epidemiology, logistic regression analysis,
feature set selection, K-NN analysis

Santoso HANDRI[*,**], Shusaku NOMURA[*],
C.M. Althaff IRFAN[***], Sanae FUKUDA[****],
Emi YAMANO[****,*****], Yasuyoshi WATANABE [****,******]

AN ANALYSIS OF A MULTIDIMENSIONAL DATASET OF AN EPIDEMIC STUDY USING SOFT COMPUTING TOOLS －A PILOT STUDY

Two contrasting approaches toward an epidemic study were illustrated as a pilot study; 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 we used for this study is obtained from a part of a cohort study which principally focused on a fatigue syndrome of the elementary and junior high school educates. 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 cross validation method in the classifier analysis provides better results than that of the regression analysis. Finally we discussed the availability of both analyses with referring the technical and conceptual limitation of both approaches.

1. INTRODUCTION

This study illustrates two contrastingly 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. On the other hand, 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 that is 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, like as any other methodologies, it has strict limitations, and which induced us to promote this pilot study.

1.1. TECHNICAL AND CONCEPTUAL LIMITATION OF REGRESSION ANALYSIS, AND OUR STUDY

There are two kinds of limitations (assumptions) on regression analysis, technically and conceptually: as for a technical limitation, the dependent variables should have normal distribution for any independent variables, all independent variables should be linearly independent to each other (multicollinearlity), etc., and as for a conceptual limitation the set of sampled data should have exactly the same statistical features as its population. The conceptual limitation (assumption) certainly gives a greatest advantage to the regression analysis in terms of inference. One can refer to the statistical features of the 100 million people (assumed population) solely by the 2000 individuals (sample). However when one looks at our real society, it might be rather bold assumption as which solely 0.2% of some groups represents the rest. Moreover, because it is a conceptual assumption, it is technically untouchable. The reliability of such a conceptual assumption would depend on following three factors: 1) target of the epidemic study, i.e., death rate for particular disease, 2) method of sampling, i.e., random sampling with regard to the target, and 3) selection of independent variables, i.e., life style, behaviour, personality, etc. At least there might be a space for introducing another approach to the analysis of epidemic study which does not entail the statistics for inference.

We then 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 was focused.

[*] Nagaoka University of Technology,Top Runner Incubation Center for Academia-Industry Fusion, 1603-1 Kamitomioka, Nagaoka, Niigata, 940-2188 Japan
[**] Multimedia Nusantara University, Tangerang, Indonesia
[***] Nagaoka University of Technology, Graduate school of Management and Information Systems Engineering
[****] Osaka City University Graduate School of Medicine, Japan
[*****] Japan Science and Technology Agency (JST)/Research Institute of Science and Technology for Society (RISTEX), Japan
[******] Center for Molecular Imaging Science, RIKEN, Japan

2. METHOD

2.1. DATASET

The dataset we used for this pilot study is obtained from a part of our cohort study which principally focused on a fatigue syndrome of the elementary and junior high school educates who were 9 to 15 in their ages. Over 2000 educates from four elementary and four junior high schools voluntary participated in this cohort study. They were asked to fill up a questionnaire consist of over 200 items including 14 items (4-point scale) for the Chalder's Fatigue Scale (CF) [4] (Japanese version was provided by Demura, 2001. [5]) which was one of the targets of this cohort study and 27 items for their life style, school life, family relationships, and diseases 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, referring to the distribution of CF scores, the subjects who had 35 point or higher in CF score were annotated in the high fatigue group and others were in low fatigue group. This study was endorsed by the ethics committee of the Osaka City University. Based on the accuracy of prediction of high and low fatigue subjects with the cross validation method (see following section for more detail), we compare the performance of the regression analysis and our classifier analysis.

2.2. CLASSIFIER ANALYSIS

Figure 1 shows the procedure of classifier analysis which was employed in this study. Normalization (Z-transform) and Feature Subset Selection (FSS) was employed as a pretreatment. Feature subset selection (FSS) is the preprocessing part of the model that selects useful features for classification. Selection is based on the individual advantages of each feature. The t-test criterion measures individual feature significance based on ranking features x using an independent evaluation criterion for binary classification. The D-dimensional input vector is denoted as, where the number of examples belonging to the effective group is n+1, the examples belonging to the ineffective group n−1, the mean of the j-th feature of the effective group μ_j,+1, the mean of the j-th feature of ineffective group μ_j,−1, and their standard deviations $\sigma_{j,+1}$ and $\sigma_{j,-1}$. The significance of each feature x_j is measured as follows:

