International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206442

Approach of Analysis of Data Mining Prediction In Earthquake Case Using Non Parametric Adaptive Regression Method

Dadang Priyanto [1], Muhammad Zarlis [2*], Herman Mawengkang [3], and Syahril Efendi [4]

[1] Graduate Program of Computer Science, Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

[2,3,4] Department of Computer Science, Faculty of computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

Article Info
Volume 6, Issue 4
Page Number: 247-253
Publication Issue :
July-August-2020

Article History
Accepted : 20 July 2020
Published : 27 July 2020

ABSTRACT

Data Mining is the process of finding certain patterns and knowledge from big data. In general, the data mining process can be grouped into two categories, namely descriptive data mining and data mining prediction. There are several Math functions that can be used in the data mining process, one of which is the Classification and Regression function. Regression Analysis is also called Prediction analysis, which is a statistical method that is widely used to investigate and model relationships between variables. Regression analysis to estimate the regression curve can be done by analyzing Nonparametric Regression. One well-known method in non-parametric regression is MARS (Multivariate Adaptive Regression Spline). The MARS method is used to overcome the weaknesses of the Linear Regression method. The use of a stepwise backward algorithm with the CQP quadratic programming framework (CQP) from MARS resulted in a new method called CMARS (Conic Multivariate Adaptive Regression Splines). The CMARS method is able to model high dimensional data with nonlinear structures. The flexible nature of the CMARS model can be used in the process of analyzing earthquake predictions, especially in Lombok, West Nusa Tenggara. Test results Obtained a mathematical model of four independent variables gives significant results to the dependent variable, namely Peak Ground Acceleration (PGA). Contributions of independent variables are the distance of the epicenter 100%, magnitude 31.1%, the temperature of the incident location 5.5% and a depth of 3.5%.

Keywords: MARS, Data Mining, CMARS, Prediction Analysis.

I. INTRODUCTION

Data mining is a well-known discussion in the digital era today, because it requires quite a lot of data, but its processing and utilization have not been maximized to produce new and useful information. Data mining according to Turban is a process that uses statistical, mathematical, artificial intelligence

and machine learning techniques to extract and add useful information and knowledge related to large databases. [1] Other experts say Data Mining is the process of finding certain patterns and knowledge from a large amount of data (Big Data). [2] Data sources can include basic data, data warehouse, web, repository information and others including data that is passed dynamically through the internet. Some things that can be done in data mining, such as Description, Estimation, Prediction, Classification, Clustering, and Association. [3] Description is a data mining activity that provides information as transparent as possible, and the results of the data mining model must illustrate clear patterns and be able to accept intuitive interpretations and explanations. Estimation is almost the same as the classification, the model is made with complete records and provides target variable values and predictors. To make new observations based on estimates of the target variable values based on predictor values. Prediction is a data mining activity that is similar to classification and estimation, the difference is that predictions provide results in the future. Classification has a target variable category, the data mining model examines a large number of records, and each record contains information about the target variable and a series of input or variable predictors. Clustering is a grouping of observations, notes, or cases in similar or homogeneous classes. Clustering is not the same as classification because there are no target variables in the grouping. Association is a data mining activity that finds the same attributes in the same activity. The aim is to uncover the rules for measuring the relationship between two or more attributes. There are several Math functions that can be used in the data mining process such as the Association and Correlation functions, Classification and Regression, Classification and outlier analysis. [2, 3] These mathematical functions can be used to find certain patterns as desired in the data mining process. In general, data mining processes can be grouped into two categories, namely descriptive data mining and Predicted data mining. [2] Descriptive Data Mining is the process of characterizing data properties into a target data set, while the Data Mining Prediction process is an induction on current data to make a prediction of the future. Many Data Mining Prediction methods can be used in forming mathematical or statistical relationships between several desired factors. Among them are Classification and Regression Trees (CART) methods, Artificial Neural Networks (ANNs) methods, Multiple Linear Regression (MLR) methods, Support Vector Machines (SVMs) methods, Multivariate Adaptive Regression Spline (MARS) methods and others. [4] Prediction Analysis is also called Regression analysis, which is a statistical method that is widely used to investigate and model relationships between variables. [5] In Regression Analysis to estimate the regression curve can be done in two ways, namely Parametric Regression and Nonparametric Regression. Nonparametric regression is one approach that is used to determine the pattern of relationships between predictor variables and responses that are not known for their regression curves or that there is no complete past information about the shape of data patterns. [6] Multivariate Adaptive Regression Spline (MARS) is one of the nonparametric regression models, which assumes the shape of the functional relationship between response variables and predictors is unknown. MARS is a complex combination of the spline method and recursive partitions to produce estimates of continuous regression functions, and is used for prediction and classification. [7] For management of high-dimensional data MARS has been developed into C-MARS (Conic Multivariate Adaptive Regression Splines Method) and is one of the effective non-parametric regression approach methods. The flexible nature of CMARS model can be implemented in various fields of application in the data mining process. [8]

