RESEARCH ARTICLE
URL of this article: http://heanoti.com/index.php/hn/article/view/hn1202

Multivariate Adaptive Regression Spline Approach for the Classification Accuracy of Drugs User in East Java

Stefanny Zulistya Wenno*, Kuntoro*, Soenarnatalina*
*Faculty of Public Health, Airlangga University, Indonesia.
Email: stef.wenno@gmail.com

ABSTRACT

Background: Classification method is a statistical method for grouping or classifying the systematically arranged data into a group so we can know that an individual are in a particular group. Multivariate Adaptive Regression Spline (MARS) introduced by (Friedman, 1991) is a methodology for approximating functions of many input variables given the value of the function at a collection of points in the input space. Although training times for this method tend to be much faster than feed forward neural networks using back propagation, it can still be fairly slow for large problems that require complex approximations (many units). **Methods:** This was a nonreactive study, which is a measurement which individuals surveyed did not realize that they are part of a study. **Result:** Based on the best model selection criteria MARS then the selected is with model BF 20, MI 1 and MO 0 with the form Y = 0.929944 + 0.912438 * BF1 - 0.218729 * BF2 + 0.886429 * BF3 + 0.215575 * BF4 + 0.0745423 * BF5 - 0.232014 * BF6 + 0.0472966 * BF7 - 0.0367996 * BF8 + 0.0188678 * BF9 + 0.0304537 * BF11. Accuracy of drugs user rehabilitation classification that non relapse and relapse status based on MARS model is calculated using precision classification value. The accuracy level of drugs user rehabilitation classification in East Java using MARS method produces accuracy of 95,71% and misclassification of 4,29%. The magnitude of the above classification accuracy is due to the large prediction in the nonrelapse class that as many as 269 people with nonrelapse status are appropriately predicted in the nonrelapse status class. **Conclusion:** There are four important variables included in the best MARS model that is age of first use of drugs, how to use drugs, marital status and jobs.
The accuracy level of drugs user rehabilitation classification in East Java using MARS method produces accuracy of 95,71% and misclassification of 4,29%.

Keywords: Multivariate adaptive regression spline, Classification accuracy, Drugs user.

INTRODUCTION

Classification method is one of the statistical methods to classify or classify a systematically organized data into a group so that it can be known an individual is in a certain group. Classification problems are often encountered in everyday life. This classification problem arises when there are a number of sizes consisting of one or more categories that cannot be directly defined but must use a size.

Multivariate Adaptive Regression Spline (MARS) introduced by (Friedman, 1991) is a methodology for approximating functions of many input variables given the value of the function at a collection of points in the input space. Although training times for this method tend to be much faster than feed forward neural networks using back propagation, it can still be fairly slow for large problems that require complex approximations (many units).

Propagation, it can still be fairly slow for large problems that require complex approximations (many units). MARS are specialization of a general multivariate regression algorithm that builds hierarchical models using a set of basis functions and stepwise selection.

Multivariate Adaptive Regression Spline (MARS) approach to multivariate non parametric regression. The goal of this procedure is to overcome some of the limitations associated with existing methodology outlined earlier. It is most easily understood through its connections with recursive partitioning regression. It will therefore be developed here as a series of generalizations to that procedure.

The MARS procedure like recursive partitioning makes heavy use of the response values to construct a basis function set. This is how it achieves its response values to construct a basis function set. This is how it achieves its power and flexibility. This usually reduces the bias of model estimates, but at the same time increase the variance since additional parameters of the basis functions are being adjusted to help better fit the data at hand. The reduction in bias is directly reflected in reduced expected average squared residual.

The MARS procedure as well as recursive partitioning can be viewed as a technique for developing in a multivariate model, based on sums of products of univariate functions, through the use of a univariate smoother.

It is estimated that the number of drugs abusers is 3.8 million to 4.1 million people or about 2.10% to 2.25% of the total population of Indonesia at risk of exposure to drugs in 2014. Compared to the 2011 study, the prevalence rate is relatively stable (2.2%) but an increase compared to the 2008 study (1.9%). The contribution of the largest number of abusers comes from the group of workers, because it has the ability to finance and work pressure is great so that the level of high stress. Used wearers have the largest proportion of student groups.

