Employing Data Mining to Identify the Significant Rules for Classifying Body Types

CHIH-HUNG HSU[1], SU-CHIN CHEN[2], BOR-SHONG LIU[3]
[1]Department of Industrial Engineering and Management, Hsiuping Institute of Technology, Taiwan, R.O.C.
[2]Department of Industrial Engineering and Management, St. John's University, Taiwan, R.O.C.
[3]Department of Fashion Imaging, Ming-Dad University, Taiwan, R.O.C.

Abstract: - The goal of this study was to identify the significant rules of human body types from the anthropometric data of adult males, using novel two stage-based data mining procedure. The development procedure included two phases. First, cluster analysis was conducted to sort cases into clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Second, the decision tree uses rules created in accordance with input variables and is executed with data classification by a tree type demonstration to extract the most significant factors and the significant rules based on the results of cluster analysis. Certain advantages may be observed when the significant rules are identified, using two stage-based data mining procedure. Body types could be accurately classified for physiology, medical treatment, sports talent and garment manufacturing according the newly classification rules. The results of this study can provide an effective procedure of identifying the significant rules for classifying human body type to satisfy the demands for industrial and commerce.

Key-Words: - Data mining; Classification rules; Human body types; Industrial and commercial demand

1 Introduction

Human body type classifications are very crucial issue, play an even important role for physiology, medical treatment, sports talent and garment manufacturing. Taking the sizing systems used in the garment manufacturing industry as an example, garment manufacturers have never developed standard sizing systems according classified body type, finally resulting in heavy stock burden. Human body types could be accurately classified, garment manufacturers can correctly predict numbers of items and ratio of sizes to be produced, resulting in accurate inventory control and production planning according standard sizing systems [1]. With the quick advances in the social and economic environment, people’s body dimensions and shape are changeable. Each nation requires realizing the people’s body types for industrial and commercial demand. Thus, the classification of human body type is long overdue.

Body type classifications have different goals. The earliest exploration body types were most focused on physiology, medical treatment and sports talent. Until 1950, body type data was only applicable to garment sizing systems [2].

In related physiological and medicinal studies, Chau et al. (1993) studies 300 healthy adults to discover factors that affect figure type, and expecially body weight [3]. Kalichman and Kobyliansky (2006) described the age- and sex-related variations of the figure type in a Chuvasha population residing in a rural region in central Russia [4]. Zalleg et al. was to study the relationship between the cardiac output and the figure type in elites handball players [5].

In related sport studies, Guo and Xu analysed the relation between figure type indexes and athletic level [6]. Chen and Chao analyzed the difference of explosive power and figure type between young female aboriginal and non-aboriginal students [7]. Liou analyzed 47 female weightlifters subjects in his study. Collected their age, playing year and figure type data to study the relationship between them and to offer the reference for coaches [8]. Bayios et al. in their study were: a) to determine the anthropometric profile, body composition and figure type of elite Greek female basketball, volleyball and handball players, b) to compare the mean scores among sports and c) to detect possible differences in relation to competition level [9]. Kawashima compared the physical characteristics and figure types of 4 Japanese male golfer groups with 2 non-golfer control groups in his investigation [10].

In the ancient Greek era, Hippocrates divided people into two body types: stout, referred to as Habitus Apoplecticus, and slender, called Habitus Phythesicus. Siguad proposed four body types, Respiratory, Digestive, Muscular and Cenebral. Kretchmer presented three body types, Athenic, Athletic and Pyknic. The above classification methods are all based on observable appearance. By contrast, Sheldon (1940) classified body type by measurement. Sheldon compare and analyze the results with the body type of people divided as Endomorphy, Mesomorphy and Ectomorphy by human body photography. Sheldon's classification is based only on feature description without concrete or qualified classification reference baselines [11]. Skerli et al. (1953) categorized people into eight body types by measuring from photography and the hypodermis: Norma (normal figure), Rubens (obese figure), Superior (obese above the waist and normal below the waist), Inferior (normal above the waist and obese below the waist), Truncic (obese trunk obesity with normal extremities), Exterminable (normal trunks with obese extremities), Mammary (figure with fat accumulated on the chest) and Trochanteric (with fat accumulated on the legs). Skerli emphasized partial variances. In 1984, Ronald studied the bodily density of various humane body types, classifying human body types into Obesity, Robust and Slender and subsequently measured the density of human body [12].