A. Data Sets

The data mining process is very dependent on the use of data, because the data is used as a basis for input sources or inputs to be processed with an algorithm to provide output or output as desired inference. The data mining process usually uses a data set that is divided into three categories namely Data Training Set, Data Development Set and Data Testing Set. [9] Training data is data used in the process of building models. Data Development set or validation set is a set of data that is used to optimize when training the model. Data Development Set is used as a generalization model to find out generic patterns. Testing data is data used in the process of testing the performance of a model. In the process a dataset is needed that is a collection of data (sample in statistics). This sample is the data that we use to create models and evaluate models in the data mining process. One sample in a data set is called a data point or instance that represents a statistical event (event). The process of training, development, and testing, data is taken from the same distribution and has the same characteristics (independently and identically distributed). The distribution of each dataset should also be balanced and contain all cases. For example, a binary classification dataset should contain 50% of positive cases and 50% of negative cases. In general, the dataset distribution ratio is (80%: 10%: 10%) or (90%: 5%: 5%) (training: development: testing). Development sets in general can not be used if the dataset is small (only divided into training and testing sets only).

The data set of this study is earthquake data that occurred in Indonesia, particularly the events in Lombok, West Nusa Tenggara, in the events between 2010 and 2019. The total data set was 8.053 records in the form of an earthquake catalog. Data sets need to be processed to produce Peak Ground Acceleration (PGA) values with empirical calculations using the Joyner and Boore Attenuation functions. Filtering and classification need to be done, to get valid data to do the next process is prediction analysis.

I. METHODS AND MATERIAL

Non-parametric regression models can generally be written as follows: [10]

$$yi = f(xi) + \mathcal{E}i \qquad (1)$$

With yi = the response variable on observation i,
$f(xi)$ = $vector\ predictor\ function.$
$\mathcal{E}i$ = $is\ a\ free\ error\ i.$

For functions f(xi) with each component input xi (i = 1,2, ..., p) where p is the knots dimension $\tau = (\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,p})^T$, (i = 1,2, …, N) each input vector $xi = (xi, 1, xi, 2, \ldots xi, p)^T$
Furthermore, the nonparametric regression function commonly used is splines regression which is a linear piecewise function whose form is as follows:

$$c^+(x,\tau) = [+(x-\tau)]_+, c^-(x,\tau) = [-(x-\tau)]_+, \qquad (2)$$

Where $[q]_+ := \max\{0, q\}$ and τ are univariate knot, so the base function becomes:

$$f(x) = \{(xj - \tau)+, (\tau - xj) + |\tau \epsilon \{x1, j, x2, j, \ldots, xN, j\}, j \epsilon \{1,2, \ldots, p\}\} \qquad (3)$$