METHODS

The type of research was nonreactive research which is a measurement where the individual studied is not aware that they are part of a study. Another name for nonreactive measurements is an unobtrusive measurement that emphasizes how the individual studied is not aware of the research because the measurements do not interfere with individuals and individuals not to be disturbed.

RESULTS

To get modeling with MARS approach is as follow: (1) describe response and predictor variable in modeling, (2) get the best MARS model with trial and error with following stages which determining maximum base function and determining maximum interaction, (3) determining minimum observation of each knots, (4) getting the best model with minimum GCV value, (5) determining variable into the best model based step 3 and (6) determining classification accuracy.

Processing with trial and error is done by combining the number of BF, MI and MO. The amount of BF used in this treatment is 2 to 4 times the predictor variable. MI used is 1, 2 or 3 with assumption if more than 3 will produce a very complex model. MO between knots used is 0, 1, 2, or 3. The modeling stage is performed by combining the specified BF, MI and MO values. The best model selection is seen from the smallest GCV value, but if GCV is equals then it seen at model which has the greatest classification accuracy.

Based on the best model selection criteria MARS then the selected is with model BF 20, MI 1 and MO 0 with the form:

Y = 0.929944 + 0.912438 * BF1 - 0.218729 * BF2 + 0.886429 * BF3 + 0.215575 * BF4 + 0.0745423 * BF5 - 0.232014 * BF6 + 0.0472966 * BF7 - 0.0367996 * BF8 + 0.0188678 * BF9 + 0.0304537 * BF11;

With:

BF1 = max (0, X5 - 2); Coefficient of BF1 is significant if the value of X5 (age of first use of drugs) greater than 2 is larger than 30 years old but if the value of X5 is smaller than 2 is 15-19 years old then BF1 coefficient is not significant
BF2 = max (0, 2 - X5); Coefficient of BF2 is significant if the value of X5 (age of first use of drugs) smaller than 2 i.e. 15-19 years old but if the value of X5 is greater than 2 then BF2 coefficient is not significant
BF3 = max (0, X8 - 2); Coefficient of BF3 is significant if the value of X8 (how to use drugs) greater than 2 i.e. inhale but if the value of X8 is smaller than 2 i.e. oral then BF3 coefficient is not significant
BF4 = max (0, 2 - X8); Coefficient of BF4 is significant if the value of X8 (how to use drugs) smaller than 2 i.e. oral but if the value of X8 is greater than 2 i.e. inhale then BF4 coefficient is not significant
BF5 = max (0, X3 - 1); Coefficient of BF5 is significant if the value of X3 (jobs) greater than 1 is privates/farmers/laborers or college student/students.
BF6 = max (0, X4 - 1); Coefficient of BF6 is significant if the value of X4 (marital status) is married.
BF7 = max (0, X1 - 1); Coefficient of BF7 is significant if the value of X1 (sex) greater is female.
BF8 = max (0, X7 - 1); Coefficient of BF8 is significant if the value of X7 (types of drugs) is drug list G.
BF9 = max (0, X6 - 1); Coefficient of BF9 is significant if the value of X6 (frequency of use) greater than 1 i.e. regular users or addict.
BF11 = max (0, X2 - 3); Coefficient of BF11 is significant if the value of X2 (education) is college.

Table 1. Variables that influence the reduction GCV value of drugs user rehabilitation

No	Variable	Level of importance (%)	-GCV
1	Age of first use of drugs (X5)	100.000	0.07828
2	How to use drugs (X8)	90.44231	0.06581
3	Marital Status (X4)	5.56588	0.00999
4	Jobs (X3)	4.87755	0.00994
5	Education (X2)	0.000	0.00959
6	Type of clinical care (X10)	0.000	0.00978
7	Frequency of use (X6)	0.000	0.00961
8	Illness Companion (X9)	0.000	0.00978
9	Types of drugs (X7)	0.000	0.00971
10	Sex (X1)	0.000	0.00971

In table 1 above it can be seen that the age of first use of drugs is the most variable on the MARS model with the importance level 100%. Then followed by successive how to using drugs, marital status and jobs with a large contribution to the model amount to 90.44231%, 5.56588% and 4.87755. Six variable have an importance level of 0.000% it means the variables are not included in the model because it is represented by variables that enter the MARS model. The minus GCV value indicates that if the first age drugs use variable is included in the model, the GCV value will decreased by 0,07828. If how to use drugs variable is included in the model, the GCV value will decrease by 0.06581. If marital status variable is included in the model, the GCV value will decrease by 0.00999. Also, if jobs variable is included in the model, the GCV value will decrease by 0.00994.