Most studies discovered relationships between figure types and other areas for different goals. However, little research has been done on the classification of figure types. On the other hand, data mining has been successfully applied in many fields. The application domain is quite broad and plausible in marketing [13], production [14], human resource management [15], risk prediction [16], biomedical technology [17] and health insurance [18]. However, research on identifying body type classifications using data mining is lacking. Accordingly, this study attempts to classify body types using the difference between waist girth and hip girth and the waist-to-hip ratio as obtained from anthropometric data by employing two stages-based data mining procedure to identify unknown rules for body type classifications. By applying two stages-based data mining procedure, body types can be classified from an anthropometric database.

2 Two stage-based data mining

Berry and Linoff defined data mining as the analysis of huge amounts of data by automatic or semi-automatic means, in order to identify significant patterns or rules [19]. Data mining has the following features:

(1).The database to be analyzed is large.

(2).The data mining process is created automatically or semi-automatically.

(3).The information excavation must be meaningful or usable.

(4).The formats of new information are composed of correlation and the modeling demonstration.

One of the most important data mining techniques is cluster analysis, which is an exploratory data analysis tool for solving classification problems. Its object is to sort cases into clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. The cluster analysis includes both hierarchical and non-hierarchical methods [20]. Ward's minimum variance is an important agglomerative hierarchical algorithm method, as the smallest increase in total within-group variance has the highest priority of combination. On the other hand, the most widely used method for non-hierarchical algorithms is the K-means method. The K-means method is known as a partitional method since the user must first determine the number of clusters after which the algorithm partitions the data iteratively until a solution is found [21]. This study has integrated Ward's minimum variance method with the K-means method to conduct the first stage data mining.

On the other hand, the decision tree, one of the important data mining techniques, uses rules created in accordance with input variables and is executed with data classification by a tree type demonstration to extract the most significant factors. Thus, decision tree is the most suitable method for classification.

Data types can be classified as discrete and continuous. Various types of data will come with appropriate decision tree algorithms. The algorithm most appropriate for continuous data processing is exactly the CART (Classifucation and Regression Tree) as the best representative [22]. Because all the anthropometric data are continuous, thus, CART was used to classify the anthropometric data for data mining.

The most important advantage of CART is that it improves the accuracy percentage, and best decreases the complexity of decision trees including the tree depth and number of nodes. The most critical part of establishing the CART algorithm is to identify the important attributes, particularly the predict variable. The target variable can be subdivided into child nodes based on the predict variable data, with each split aimed to reduce the impurity of the child node until all samples within each child node reach the congruence, dividing the divergent impure data into several very pure child nodes [22].

CART can generate a tree structure with the availability for data classification. The data are placed into the root after pre-processing. A tree structure is therefore reached through a series of classification operations and locations for best divided points. The relevant classification operations are according to the purity of the child node, dividing the data to create the child node with similarities and also create the classification rules. These rules can be used to obtain the most valuable information obtained by data mining for data classification [23].

This study has integrated cluster analysis with decision tree to conduct data mining. A two stage-based data mining procedure was proposed, in order to mine the patterns of anthropometric data for the significant rules of human body types.

3 The Data Mining Procedure

The data mining procedure involves a series of activities, from defining the goal to evaluating the results. The previous steps can be served as the baseline reference for the next step.