The function $f(x)$ is a representation of equation (1) which involves intercepts and linear combinations so that the model generally becomes: [10]

$$y = \theta_0 + \sum_{m=1}^{M} \theta_m \beta_m(x^m) + \varepsilon \qquad (4)$$

Where β_m with (m = 1,2, ..., M) is the basis of the function or product of equation (3). For θ_m is the basis function parameter to m (m = 1,2, ..., M) or a constant value (m = 0). The interaction of basis functions is

obtained by multiplying the basis functions by the truncated linear functions by involving new predictors. For the data provided ($\tilde{x}_i$, $\tilde{y}_i$) (i = 1,2, ..., N) with the base function m being: [11]

$$\beta_m(x^m) = \prod_{j=1}^{K_m} \left[S_{k_j}^{\ m} \left(X_{k_j}^{\ m} - \tau_{k_j}^{\ m} \right) \right] +, \qquad (5)$$

Where K_m is the number of truncated linear functions multiplied in basis functions to m. For $X_{k_j}^{\ m}$is an input variable that corresponds to j, and is related to a truncated function on the m to function base. $\tau_{k_j}^{\ m}$ is the value of knots in $X_{k_j}^{\ m}$ and $S_{k_j}^{\ m}$ is +/-. According to Friedman MARS has two algorithms for computational computation namely the first Forward Stepwise Algorithm where base functions are chosen to minimize the lack of fit until the maximum number of user-defined basic functions, M_{max}, is reached. The second algorithm, the Backward Stepwise Algorithm, which is to overcome weaknesses in the forward stepwise basic function that contributes the least amount to the residual squared error is eliminated, thus making the model simpler. Of the two algorithms, MARS in variable selection uses Generalized Cross-Validation (GCV). [12] in [10]

The C-MARS model as a form of backward stepwise development of the algorithm of the MARS model with Penalized Residual Sum of Squares (PRSS) which is described as follows: [10, 13]

$$PRSS = \sum_{i=1}^{N} \left(y_i - f(x_i) \right)^2 + \sum_{m=1}^{Mmax} \lambda m \sum_{\substack{|\alpha|=1 \\ \alpha=(\alpha 1,\alpha 2)^T}}^{2} \sum_{r<s} \int \theta^2_{m[D^{\alpha}_{r,s}\ \beta_m\,(t^m)]^2\, dt^m} \qquad (6)$$

In this equation, the set element $V_{(m)} = \left\{ \left(K_j^m \right) | j = 1,2, \dots, K_m \right\}$ calculate variables related to Function Base to m. β_m, and $t^m = \left(t_{m1,} \dots t_{mK_m} \right)^T$ represents the variable predictor vector which gives the basis function to m. Next :

$$D^{\alpha}_{r,s}\ \beta_m(t^m) \coloneqq \frac{\partial |\alpha| \beta_m}{\partial \alpha 1\ t_r^m\ \partial \alpha 2\ t_s^m}\ (t^m), \qquad (7)$$

For $\alpha = (\alpha_1, \alpha_2)^T$, $|\alpha| := \alpha_1 + \alpha_2$, where $\alpha_1, \alpha_2 \in \{0,1\}$

In equation (6), there are two parts of PRSS that are related to accuracy and complexity using the λm penalty parameter. The integral symbol " ∫ " indicates the dummy, which is ∫ Qm where Q_m is a number of parallel-pipe dimensions of integrated K_m. Because the integrals in equation (6) are multi-dimensional and difficult to evaluate, the integrals in the PRSS equation are set to

$$PRSS \approx \sum_{i=1}^{N} (yi - \theta^T \beta(d_i))^2 + \sum_{m=1}^{Mmax} \lambda_m \theta_m^2 \sum_{i=1}^{(N+1)^{Km}} \left(\sum^2_{\substack{|\alpha|=1 \\ \alpha=(\alpha 1,\alpha 2)^T}} \sum_{\substack{r<s \\ r,s \epsilon Vm}} \left[D^{\alpha}_{r,s}\, \beta m(X_i^m) \right]^2 \right) \Delta X_i^m, \qquad (8)$$

