Norminal scale)	
	Accuracy (%)	Time	Accuracy (%)	Time
KNN	88.9764	0.00	85.8268	0.00
Multilayer Perceptron	87.4016	1.28	85.8268	1.20
NaiveBayes	84.252	0.02	83.4646	0.02
Bayesian Networks	83.4646	0.00	85.8268	0.00
Multinomial Logistic Regression	81.8898	0.25	84.252	0.16

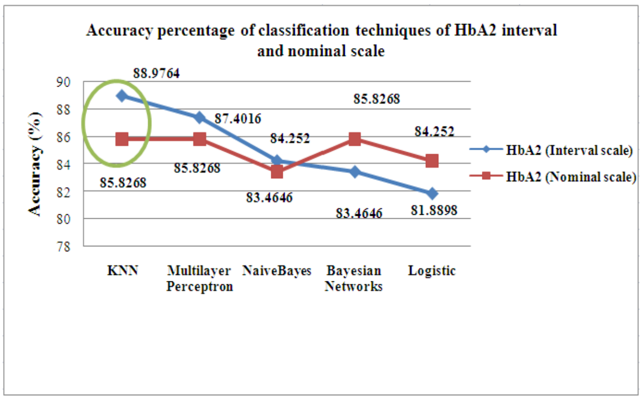

Fig. 2. The accuracy percentage of using pearson's chi square and machine learning techniques.

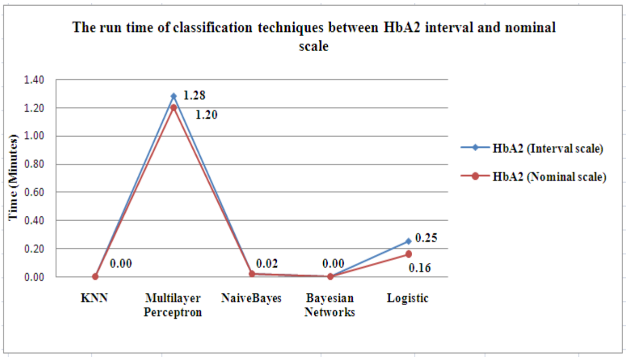

Fig. 3. The run time of using pearson's chi square and machine learning techniques

The classification performance (Accuracy Percentage) of using PCA and Machine Learning techniques shows that the best algorithm is MLP with 86.61% of accuracy. On the other hand, KNN, NaiveBayes, BNs and MLR obtained 85.83%, 85.04%, 85.04% and 82.68% respectively for interval scale. In case of norminal scale, there are two best algorithms including, KNN, and MLP with 85.83%. BNs and MLR reached 83.47% and NaiveBayes 81.10% (Fig. 4).

The comparison of execution time of using PCA and Machine Learning techniques shows that fastest algorithm is KNN and NaiveBayes for interval scale with the rum time 0.00 Minute, for Norminal scale, they are KNN and BNs with the run time 0.00 Minute (Fig. 5).

These comparisons reveal that KNN is the best algorithm in both interval and norminal scale in case of using PCA and Machine Learning techniques.

TABLE III: ACCURACY PERCENTAGE OF USING PCA AND MACHINE LEARNING TECHNIQUES

Method	HbA2 (Interval scale)		HbA2 (Norminal scale)	
	Accuracy (%)	Time	Accuracy (%)	Time
KNN	85.8268	0.00	85.8268	0.00
Multilayer Perceptron	86.6142	6.05	85.8268	9.84
NaiveBayes	85.0394	0.00	81.1024	0.02
Bayesian Networks	85.0394	0.30	83.4646	0.00
Multinomial Logistic	82.6772	0.50	83.4646	0.48

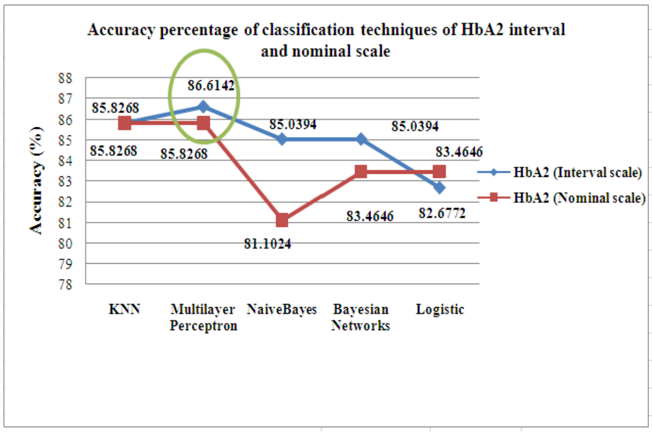

Fig. 4. The accuracy percentage of using PCA and machine learning techniques.