3.1 Defining the goal

Owing to outdated and incomplete the classification of human body type, an anthropometric database was created for Taiwanese adult males. The ages of these samples are from 40 to 60 years old. The anthropometric database based on 52 anthropometric variables measured in each of 495 males according to the definition of the ISO 8559 [24].The goal of this study was to explore and analyze a huge amount of data, by employing two stage-based data mining procedure, so as to identify significant rules within body dimensions. Based on these rules, the lower body types classification of Taiwanese adult males may be classified.

3.2 Data preparation

Before mining the data, the data had to be processed, with all missing data being separated out [25]. As a result, of the 502 samples of adult males, 8, which had missing data, were deleted; this left a total of 494 valid samples.

Not all of the 52 anthropometric variables were suitable for use in identifying significant rules within body dimensions; therefore, in coordination with the judgment of domain experts, this study identified 10 variables.

To use all of the 10 anthropometric variables, as a basis for identifying significant rules, would make things too complicated; therefore, this study attempts to identify body types using the difference between waist girth and hip girth and the waist-to-hip ratio as obtained from anthropometric data by using two stage-based data mining procedure.

3.3 Data mining by cluster analysis

Data mining was undertaken through data preparation, using the cluster analysis, which included both hierarchical and non-hierarchical clustering. Ward's minimum variance method was integrated with the K-means method, to mine the patterns of anthropometric data for identifying significant rules within body dimensions. Ward's minimum variance method was used to determine the initial clustering information for the K-means method, while the K-means method determined the final clusters.

In the first hierarchical clustering, this study analysed the difference between waist girth and hip girth and the waist-to-hip ratio to decide the cluster numbers, using Ward's minimum variance method. A tree diagram, shown in Figure 1, presents the results. As shown, a total of 494 males were grouped three or five obvious clusters; thus, three or five cluster numbers were chosen for the next stage of processing.

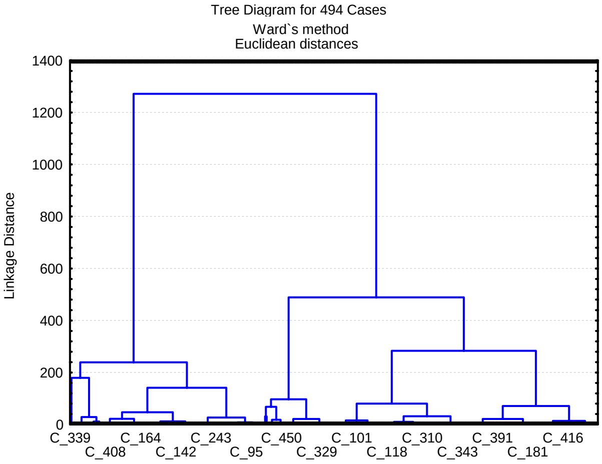
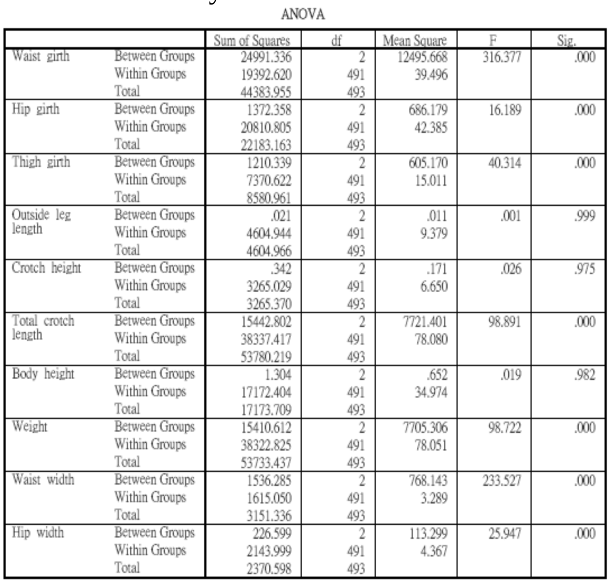


Fig. 1. The tree diagram of the difference between waist girth and hip girth and the waist-to-hip ratio by using Ward's minimum variance method