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

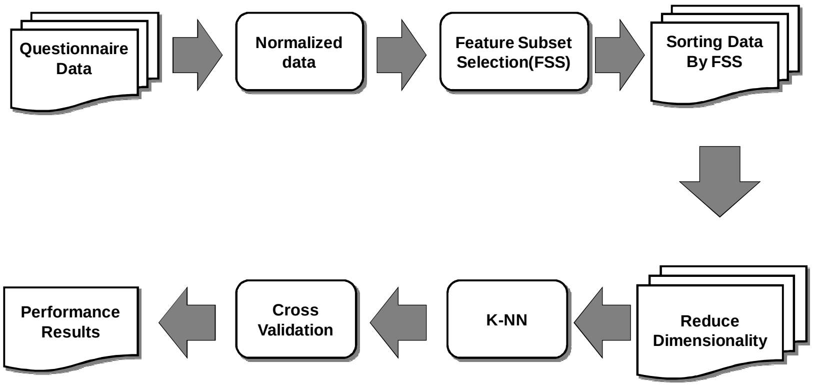


Fig. 1. The procedure of the classifier analysis

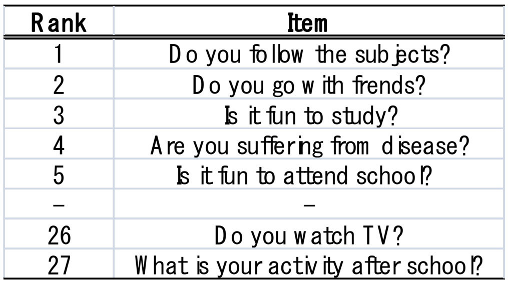
This criterion is interpreted as finding the one single feature that best discriminates among both groups in feature space. The greater this score, the better the feature's discrimination. Based on this score, individual features are assigned by rank of significance. Features are selected using a certain number of features from the top.

FSS gave an order among 27 items according to the importance and independency of the each item. This order was referred in the regression analysis which was performed for making a comparison with our classifier analysis. After FSS, 90 % of the given data was randomly selected for following steps and 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 Neighbour (K-NN) analysis. The K-NN is a supervised learning algorithm and is amongst the simplest of all machine learning algorithms, but has high performance and low computational cost. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours. k is a positive integer, typically small. If k=1, then the object is simply assigned to the class of its nearest neighbour. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes.

The neighbours are taken from a set of objects for which the correct classification is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbours, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidean distance, though other distance measures, such as the Manhattan distance could in principle be used instead.

Annotated data, either high or low fatigue, for each subject was used in the supervised learning process 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 in the 10% of test dataset (10 times), the number of PCA components (from 2 to 27), and the number of k (from 2 to 15).

Table 1. The top 5 and the worst 2 items ordered by FSS



Rank	Item
1	Do you follow the subjects?
2	Do you go with friends?
3	Is it fun to study?
4	Are you suffering from disease?
5	Is it fun to attend school?
–	–
26	Do you watch TV?
27	What is your activity after school?

With regard to the regression analysis, for the purpose of better comparison with classifier analysis, the single dimensional logistic regression analysis was employed in which each one of the items was an independent variable and the high/low fatigue score was dependent variable.

3. RESULT AND DISCUSSION

Table 1 shows the top 5 and the worst 2 items ordered by FSS. As a matter of fact, 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. Figure 2 shows the result of the logistic regression analysis for each top 4 items in FSS. The “evaluation” and “identification” in this figure represents the accuracy rate obtained by the 10% of test dataset and 90% of the data that is used for the analysis, respectively. The accuracy rate is the average of 10 times of the cross validation (error bar represents the standard deviation). There were almost no difference in the accuracy rate between the evaluation and identification. Also its deviations were small. The regression analysis achieved a certain and stable performance in this regard.

The result of classifier analysis, the average accuracy rate and its deviation are shown in Figure 3. The accuracy rate both in the evaluation and identification were higher than those of the regression analyses, while the difference in the accuracy between the evaluation and identification and its deviations were larger. Note 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 actual fact, as shown in Figure 4, the only two PCA components, first and second components, well explains the difference in high/low fatigue whether in the identification and/or evaluation. Moreover it should be note that the classifier analysis does not focus on a particular single item as it is in the regression analysis but encompasses all the items. In other words, the classifier analysis can refer to the balance or combination of the multiple items unlike as the regression analysis.