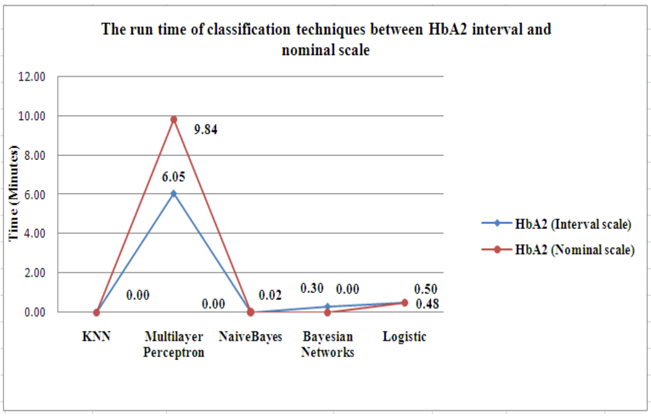

Fig. 5. The run time of using PCA and machine learning techniques.

VII. CONCLUSIONS

In conclusion, the comparison of using KDE (Pearson's Chi square, PCA and Machine Learning Techniques) for screening β-Thalassemia reveals that the best algorithm of using Pearson's Chi square obtained the best classification performance (KNN, 88.97%) in case of interval scale compare to the results of using PCA and the algorithms that obtained the fastest run time include KNN, BNs and NaiveBayes. In the previous paper, there are many classification algorithms that were used to classify on this data set for example the using Reasoning metrics, Polychromatic set and BNs for representing Tlassemia knowledge in graphical model [7], [12]. Moreover, Binomial Logistic Regression based on Classical (Maximun Likelihood) and Bayesian (MCMC) statistics were used to classify these types of Thalassemia and obtained the complacent results (More than 90% of accuracy in each type) [13], [14]. In addition, the rule induction techniques such as C5.0 and CART were used to elicit the Thalassemia knowledge to improve the quality of data to achieve the objectives of this study [15].

In the future, the hybrid classification methods such as BLR-SVM and Fuzzy-GAs and etc. were applied on this data set to increase the classification performance.

ACKNOWLEDGMENT

I would like to thank Associate Professor Dr. Somdech Srichairattanakool, Assistant Professor Dr. Napat Harnpornchai, Associate Professor Dr. Michele Ceccarelli, Dr. Nopasit Chakpitak and Professor Stefano Pagnotta for some suggestions on this topic.

REFERENCES

[1] B. J. Garner, G. J. Ridley, and P. J. Lowe, "Data engineering for neural net analysis of glass furnace characteristics," in *Proc. The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, 1993, pp. 317 – 320.

[2] P. Turney, "Data engineering for the analysis of semiconductor manufacturing data," *IJCAI-95 Workshop on Data Engineering for Inductive Learning*, 1995, pp.1-10.

[3] P. E.-S. Chen and R. E V. D. Riet. "Introduction to the special issue celebrating the 25th volume of data and knowledge engineering: DKE," *Data and Knowledge Engineering*, vol. 25, pp. 1-9, 1998.

[4] C. Chen , Il-Y. Song, X. Yuan, and J. Zhang, "The thematic and citation landscape of data and knowledge engineering (1985–2007), *Data and Knowledge Engineering*, vol. 67, pp. 234–259, 2008.

[5] M. S. Chen, J. Han, and P. S. Yu, "Data Mining: an overview from a database perspective," *IEEE Trans. On Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866-883, 1996.

[6] C. V. Ramamoorthy and B. W. WAH, "Knowledge and data engineering," *IEEE Trans. On Knowledge and Data Engineering*, vol. 1, no. 1, pp. 9-16, 1989.