In the second non-hierarchical clustering, this study discovered that three and five clusters are appropriate for clustered result by the K-means method iteratively. To gain a better insight into the differences between the three and five clusters resulting from the cluster analysis, the Analysis of variance (ANOVA) was then conducted. In order to verify the anthropometric variables of all body types, and to determine whether notable differences existed among them. As the reaults that the three and five clusters bear significant differences in the girth anthropometric variables. The height anthropometric variables did not have significant differences. The results also uncovered the fact that differences did, indeed, exist between the girth anthropometric variables of the three and five clusters. The result is shown in Table 1 and Table 2.

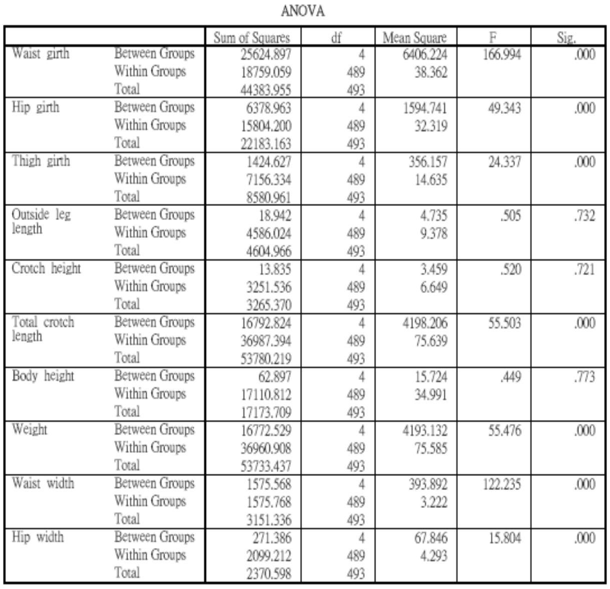
Table 1. The Analysis of variance for three clusters



ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Waist girth	Between Groups	24991.336	2	12495.668	316.377	.000
	Within Groups	19392.620	491	39.496		
	Total	44383.955	493			
Hip girth	Between Groups	1372.358	2	686.179	16.189	.000
	Within Groups	20810.805	491	42.385		
	Total	22183.163	493			
Thigh girth	Between Groups	1210.339	2	605.170	40.314	.000
	Within Groups	7370.622	491	15.011		
	Total	8580.961	493			
Outside leg length	Between Groups	.021	2	.011	.001	.999
	Within Groups	4604.944	491	9.379		
	Total	4604.966	493			
Crotch height	Between Groups	.342	2	.171	.026	.975
	Within Groups	3265.029	491	6.650		
	Total	3265.370	493			
Total crotch length	Between Groups	15442.802	2	7721.401	98.891	.000
	Within Groups	38337.417	491	78.080		
	Total	53780.219	493			
Body height	Between Groups	1.304	2	.652	.019	.982
	Within Groups	17172.404	491	34.974		
	Total	17173.709	493			
Weight	Between Groups	15410.612	2	7705.306	98.722	.000
	Within Groups	38322.825	491	78.051		
	Total	53733.437	493			
Waist width	Between Groups	1536.285	2	768.143	233.527	.000
	Within Groups	1615.050	491	3.289		
	Total	3151.336	493			
Hip width	Between Groups	226.599	2	113.299	25.947	.000
	Within Groups	2143.999	491	4.367		
	Total	2370.598	493			