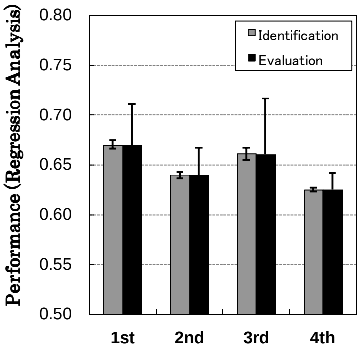


Fig. 2. The performance of the regression analysis for each item

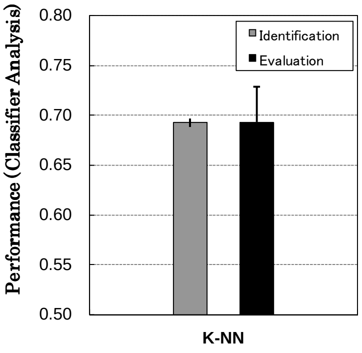


Fig. 3. The performance of the classifier analysis for each item

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 the confidence of the relationship between the target and assumed factor (e.g., [2]). The odds rate was obtained by logistic regression analysis. It thus gives a strong suggestion on the statistical futures of the target population. Moreover it is technically possible to introduce multi dimensional logistic regression analysis even though it requires strong assumptions as mentioned earlier. However when one goes back to the conceptual and untouchable limitation of the regression analysis which entails the idea of statistics for inference, the classifier analysis which shows rather better performance in this study can be considered. Regression analysis gives no result when there were no linear relationship between the target and assumed factors. Above all, because the target (endpoint) of our epidemic study is "fatigue" neither death nor disease, therefore the reliability of such a conceptual assumption would be impaired; no objective observation could be made on the "fatigue" unlike other physical conditions. The threshold of high/low fatigue score is not necessary to consider as a fixed constant but a parameter. Therefore exploratory data analysis including classifier analysis made purely by the given dataset could be practical even though such a method entails more like descriptive statistics; hence one cannot refer to the population.

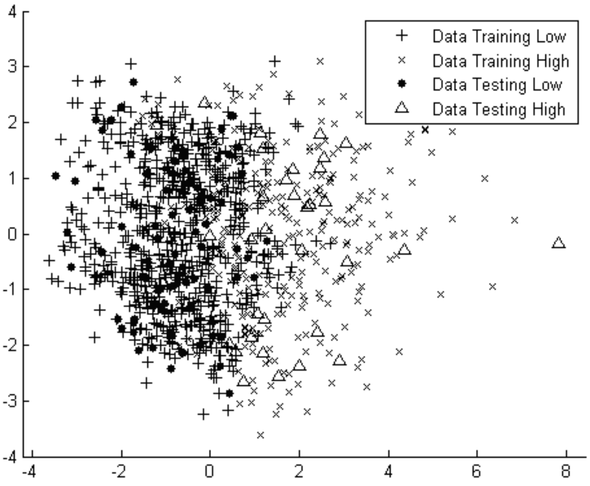


Fig. 4. The result of Principal Component analysis (PCA): vertical and horizontal axis represents the first and the second component of PCA. Data Training and Testing is the dataset used for identification and evaluation, respectively.

4. CONCLUSION AND FUTURE WORKS

Two contrasting approaches toward an epidemic study were illustrated as a pilot study. The classifier analysis introduced in this study shows better result than conventional logistic regression analysis. 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 epidemiology. However it would rather practical when there were no linear relationship and give strong suggestions to decide next experimental design.

ACKNOWLEDGEMENT

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.

BIBLIOGRAPHY

[1] TANAKA M., MIZUNO K., FUKUDA S., TAJIMA S., WATANABE Y., Personality traits associated with intrinsic academic motivation in medical students, Medical Education, Vol. 43, pp.384-387, 2009.

[2] TANAKA M., MIZUNO K., FUKUDA S., SHIGIHARA Y., WATANABE Y., Relationship between dietary habits and the prevalence of fatigue in medical students, Nutrition, Vol. 24, pp.985-989, 2008.

[3] MEIJER A.M., HABEKOTHÉ H.T., VAN DEN WITTENBOER G.L.H., Time in bed, quality of sleep and school functioning of children, J. Sleep Res., Vol. 9, pp.145-153, 2000.

[4] CHALDER T., BERELOWITZ G., PAWLIKOWSKA T., WATTS L., WESSELY S., WRIGHT D., WALLACE E.P., Development of a fatigue scale. J. Psychosom. Res., Vol.37, pp.147-153, 1993.

[5] DEMURA S., KOBAYASHI H., SATO S., NAGASAWA Y., Examination of validity of the subjective fatigue scale for young adults. Nippon Koshu Eisei Zasshi, Vol.48, pp.76-84, 2001. [Japanese]