K-NN Classifier Analysis for an Epidemic Study on Fatigue Syndrome of Juvenile Educatees

Shusaku Nomura
Nagaoka University of Technology, Nagaoka, Japan
nomura@kjs.nagaokaut.ac.jp

Santoso Handri
Nagaoka University of Technology, Nagaoka, Japan

C.M. Althaff Irfan
Nagaoka University of Technology, Nagaoka, Japan

Sanae Fukuda
Osaka City University, Osaka, Japan

Emi Yamano
Osaka City University, Osaka, Japan

Yasuyoshi Watanabe
Osaka City University, Osaka, Japan
and
Center for Molecular Imaging Science, RIKEN, Japan

Abstract
Two contrasting approaches toward an epidemic study were illustrated in this study; one is the regression analysis which is rather conventional methodology used in the past/present epidemic studies, and the other is the classifier analysis which is in the soft computing toolbox. The dataset analysed is a part of a cohort study which principally focused on a fatigue syndrome of the elementary and junior high school educatees. In the classifier analysis we employed a major supervised machine-learning algorithm, K-Nearest Neighbour (K-NN), coupled with Principal Component Analysis (PCA). As a result, the performance that was found by the classifier analysis provides rather better results than that of the regression analysis. Finally we discussed the availability of both analyses with referring to the technical and conceptual limitation of both approaches.
Keywords: K-NN, PCA, cohort study, fatigue, epidemic study.

1 Introduction

This study illustrates two contrasting approaches toward analyzing a multidimensional dataset of an epidemic study; one is the regression analysis which is rather conventional methodology and frequently used in the past/present epidemic studies [1-3], and the other is the classifier analysis which is in the soft computing toolbox. These two approaches are sometimes technically fused and used as an integrated approach. However, by virtue, these are standing on different concepts to each other. The regression analysis is not merely a method to reveal a linear relationship between dependent and independent variables but it can refer to some statistical features of the whole *population* from the limited number of samples, as it is standing on the concept of *statistics for inference*. Meanwhile, classifier which is constructed by the given dataset via supervised or unsupervised machine-learning is a method to classify all-new data for the system. It does not refer to the *population* but makes a decision purely by such a classifier constructed by the limited number of given data, thus it would rather be said that it is standing on the concept of *descriptive statistics*.

Because the eventual goal of the epidemic study is the prediction of the future states from what has happened in the past or what it is in the present, the regression analysis with inference fits for this purpose to some extent. However in contrast to its availability it has strict limitations, and which induced us to promote this study.

1.1 Technical Limitation of Regression Analysis, and Our Study

There are technical limitations (assumptions) on regression analysis; as such the dependent variables should have normal distribution for any independent variables, all independent variables should be linearly independent to each other (multicollinearlity), the set of sampled data should have exactly the same statistical features as its population, etc. These technical limitations (assumption) certainly yield a great advantage to regression analysis in terms of inference. One can refer to the statistical features of the hundreds million people (assumed *population*) solely by the thousands individuals' profile (*sample*).

However when one looks at our real society, it would be a bold assumption as which solely 0.1% of some groups represents the rest. Moreover, since these technical assumptions are practically "conceptual" assumptions as well, it is untouchable in principle. It is ironically unknown if one would belongs to the assumed *population* by which he/she estimates the risk factor of their behaviours. Especially when there is no statistically significant result by regression analysis, it is indistinctive whether it would be resulted in the absence of relevance among independent variables, e.g., the behaviours accounted by questionnaires, and a dependent variable, e.g., the morbidity for a particular disease, or be resulted in the inappropriate choice of the assumed population; which encompasses the inappropriate choice of the independent variables in the study.

The regression analysis is conventional and even sophisticated approach to interpret the huge dataset of an epidemic study. However it should be paying attention to the fact that it is standing on the strict limitations. At least there might be a space for introducing an alternative approach, which does not entail the statistics for inference, to interpret the epidemic data. In this study we introduced classifier analysis in the soft computing toolbox to deal with a multidimensional dataset of an epidemic study on which a fatigue syndrome of the elementary and junior high school educatees were focused. We discuss the difference between our approach and conventional regression analysis in terms of its limitations, performances, and interpretations and finally introduce our proprietary web-application developed on the idea of classifier analysis, by which each user can predict the level of motivation for learning on-site.