Table 2. The Analysis of variance for five clusters



ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Waist girth	Between Groups	25624.897	4	6406.224	166.994	.000
	Within Groups	18759.059	489	38.362		
	Total	44383.955	493			
Hip girth	Between Groups	6378.963	4	1594.741	49.343	.000
	Within Groups	15804.200	489	32.319		
	Total	22183.163	493			
Thigh girth	Between Groups	1424.627	4	356.157	24.337	.000
	Within Groups	7156.334	489	14.635		
	Total	8580.961	493			
Outside leg length	Between Groups	18.942	4	4.735	.505	.732
	Within Groups	4586.024	489	9.378		
	Total	4604.966	493			
Crotch height	Between Groups	13.835	4	3.459	.520	.721
	Within Groups	3251.536	489	6.649		
	Total	3265.370	493			
Total crotch length	Between Groups	16792.824	4	4198.206	55.503	.000
	Within Groups	36987.394	489	75.639		
	Total	53780.219	493			
Body height	Between Groups	62.897	4	15.724	.449	.773
	Within Groups	17110.812	489	34.991		
	Total	17173.709	493			
Weight	Between Groups	16772.529	4	4193.132	55.476	.000
	Within Groups	36960.908	489	75.585		
	Total	53733.437	493			
Waist width	Between Groups	1575.568	4	393.892	122.235	.000
	Within Groups	1575.768	489	3.222		
	Total	3151.336	493			
Hip width	Between Groups	271.386	4	67.846	15.804	.000
	Within Groups	2099.212	489	4.293		
	Total	2370.598	493			

In analysis of variance, if the result is significant. we can use the Scheffe's test to analysis which specific mean differs from which other specific mean. In the study, the Scheffe's test was used to detect the means of anthropometric variables differences among body types. We conduct the Scheffe's test to see that there is a significant difference between the means of cluster 1, cluster 2 and cluster 3 in the girth anthropometric variables. There is no significant difference among the means of cluster 1, cluster 2 and cluster 3 in the height anthropometric variables. The result is shown in Table 3. The variables means of various body type was arranged according the value.

Table 3. The Scheffe's test

Anthropometric variables	Big value		Small value
Waist girth	cluster 3	cluster 1	cluster 2
Hip girth	cluster 3	cluster 1	cluster 2
Thigh girth	cluster 3	cluster 1	cluster 2
Outside leg length	cluster 1	cluster 3	cluster 2
Crotch height	cluster 1	cluster 3	cluster 2
Total crotch length	cluster 3	cluster 1	cluster 2
Body height	cluster 1	cluster 3	cluster 2
Weight	cluster 3	cluster 1	cluster 2
Waist width	cluster 3	cluster 1	cluster 2
Hip width	cluster 3	cluster 1	cluster 2

Therefore, this study defined the body type, formed by cluster 3, with large girth anthropometric variables, as type L; the body type, formed by cluster 1, with medium girth anthropometric variables, were defined as type M; and the body type, formed by cluster 2, with small girth anthropometric variables, were defined as type S. This definition of the three body types, used in this study, is shown in Table 4. No significant differences among the three body types in the height anthropometric variables.

Table 4. Definitions of three body types

Clusters	3	1	2
Numbers	115	254	125
Girth variables	Large	Medium	Small
Height variables	-	-	-
Body types	L	M	S

3.4 Data mining by decision tree

The decision tree was used to mine data based on the results of cluster analysis. This study takes the waist-to-hip ratio as the target variable, with waist girth and hip girth being predictors used to classify the target variable. The following stopping rules were set as follow.

- The greatest depth of the tree extends to the fourth level beneath the root node.
- The minimum number of samples in the parent node is 2, and the minimum number of samples in the child node is 1.

Taking the cluster 1 (body type M) as an example, Figure 2 presents the results of decision tree analysis. The root node was split according to the waist girth, resulting in the first level. A total of 100 samples whose waist girths were smaller than or equal to 88 cm were grouped into Node 1, and 154 samples with waist girths greater than 88 cm were grouped into Node 2. Eventually, only the two nodes generated at the first level were chosen to represent the classification rules. The classification rules of all body types shows Table 5.