DISCUSSION

Rehabilitation is a process of comprehensively recovering dependence on drugs covering by psychosocial and spiritual aspects, so it takes time, hard will, patience, consistency and continuous learning. The goal of this rehabilitation service is drugs user, drugs abuse victims and the nearest person or family of drugs user. The purpose of this rehabilitation was to change the behavior toward positive and healthy living to improve the ability of emotional control thus avoiding legal issues, live more productively so that they are able to carry out their social function and as far as possible stop using drugs.

The majority of drugs user first use drugs in their teens due to social condition, psychologically requiring recognition identity and emotional instability. This is because was often identical with the search period of identity so that driven desire to try something new including using drugs. How to use drugs also affects the emergence of certain diseases. Some of the more risky drugs to brain such as decrease ability to think, remember and cognitive function.

Besides the strong desire of drugs user, family support is also important. Families who support drugs user to join the rehabilitation program can be used as a motivation in achieving success of the recovery program. Harmonious family relationship and gaining great support from the family is one of the success factors of rehabilitation. Family members should intensively assist and support drugs user in the rehabilitation program.

Lack of family support during the rehabilitation program or a degrading environment and disrespect for they efforts to recovery can add stress can make them vulnerable to use drugs or relapse.

The modeling of drugs user using MARS method processing by trial and error by combining the number of BF, MI and MO. The amount of BF used in this treatment is 2 to 4 times the number of predictor variables. MI which is 1, 2 or 3 under consideration if more than 3 will produce a very complex model. MO between knots used is 0, 1, 2 or 3.

The importance of each variable was estimated by the increase of GCV value. The importance of variable has a role to minimize the GCV value in the model. The classification table is another interesting way to state the feasibility of a model, i.e. how large the observation percentage precisely classified. Classification accuracy using the MARS method has been good at classifying the status of drugs user.

CONCLUSION

There are four important variables included in the best MARS model that is age of first use of drugs, how to use drugs, marital status and jobs. The accuracy level of drugs user rehabilitation classification in East Java using MARS method produces accuracy of 95,71% and misclassification of 4,29%.

REFERENCES

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* USA: Wiley.

Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley and Sons Inc

Breiman, L. (2001). *Random forest, machine learning.* Kluwer Academic Publisher.

Eubank, R. L. (1988). *Spline smoothing and non parametric regression.* New York: Marcell Dekker.

Friedman, J. H. (1991). *Multivariate adaptive regression spline.* Annal of Statistics.

Friedman, J. H. (1993). *FAST MARS.* Stanford University Department of Statistics, Technical Report 110.

Hastie, T., Tibshirani, R., Friedman, J. H. (2001). *The elements of statistical learnig: data minning, inference and Prediction. Second Edition* . New York: Springer-Verlag.

Hardle, W. (1990). *Applied nonparametric regression.* Cambridge University Press.

Johnson., W. D. (2002). *Data minning overview. applied multivariaate statistical analysis.* 5th ed. Prentice Hall.

Kutner, M. H. (2004). *Applied linear statistical models.* McGraw Hill.

Lewis, R. J. (2000). *An introduction to classification and regression trees (CART) analysis. annual meeting of the society for academic emergency medicine.* California: UCLA Medical Center.

Liaw, A. W. (2002). *Classification and regression by random forest.*

M. Nash, D. Bradford. (2001). *Parametric and nonparametric logistic regressions for prediction of presence/absence of an amphibian.* EPA Oct.

Rogers, D. (1992). Data Analysis using G/SPLINES. *In: Advance in neural processing information systems, 4, 1088-1095.* Morgan Kaufmann, San Mateo, CA.

Sutton, C. O. (2005). *Classification and regression trees, bagging and boosting.* Handbook of Statistics

Wezel, M. P. (2007). Improved customer choice predictions using ensemble methods. *European Journal of Operational Research*, 181 (1).