2 Method

2.1 Dataset

The dataset we used for this study is obtained from a part of our cohort study which principally focused on a fatigue syndrome of the elementary and junior high school educatees who were 9 to 15 in their ages as shown in Fig. 1. Every half a year for 2.5 years (five times in total), over 2000 educatees from four elementary and four junior high schools voluntary participated in this cohort study. They were asked to fill up a questionnaire which consists of over 200 items including 14 items (4-point scale) for the Chalder's Fatigue Scale [4] (Japanese version was provided by Demura, 2001. [5]), which is the target of this cohort study, and 27 items for their lifestyle as shown in Table I, which were assumed as contributing factors to the fatigue syndrome. Chalder's Fatigue score was found by summation of the pointed scale in each item. In this study, for the purpose of simplification, we used the dataset of 202 subjects who completed all items required in the questionnaire during 2.5 follow-up surveys.

This study was endorsed by the ethics committee of the Osaka City University.

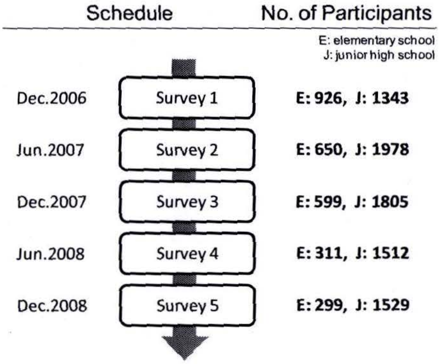

Figure 1: Schema of our cohort study

2.2 Regression Analysis

With regard to a conventional analysis we introduced multiple linear regression analysis with the least square method; for which 27 items for lifestyle at the beginning of the survey were assumed as independent variables and Chalder's Fatigue score at 2.5 years later was assumed as the dependent variable. Since Chalder's Fatigue score is quantitative variables, we introduced linear regression analysis instead of logistic regression analysis; it has been frequently used in the analysis of epidemic data since the dependent variables of epidemic studies are usually in the form of categorical data such as dead/alive. The step-wise selection of the dependent variables was also introduced for the purpose of optimization.

Table I: Items and FSS ranks.

ID	Item	*FSS rank* [a]
1	Hours of sleep (weekday)	15
2	Hours of sleep (weekend)	23
3	Regularity of breakfast	10
4	Appetite for breakfast	7
5	Regularity of taking a dinner with family	11
6	Meal time	6
7	Likes and dislikes in food	4
8	School attendance	24
9	Tardiness for school	14
10	Enjoyment in school	5
11	Getting along with classmates	16
12	Falling behind in class	26
13	Motivation for study	12
14	Enjoyment in studying	8
15	After-school activities (weekday)	27
16	After-school activities (weekend)	22
17	Frequency of exercise	1
18	Hours of using computer	17
19	Hours of watching TV	3
20	Frequency of video game	21
21	Frequency of going shopping	18
22	Number of family member	2
23	Hours of spending time with family	25
24	Being praised by family for studying hard	9
25	Having chronic disease	19
26	Absent from school for no special reason	13
27	Absent from school by physical problem	20

a. FSS ranks: the rank ordered by Feature Subset Selection (described in Result section).