[7] P. Paokanta, "Reasoning matrices and polychromatic set for screening thalassemia," *Journal of Medical Research and Science*, vol. 1, no. 1, pp. 144-152, 2012.

[8] S. Westendorf, "Literature Review on Knowledge Engineering, Data Clustering and Computational Creativity," pp. 1-37, June, 2012. [Online]. Available: http://dmu-ca.ioct.dmu.ac.uk/publication/papers/literature_review_sascha.pdf.

[9] P. Paokanta, N. Harnpornchai, S. Srichairattanakool, and M. Ceccarelli, "The Comparison of classification performance of machine learning techniques using principal components analysis: PCA for screening β-thalassemia," in *Proc. IEEE ICCSIT's 11 The 4th IEEE International Conference on Computer Science and Information Technology*, 2011, pp. 316-319, Chengdu, China.

[10] P. Paokanta, N. Harnpornchai, S. Srichairatanakool, and M. Ceccarelli, "The knowledge discovery of β-thalassemia using principal components analysis: PCA and machine learning techniques," *International Journal of e-Education, e-Business, e-Management and e-Learning: IJEEEE*, vol.1, no.2, pp. 175-180, 2011.

[11] P. Paokanta, M. Ceccarelli, and S. Srichairatanakool, "The effeciency of data types for classification performance of machine learning techniques for screening β-thalassemia," in *Proc. ISABEL's 10 The 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, 2010, pp. 1-4, Rome, Italy.

[12] P. Paokanta and N. Harnpornchai, "Risk analysis of thalassemia using knowledge representation model: diagnostic bayesian networks," in *Proc. IEEE-EMBS BHI' 12 International Conference on Biomedical and Health Informatics*, 2012, pp. 61-61, Shenzhen, Hong Kong.

[13] P. Paokanta, N. Harnpornchai, and N. Chakpitak, "The classification performance of binomial logistic regression based on classical and

bayesian statistics for screening β-thalassemia," in *Proc. ICMIA' 11 The 3rd International Conference on Data Mining and Intelligent Information Technnology*, 2011, vol. 2, pp. 427-432, Macau, China.

[14] P. Paokanta, N. Harnpornchai, N. Chakpitak, M. Ceccarelli, and S. Srichairatanakool, "Parameter estimation of binomial logistic regression based on classical (maximum likelihood) and bayesian (mcmc) approach for screening β-thalassemia," *International Journal of Intelligent Information Processing: IJIIP,* vol. 3, no. 1, pp. 90-100, 2012.

[15] P. Paokanta, M. Ceccarelli, N. Harnpornchai, N. Chakpitak, and S. Srichairatanakool, "Rule induction for screening thalassemia using machine learning techniques: C5.0 and CART," *ICIC Express Letter: An International Journal of Research and Surveys*, vol. 6, no.2, pp. 301-306, 2012.

Patcharaporn Paokanta is a member of IEEE and a lecturer in the area of Data Management, E-Commerce, Rapid Application and Development, System Analysis and Design, and Information Technology at the College of Arts, Media and Technology, Chiang Mai University (CMU), Thailand. She is studying for a Ph.D. in Knowledge Management and obtained her M.S. in Software Engineering in 2009 from the College of Arts, Media and Technology, CMU, Thailand. In addition, she obtained a B.S. in Statistics from the Faculty of Science, Chiang Mai University, Thailand, in 2006. She was awarded an ERASMUS MUNDUS scholarship (E-Link Project) to study and perform research at the University of Sannio in Italy for 10 months. Her research interests include data mining, machine learning, statistics, biomedical sciences, buyer behaviour, knowledge management, risk management, software engineering, expert systems, artificial and computing intelligence, and applied mathematics. Patcharaporn Paokanta has published articles in international journal and conference proceedings, including ICIC Express Letter: An International Journal of Research and Surveys, International Journal of e-Education, e-Business, e-Management and e-Learning (IJEEEE), International Journal of Intelligent Information Processing (IJIIP), SKIMA 2009, ISABEL 2010, ICCSIT 2011, ICMIA 2011, and BHI 2012.