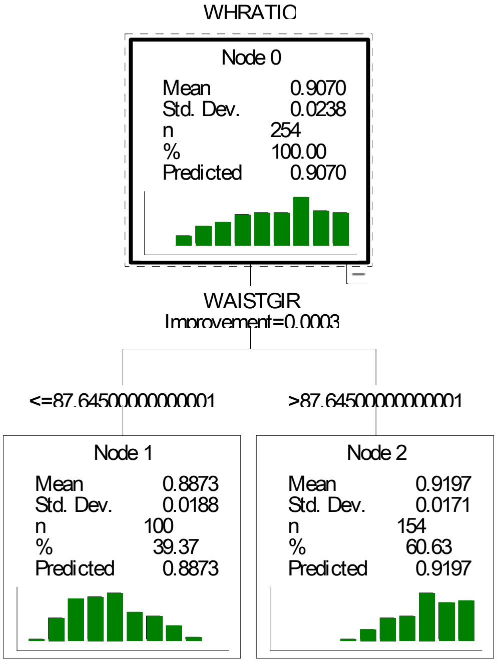


Fig. 2. The decision tree branching

Table 5. The classification rules of all body types

Clusters	Body types	Classified rules	Numbers
3	Lw1	105 < Waist girth	18
	Lw2	Waist girth≦ 105	97
1	Mw1	88 < Waist girth	144
	Mw2	Waist girth≦ 88	110
2	Sw1	80 < Waist girth	46
	Sw2	Waist girth≦ 80	79

3.5 Evaluation of Results

Having classified the males into three body types and six classification rules, the new classified body types were identified using two stage-based data mining procedure. Body types could be accurately classified for physiology, medical treatment, sports talent and garment manufacturing according the newly classification rules.

Therefore, this study defined the precise figure type, formed by cluster 3 with waist girths greater than 105 cm, as type Lw1; with waist girths were smaller than or equal to 105 cm, as type Lw2. The figure type, formed by cluster 1, with waist girths greater than 88 cm, as type Mw1; with waist girths were smaller than or equal to 88 cm, as type Mw2. The figure type, formed by cluster 2, with waist girths greater than 80 cm, as type Sw1; with waist girths were smaller than or equal to 80 cm, as type Sw2.

Taking the sizing systems used in the garment manufacturing industry as an example, garment manufacturers can correctly predict numbers of items and ratio of sizes to be produced, resulting in accurate inventory control and production planning according the newly classified body types. Furthermore, the sizing systems thus developed provide the percentage of males within each size group, and the distribution of body types, enabling manufacturers to access reference points and facilitating garment production for specific markets, and supplying effective manufacturing information, improving production planning and material control. The standard sizing systems can then be developed to facilitate garment production according to the new classified body types. The newly classification rules play an even important role for industrial and commerce.

4 Conclusion

Human body type classifications play an even important role for physiology, medical treatment, sports talent and garment manufacturing. The application domain of data mining has been quite broad. However, little research has been done in the area of identifying the significant rules of human body types, using data mining. This study applied novel two stage-based data mining procedure using the difference between waist girth and hip girth and the waist-to-hip ratio, to identify the significant rules of human body types. Certain advantages may be observed when the significant rules are identified, using two stage-based data mining procedure. Body types could be accurately classified for physiology, medical treatment, sports talent and garment manufacturing according the newly classification rules. The results of this study can provide an effective procedure of identifying the significant rules for classifying human body type to satisfy various demands.

Acknowledgments

The author acknowledges the partial support from the National Science Council of Taiwan, R.O.C., under project number NSC 95-2221-E-164-002.

References:

[1] Burns LD, Bryant NO, *The Business of Fashion: Designing, Marketing, and Manufacturing*, Fairchild, 2000.

[2] Hsu CH and Wang MJ, Using decision tree based data mining to establish a sizing system for the manufacture of garments., *International Journal of Advanced Manufacturing Technology*, 26, 2005, pp.669-674.

[3] Chau TT, Hsu SM and Lin CL, A preliminary report about figure shape and composition of 300 citizens in Kaohsiung city, *The Kaohsiung Journal of Medical Sciences*, 9, 1993, pp.296-304.