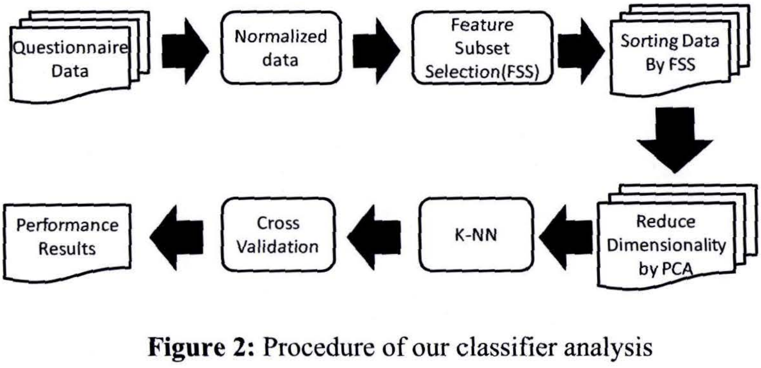

Figure 2: Procedure of our classifier analysis

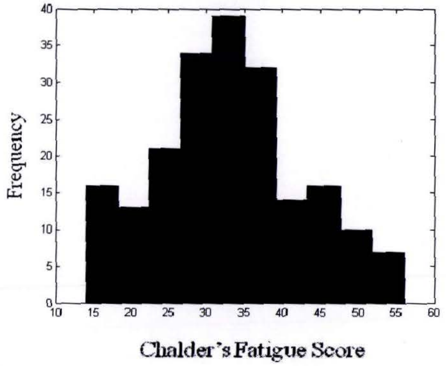

Figure 3: Distribution of Chalder's Fatigue score

2.3 Classifier Analysis

Fig. 2 shows the procedure of classifier analysis in this study. Normalization (Z-transform) and Feature Subset Selection (FSS) was employed as a pre-treatment. FSS is a practical pre-treatment processing for selecting useful features. Selection is made by the significance of each feature and it is ranked based on the efficacy of individually conduced binary classification.

The significance of each feature is found by

$$F(x_j) = \left| \frac{(\mu_{j,+1} - \mu_{j,-1})}{\sqrt{\left(\dfrac{\sigma_{j,+1} - \sigma_{j,-1}}{n_{j,+1} - n_{j,-1}} \right)}} \right| \qquad (1)$$

where n_{+1} and n_{-1} is the number of examples belonging to the effective and ineffective group, $\mu_{j,+1}$ and $\mu_{j,-1}$ is the mean of the j-th feature of the effective and ineffective group, and $\sigma_{j,+1}$ and $\sigma_{j,-1}$ is that of standard deviation, respectively. This criterion is interpreted as finding the one single feature that best discriminates among both groups in feature space. The greater this score, the better the feature's discrimination. Based on this score, each feature is assigned by rank of significance. Features are selected using a certain number of features from the top.

After FSS, 90% of the given data was randomly assigned into the training dataset in the next steps and the rest of 10% was kept for the validation (cross validation). In the next the Principal Component Analysis (PCA) was made so as to reduce the dimension of feature space for making a better performance in the subsequent K-Nearest Neighbor (K-NN) analysis. The K-NN is a supervised learning algorithm. Although it is the simplest of all machine learning algorithms, it has high performance and low computational cost. An object is classified by a majority vote of its k nearest neighbors.

For the purpose of simplification, we applied K-NN in binary classification problem with employing the Euclidean distance in order to identify neighbors. With regard to the annotation of the training data into binary class, i.e. high or low fatigue, by referring to the distribution of Chalder's Fatigue score (Fig. 3), the subjects who had 35 point or higher in the score were annotated in the high fatigue group and others were in low fatigue group. Annotated data for each subject, either high or low fatigue, was used in the supervised learning processing of K-NN. Finally the cross validation was performed by 10% of the remained test data. All the steps subsequent to FSS were iterated by changing the test dataset, the number of PCA components (from 2 to 27), and the number of k (from 2 to 15).

3 Result and Discussion

3.1 Regression Analysis

As a result of multiple regression analysis with step-wise method, the statistically significant relationship ($\acute{R}^2$=0.098, $F_{7,194}$=4.135; p<0.001) between a dependent variable (Chalder's Fatigue score) and reduced independent variables (lifestyle) was found as shown in Table II. However when one takes account for adjusted coefficient of determination ($\acute{R}^2$), it is unreasonable to claim that there is relationship between fatigue and lifestyle even though some statistical significance was found in the analysis. In fact, it should be noted that the statistical test normally used in the multiple regression analysis based on rather bold *null-hypothesis*, i.e. H_0: *NOT ALL* partial regression coefficient is zero; in other words, H_0 is accepted only if *ALL* the independent variables has no relevance at all with the dependent variable. Moreover the greater the sample size, the greater the chance of significance in this hypothesis test. With regard to the inference such poor relation is illustrated more clearly as shown in Fig.4 in which 95% conference interval of the regression line for an item, "Hours of using computer", is depicted. Although the regression coefficient of the item is statistically "*not zero*" (Table II), it could not explain anything in terms of prediction of the fatigue.

Therefore it could be impartially considered that no distinctive relevance was obtained by the result of multiple linear regression analysis in this study.