Where $\beta(\tilde{d}_i) \coloneqq \left(1,\ \beta_1\ (\tilde{x}_i^1), \dots, \beta M(\tilde{x}_i^M), \beta M + 1\ (\tilde{x}_i^{M+1}), \dots, \beta M_{max}(\tilde{x}_i^{Mmax})\right)^T$,
$\theta := (\theta_0, \theta_1, \dots, \theta_{Mmax})^T$ with point $\tilde{d}_i := (\tilde{x}_i^1, \tilde{x}_i^2, \dots, \tilde{x}_i^M, \tilde{x}_i^{M+1}, \dots, \tilde{x}_i^{Mmax})^T$ with argument $\left(\sigma^{Kj}\right) j \epsilon \{1,2, \dots, p\} \epsilon \{0,1,2, \dots, N+1\}^{Km}$ and

$$\tilde{x}_i^m = \begin{pmatrix} \tilde{x}_{l^{K_1^m}\sigma K_1^{m,K_1^m}}, \dots, \tilde{x}_{l^{K_{Km}^m}\sigma^{K_{Km}^m,K_{Km}^m}} \end{pmatrix}, \quad \Delta\tilde{x}_i^m := \prod_{j=1}^{Km} \begin{pmatrix} \tilde{x}_{K_j^m} & -\tilde{x}_{l^{K_j^m}\sigma^{Kj,K_j^m}} \\ l_{\sigma^{Kj+1,K_j^m}} & \end{pmatrix} \qquad (9)$$

Equation (8) can be simplified by using λ penalization on each derivative. Furthermore, the PRSS equation with Tikhonov regularization problem is as follows:

$$PRSS \approx \|y - \beta(\tilde{d})\theta\|_2^2 + \lambda \| L\theta \|_2^2, \qquad (10)$$

Can be reformulated into Conic Quadratic Problem (CQP) as follows [14, 15] in [10]

$$\begin{aligned} &\min_{t,\theta} \quad t, \\ &subject\ to \left|\left|\beta(d)\theta - y\right|\right|_2 \leq t, \\ &\qquad ||L\theta||_2 \leq \sqrt{\widetilde{M}} \end{aligned} \tag{11}$$

with t ≥ 0.

In equation (11) optimization can be solved by the interior point method (IPMs). There are several solutions to the problem of different M values, one of which is closest to the angle of the efficient curve or (L) with ‖L‖2 plotted versus ‖y-β (d̃) θ‖2. [5, 13]

II. RESULTS AND DISCUSSION

The results of empirical calculation of earthquake data processing that occurred in Lombok can be seen the pattern of earthquake spread as shown in Fig. 1 below:

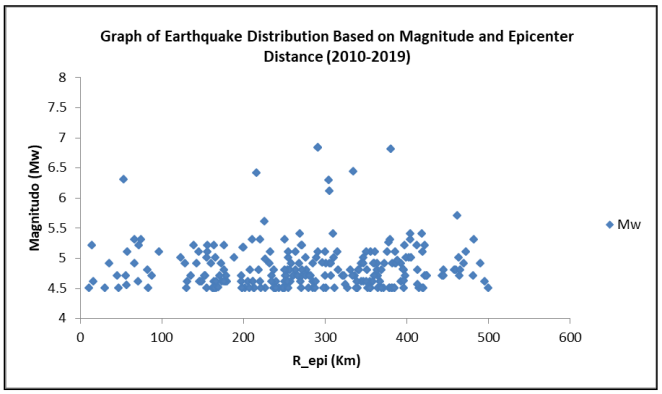

Figure 1. Distribution of earthquake spread in Lombok from 2010 to 2019

Seen in Figure 1, the magnitude that appears between 4 Mw to 7 Mw, earthquakes with this magnitude can be felt and can have a devastating effect, for earthquake data that has a smaller magnitude is removed. The distance of the epicenter as the x-ordinate axis is raised with a distance of 0 Km to 500 Km. This range is possible for earthquakes to have a detrimental impact, and epicenter distances greater than 500 Km are removed because it is confirmed that the vibrations are getting weaker and do not have any impact. The frequency of earthquakes in Lombok with a magnitude of 4 Mw to 7 Mw can be seen in the following table 1:

Table 1. Frequency of Earthquakes in Lombok

No	Magnitude (Mw)	Frequency
1	4.5 – 5	283
2	5 – 6	121
3	6 – 7	15

Furthermore, the solution using the non parametric MARS and CMARS models begins with the first step, namely the Forward Stepwise algorithm approach. This first step is for the combination of the input number of base functions (BF), input of the maximum interaction (MI) and the minimum number of observations (MO) to get the best model. The second step runs the Backward Stepwise algorithm by selecting the base function (BF) by removing the base function that does not contribute to the model. Of the 16 base functions entered there are 4 base functions that are omitted, namely the base function (BF) 4, 6, 8, 12. From the results of the selection in the second stage obtained a mathematical model for predictive analysis on Peak Ground Acceleration (PGA) variables such as equation (12) follows:

$$\begin{aligned} Y_{(PGA)} = {} & -0.0175733 - 0.00211487 * BF1 + 0.0029936 * BF2 + \\ & 0.000556472 * BF3 + 0.00172513 * BF5 + 0.000373726 * BF7 \\ & + 0.000369563 * BF9 - 0.000160793 * BF10 - 0.000689482 * \\ & BF11 + 0.000676173 * BF13 + 0.00329239 * BF14 - \\ & 0.00125948 * BF15 + 6.46282e\text{-}05 * BF16 \end{aligned} \tag{12}$$

MODEL PGA_G_ = BF1, BF2, BF3, BF5, BF7, BF9, BF10, BF11, BF13, BF14, BF15, BF16

Where Y (PGA) is the result of PGA prediction with MARS models with the contribution of each basis function (BF) is as follows:

BF1 = max(0, R_EPI - 64.642);

BF2 = max(0, 64.642 - R_EPI);

BF3 = max(0, MW - 4.7) * BF2;

BF5 = max(0, R_EPI - 31.8373);

BF6 = max(0, 31.8373 - R_EPI);

BF7 = max(0, R_EPI - 153.284);

BF8 = max(0, 153.284 - R_EPI);

BF9 = max(0, MW - 5.8) * BF8;

BF10 = max(0, 5.8 - MW) * BF8;

BF11 = max(0, SUHU__O_ - 27.4) * BF2;

BF13 = max(0, SUHU__O_ - 24.2) * BF6;

BF14 = max(0, MW - 5.1) * BF6;

BF15 = max(0, 5.1 - MW) * BF6;

BF16 = max(0, DEPTH - 1) * BF6;

After predicting the analysis using four dependent variables with a contribution level, namely the distance of the epicenter (R_epi) of 100%, Magnitude (Mw) of 31.1%, the temperature of the incident location (temperature) of 5.5% and depth (Depth) of 3.5 %. The results of the analysis obtained areas that have earthquake hazard level in Lombok as shown in table 2 below:

Table 2. Areas of earthquake hazard in Lombok

No	Lat	Long	Depth	Mw	R-epi	PGA(g)	SUHU (°)	Area location
1	-8.44	116	16	5.2	14.42382	0.18371517	26.7	Malaka, Pemenang
2	-8.36	116.22	12	6.2	27.391756	0.16732396	24.9	Genggelang, Gangga KLU
3	-8.41	116.16	17	5.5	19.222547	0.16606805	24.9	Tegal Maja, Tanjung, Pemenang KLU
4	-8.52	115.99	12	4.5	10.606472	0.16060569	27.5	Senggigi, Meninting, Mataram.
5	-8.42	116.03	23	5	16.880738	0.14393797	24.9	Senggigi, Malimbu
6	-8.43	116	13	4.6	15.833024	0.12335761	26.3	Mangsit, Senggigi

III.CONCLUSION

After going through the process of prediction analysis and empirical calculations it can be concluded that a mathematical model for the Peak Ground Acceleration (PGA) prediction analysis can be obtained, involving four independent variables that contribute, namely the epicenter distance of 100%, magnitude 31.1%, temperature of the event location 5.5 % and depth of 3.5%. The regions that have the highest values with earthquake hazard are Malacca, Genggelang, Tegal Maja, Senggigi, Mangsit and parts of Mataram City.