[4] Kalichman L and Kobyliansky E, Sex- and age-related variations of the figure type in a Chuvasha population, *Journal of Comparative Human Biology*, 57, 2006, pp.151-162.

[5] Zalleg D, Bouassida A and Jabrallah M, Figure types and cardiac output in elites handball players, *Science & Sports*, 20, 2005, pp.275-278.

[6] Guo HB and Xu SY, The figure type analysis on male adolescent dash players in Gansu Province, *Journal of Northwest Normal University*, 41, 2005, pp.89-92.

[7] Chen WY and Chao SY, The analysis of explosive power and figure type between young female aborigine and non-aborigine, *Journal of Exercise Physiology and Fitness*, 4, 2006, pp.93-105.

[8] Liou YC, An relationship analysis of R.O.C. female weightlifters' age, playing year and figure type, *Physical Education Journal*, 31, 2001, pp.1-12.

[9] Bayios IA, Bergeles NK and Apostolidis NG, Anthropometric, figure composition and figure type differences of Greek elite female basketball, volleyball and handball players, *Journal of sports medicine and physical fitness*, 46, 2006, pp.271-280.

[10] Kawashima K, Figure size and figure type characteristics of male golfers in Japan, *Journal of sports medicine and physical fitness*, 46, 20033 pp.334-341.

[11] Laubach LL and Marshall ME, A computer program for calculation parnell's anthropometric phenotype, *Journal of Sports and Medicine*, 1970, pp.217-214.

[12] Skerli B, Brozek J, Edward E and Hunt J R, Subcutaneous fat and age change in body build and body form in women, *American Journal of Physical Anthropology*, 1953, pp.577-600.

[13] Wong KW, Zhou S, Yang Q and Yeung MS, Mining customer value: from association rules to direct marketing, *Data Mining and Knowledge Discovery*, 11, 2005, pp.57-79.

[14] Sha DY and Liu CH, Using data mining for due date assignment in a dynamic job shop environment, *International Journal of Advanced Manufacturing Technology*, 25, 2005, pp.1164-1174.

[15] Min H and Emam A, Developing the profiles of truck drivers for their successful recruitment and retention: a data mining approach, *International Journal of Physical Distribution & Logistics Management*, 32, 2003, pp.149-162.

[16] Becerra-Fernandez I, Zanakis SH and alczak S, Knowledge discovery techniques for predicting country investment risk. *Computer and Industrial Engineering*, 43, 2002, pp.787-800.

[17] Maddour M and Elloumi M, A data mining approach based on machine learning techniques to classify biological sequences, *Knowledge-Based Systems*, 15, 2002, pp.217-223.

[18] Chas YM, Ho SH, Cho KW, Lee DH and Ji SH, Data mining approach to policy analysis in a health insurance domain, *International Journal of Medical Informatics*, 62, 2001, pp.103-111.

[19] Berry M, Linoff G, *Data Mining Techniques: for Marketing, Sales, and Customer Support*, Wiley, New York, 1997.

[20] Giudici P, *Applied Data Mining: Statistical Methods for Business and Industry*, Wiley, England, 2003.

[21] Kuo RJ, Ho LM and Hu CM, Cluster analysis in industrial market segmentation through artificial neural network. *Computer and Industrial Engineering*, 42, 2002, pp.391-399.

[22] Breiman L, Friedman J H and Olshen R A, *Classification and Regression Tree*, Chapman and Hall, Florida, 1998.

[23] Moisen G G and Frescino T S, Comparing five modeling techniques for predicting forest characteristics, *Ecological Modelling*, 157, 2002, pp.209-225.

[24] ISO, Garment Construction and Anthropometric Surveys – Body Dimensions. International standard, ISO 8559, 1989, pp.1-9.

[25] Pyle D, *Data Preparation for Data Mining*, Morgan Kaufmann, California, 1999.