Table II: Result of regression analysis

ID	Item	*SPRC* [a]	*PCC* [b]	*t*-test [c]
6	Meal time	0.14	0.14	p<.05
7	Likes and dislikes in food	-0.17	-0.18	p<.05
11	Getting along with classmates	0.15	0.15	p<.05
14	Enjoyment in studying	0.12	0.13	n.s.
18	Hours of using computer	0.14	0.14	p<.05
19	Hours of watching TV	-0.11	-0.11	n.s.
23	Hours of spending time with family	-0.12	-0.12	n.s.

a. SPRC: standardized partial regression coefficient, b. PCC: partial correlation coefficient, c. *t*-test for SPRC for each item. "n.s." means not significant.

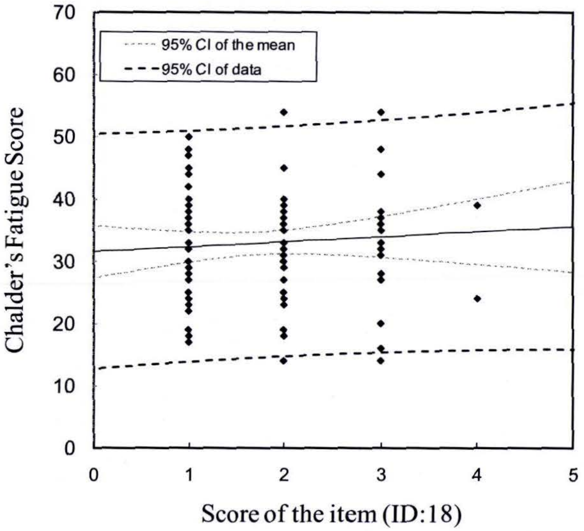

Figure 4: Linear regression between Chalder's Fatigue score and a item "Hours of using computer (ID:18)". Black solid line represents regression line. Black and grey dashed-line represents 95% confidence interval of data and the mean, respectively.

Table III: Result of classifier analysis

No. of Features [a]	Identification [b]	Evaluation [c]
3	61.1%	57.9%
5	68.1%	61.4%
10	64.3%	53.0%
15	61.9%	54.0%
20	65.2%	56.9%
25	61.9%	51.5%

a. Number of features introduced into classifier analysis, b. accuracy rate obtained by training dataset, c. accuracy rate obtained by test dataset

3.2 Classifier Analysis

Table I shows the rank of all the 27 items ordered by FSS. Actually FSS is one of the linear classifier; nevertheless it is an unsupervised classifier. FSS gives items which explain the target better/poor in terms of linear relationship. Therefore the items in the higher rank in the result of FSS are expected to give a better performance in the regression analysis as well.

The classifier analysis was conducted following the FSS. Table III shows the result of the classifier analysis. The "identification" and "evaluation" in the table represents the accuracy rate obtained by the training dataset (90%) and the test dataset (10%), respectively. The accuracy rate is the average of 10 times of the cross validation. It should be noted that the number of K-neighbour (*k*) and PCA components (*pc*) used in the process of K-NN supervised learning were evaluated in the iteration and finally optimized as *k*=15 and *pc*=3, respectively. It is suggested that a few PCA dimensions were enough to train the K-NN classifier. In fact, the only two PCA components, first and second components, well explain the difference in high/low fatigue whether in the identification and/or evaluation as shown in Fig. 5.

With regard to the performance of our analysis, the number of features used in the analysis takes certain part in the performances; however the accuracy rate, 68.1% of the highest, still remains small. Taking the result and discussion of regression analysis into account, however, it is not necessary to think this classifier analysis is poor in performance. As a matter of fact, the items on lifestyle introduced in our epidemic study have revealed to have little relevance with fatigue by its virtue as shown in Fig.4. Therefore the accuracy rate of around 60% obtained by classifier analysis gives us a reason to introduce such a classifier analysis in the soft computing tool box into conventional analytical methodology in the epidemic study, especially in case of that no linear relationship is expected by regression analysis.