IV. REFERENCES

[1]. Turban, Efram, Aronson, Jay E, dan Peng-Liang, Ting, 2005, Decision Support Systems and Intelligent Systems, pearson

[2]. Han, J. Kamber, M. Pei, J. 2012, Data mining : concepts and techniques, Morgan Kaufmann, 225Wyman Street,Waltham, MA 02451, USA

[3]. Larose, D.T, 2005, Discovering Knowledge in Data: An Introduction to Data Mining, Wiley-Interscience, John Wiley & Sons, Inc., Hoboken

[4]. Yerlikaya, F., Askan, A., Weber, G.W., 2014, An alternative approach to the ground motion prediction problem by a non-parametric adaptive regression method, Engineering Optimization, Vol. 46, No. 12, 1651–1668.

[5]. Weber, G.W., ˙I. Batmaz, G. Köksal, P. Taylan, and F.Yerlikaya-Özkurt. 2012. "CMARS:A New Contribution to Nonparametric Regression with Multivariate Adaptive Regression Splines Supported by Continuous Optimization." Inverse Problems in Science and Engineering 20 (3): 371–400.

[6]. Eubank, R.L., 1999, Nonparametric Regression and Spline Smoothing, Second Edition, Marcel Deker, New York.

[7]. Friedman, J.H., 1991, Multivariate Adaptive Regression Spline (With Discussion), The Annals of Statistics, Vol. 19, hal. 1-141.

[8]. Weber, G.W., Cevik, A., 2018, Voxel-MARS and CMARS: Methods for Early Detection of Alzheimer's Disease by Classification of Structural Brain MRI, https://www.researchgate.net/publication/327338165.

[9]. Putra, J,W,G, 2018, Pengenalan Konsep Pembelajaran Mesin Dan Deep Learning, https://wiragotama.github.io/

[10]. Yerlikaya, F., Batmaz, I., Weber, G.W., 2014, A Review and New Contribution on Conic Multivariate Adaptive Regression Splines (CMARS): A Powerful Tool for Predictive Data Mining, In book: Modeling, Dynamics, Optimization and Bioeconomics I Edition: Springer Proceedings in Mathematics & Statistics}, Volume 73, 2014 Chapter: 38 Publisher: Springer Verlag.

[11]. Weber, G.W., ˙I. Batmaz, G. Köksal, P. Taylan, and F.Yerlikaya-Özkurt. 2012. "CMARS:A New Contribution to Nonparametric Regression with Multivariate Adaptive Regression Splines Supported by Continuous Optimization." Inverse Problems in Science and Engineering 20 (3): 371–400.

[12]. P. Craven and G.Wahba, "Smoothing noisy data with spline functions : estimating the correct degree of smoothing by the method of generalized cross-validation",Numerische Mathematik 31, 1979, pp. 377–403

[13]. F. Yerlikaya, "A New Contribution to Nonlinear Robust Regression and Classification with MARS and Its Application to Data Mining for Quality Control in Manufacturing", MSc., Middle East Technical University, 2008

[14]. A. Nemirovski, "A lectures on modern convex optimisation, Israel Institute of Technology, 2002. Available at http://iew3.technion.ac.il/Labs/Opt/opt/LN/Final.pdf.

[15]. P. Taylan, G.W. Weber and A. Beck, "New approaches to regression by generalized additive models and continuoues optimisation for modern applications in finance, science and technology", Journal Optimisation 56, 2007, pp. 675–698.

Cite this article as :

Dadang Priyanto, Muhammad Zarlis, Herman Mawengkang, Syahril Efendi, "Approach of Analysis of Data Mining Prediction In Earthquake Case Using Non Parametric Adaptive Regression Method ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 247-253, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206442
Journal URL : http://ijsrcseit.com/CSEIT206442