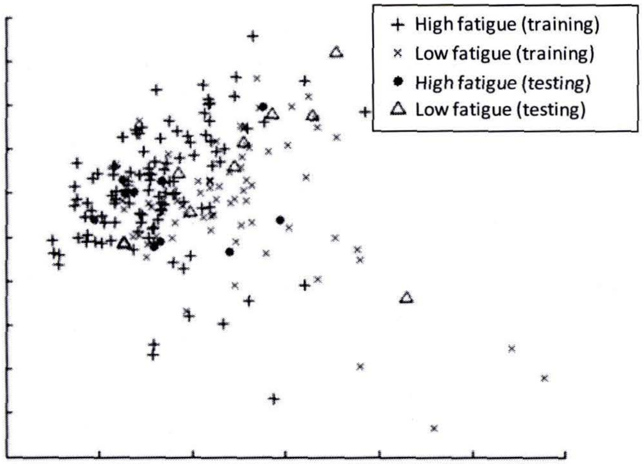

Figure 5: Result of classifier analysis presented in the first and second PCA component.

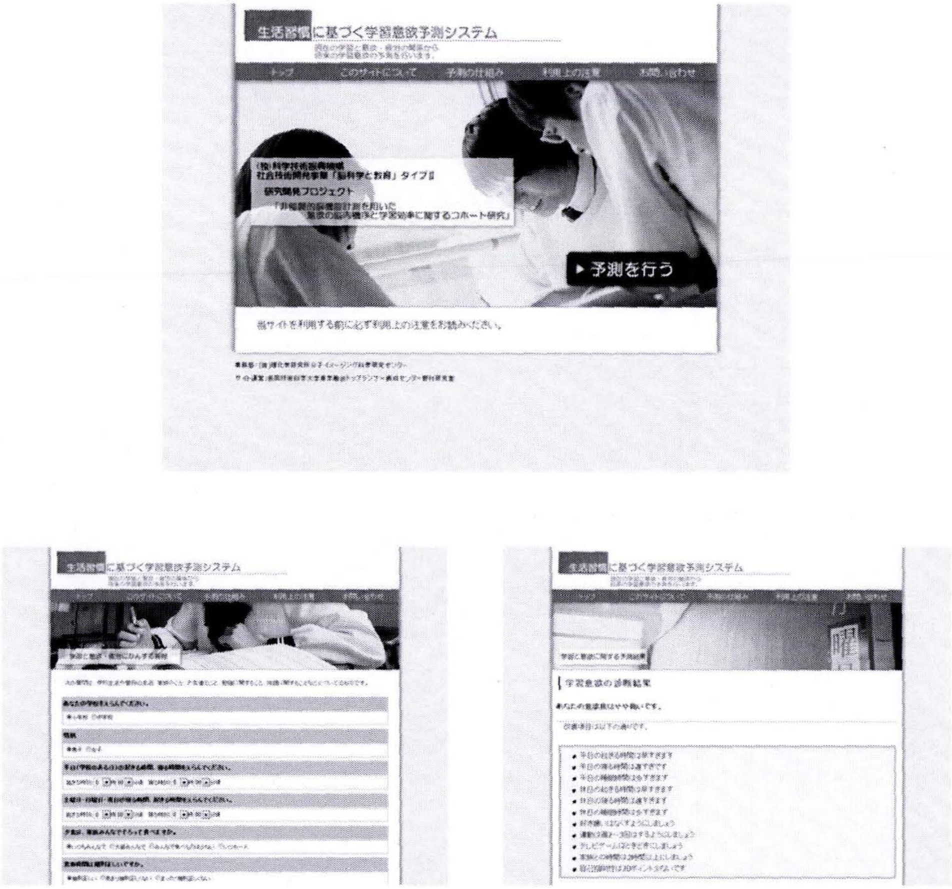
生活習慣に基づく学習意欲予測システム
トップ
このサイトについて
予測の仕組み
利用上の注意
お問い合わせ
(独)科学技術振興機構
社会技術研究開発事業「脳科学と教育」タイプII
研究開発プロジェクト
予測を行う
本サイトを利用する前に必ず利用上の注意をお読みください。
学習意欲の診断結果

Figure 6: Hard image of our web-application for the prediction of motivation for learning. By filling in all required items concerning lifestyle (lower left), the expected result and indication of unfavourable items for improving the result were presented (lower right).

3.3 General Discussion

We do not claim that the classifier analysis is better methodology than conventional regression analysis. In the stream of epidemic study the idea of *odds rate* has been frequently introduced as a reference of confidence of the relationship between target and assumed factors (e.g., [2]). The odds rate was obtained by logistic regression analysis. It thus gives a strong suggestion on the statistical features of the target population.

However when one goes back to the conceptual and thus untouchable limitation of the regression analysis which entails the idea of statistics for inference, one should be paying attention not to overestimate the results.

On the other hand classifier analysis is rather exploratory data analysis; one cannot refer to the *population*. The analysis is made purely by the given dataset. It entails more like descriptive statistics, and thus the interpretation of the analysis is made rather individualistically, not by the statistical feature of the assumed *population*. However if a particular individual could assume that *she/he is belonging to the similar group* for which the training procedure of the classifier analysis is conducted, the result of classifier analysis gives her/him a practical indication for what would happen in their future and/or how they could deal with the expected future by changing their lifestyle. The performance of our classifier analysis still remains poor (68.1% of the highest). However it is better to abandon an interpretation solely by the result of regression analysis.

The difference of above-mentioned two methodologies is the difference of conceptual assumption standing on the different idea of statistics approach. However there is a case in which the assumption of "*belonging to the similar group*" for classification is rather simple and practical than that for statistics for inference. Especially, as we demonstrated in this study, if one has failed to find clear (linear) relationship among variables, one might want to search for another methodology standing on another assumption rather than regarding the result of regression analysis as nothing.

We then developed a web-application on the idea of classifier analysis presented in this study (Fig.6), by which each user can predict the level of motivation for learning on-site [6]. It was constructed using the huge dataset obtained by our cohort study of over 2000 educatees in elementary and junior high school. So if and only if a user were an educatee in the same age which means *belonging to the similar group*, it could gives a possible future state to a user and also an indication to improve unexpected results.

4 Conclusion and Future works

Two contrasting approaches, conventional regression analysis and classifier analysis, toward an epidemic study were illustrated in this study. Comparing with the results of both analyses, it implies that there must be a reason for introducing such classifier analysis for its performance and simple assumption behind. Variety of classifier algorithms in the soft computing toolbox other than K-NN could be employed and compared in performance and stability. Such an analysis might be branded less importance as in the current epidemiology. However it would rather preferable for users to obtain a practical suggestion for checking, evaluating, and changing their behaviour.

Acknowledgements

This study was supported in part by the 21st Century COE Program "Base to Overcome Fatigue" from the Ministry of Education, Culture, Sports, Science and Technology, and the Japan Science and Technology Corporation (JST)/Research

Institute of Science and Technology for Society (RISTEX), the Japanese Government. We would like to express our profound gratitude to all participants for this study. An author would like to show special thanks to Ms. Konami, N., Ms. Kosuzu, N., Ms. Kofuki, N., and Ms. Yukiko, N. for their moral support.

References

[1] M. Tanaka, K. Mizuno, S. Fukuda, S. Tajima, and Y. Watanabe, "Personality traits associated with intrinsic academic motivation in medical students," Medical Education, vol.43, pp.384-387, 2009.

[2] M. Tanaka, K. Mizuno, S. Fukuda, Y. Shigihara, and Y. Watanabe, "Relationship between dietary habits and the prevalence of fatigue in medical students," Nutrition, vol.24, pp.985-989, 2008.

[3] A.M. Meijer, H.T. Habekothé, and G.L.H. Van Den Wittenboer, "Time in bed, quality of sleep and school functioning of children," J. Sleep Res., vol.9, pp.145-153, 2000.

[4] T. Chalder, G. Berelowitz, T. Pawlikowska, L. Watts, S. Wessely, D. Wright, and E.P. Wallace, "Development of a fatigue scale," J. Psychosom. Res., vol.37, pp.147-153, 1993.

[5] S. Demura, H. Kobayashi, S. Sato, and Y. Nagasawa, "Examination of validity of the subjective fatigue scale for young adults," Nippon Koshu Eisei Zasshi, vol.48, pp.76-84, 2001. [Japanese]

[6] Prediction of motivation for learning along with lifestyle of educatees, 2010 [Japanese]. Available at: http://motivation.nagaokaut.ac.jp/