K-means and bayesian networks to determine building damage levels

Devni Prima Sari*^{1}, Dedi Rosadi^{2}, Adhitya Ronnie Effendie^{3}, Danardono^{4}
^{1,2,3,4}Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia
^{1}Department of Mathematics, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author, e-mail: devniprimasari@fmipa.unp.ac.id^{1}, dedirosadi@gadjahmada.edu^{2}, adhityaronnie@ugm.ac.id^{3}, danardono@ugm.ac.id^{4}

Abstract

Many troubles in life require decision-making with convoluted processes because they are caused by uncertainty about the process of relationships that appear in the system. This problem leads to the creation of a model called the Bayesian Network. Bayesian Network is a Bayesian supported development supported by computing advancements. The Bayesian network has also been developed in various fields. At this time, information can implement Bayesian Networks in determining the extent of damage to buildings using individual building data. In practice, there is mixed data which is a combination of continuous and discrete variables. Therefore, to simplify the study it is assumed that all variables are discrete in order to solve practical problems in the implementation of theory. Discretization method used is the K-Means clustering because the percentage of validity obtained by this method is greater than the binning method.

Keywords*: bayesian network, buildings damage, discretization, K-Means clustering, risk of earthquakes*

1. Introduction

We are often confronted by uncertainty when we have to make a decision. This uncertainty is bad information in making a decision because this uncertainty will affect the outcome. In decision-making, we will do some process; identify problems, determine opportunities and then solve them. One way to deal with uncertainty is to use probability theory. Since the sixteenth century [1], probability theory has been used to measure uncertainty and to assist in decision making. But a lot of problems in life require decision-making with a complicated process caused by the uncertainty of the process of connection that occurs in the system. For that, we must study the interrelationship between variables on the system by constructing it into a graph. The method used in this study is Bayesian Networks because this method is able to integrate expert knowledge and empirical data to model some variables. Unlike the naive Bayes, this method represents the dependence and independence of each variable present in a case. The delineation of relationships between variables in the graph provides an easy way to understand which variables interact with each other, in addition to being more efficient at calculating conditional and marginal probability [2]. Bayesian Network (BN) is the development of a Bayesian subjective approach supported by the advancement of computing.

BN has been developed in various fields, including in medical [3, 4], chemical [5], financial [6], and technical field [7]. As well as Bayesian networking applications in terms of minimizing the risk of natural disasters: floods [8], tsunamis [9, 10], earthquakes [11-14]. Disaster risk is essentially interesting to model, due to limited knowledge about when a disaster occurs. The implementation of Bayesian networks in disaster mitigation has also increased and is constantly updated to come true. Disaster mitigation is an activity that acts as an action to mitigate the impact of a disaster, or an attempt made to reduce the victims when a disaster occurs, both lives and goods. The first step in mitigation we should take is to conduct a disaster risk assessment in the region. In calculating the disaster risk of a region, we must know the vulnerability of a region based on characteristics of physical condition and its territory. Various ways have been done to predict the extent of damage to this building, one of which is with a regression analysis approach. This method is used to determine the magnitude of earthquake

Received July 19, 2018; Revised January 27, 2019; Accepted February 5, 2019

damage by observing some independent variables such as magnitude, intensity, depth, center spacing, and earthquake duration [15]. However, if necessary the addition of new variables associated with expert analysis and the relationship between independent variables then the regression analysis approach is less flexible to use. One appropriate modeling approach used for such circumstances is the Bayesian (BN) network model.

In most approaches to studying Bayesian network (BN), simplification of assumptions is made to avoid practical problems in the implementation of theory. The assumption used is all variables are discrete. But in reality, in the application, there is often a combination of continuous and discrete variables (mixed data). The possible solution is to discretize variables. In this study, the author uses the K-clustering method to discretize because the percentage of validity obtained by this method is greater than Binning method. The difference factor of this research with others research on disaster risk is in terms of data used. In this study, the authors use individual data building.

2. Research Method

In this part, we will summarily review Bayesian network and discretization method. Discretization method is used to discrete all continuous variables. While the BN model is to see the relationship between variables.

2.1. Method Discretization

Discretization is a continuous variable transformation into a discrete variable. It is often performed in data analysis to aid understanding, grouping multiple values of continuous attributes, and continuous domain partitions into non-overlapping intervals [16]. In our study, we use unsupervised discretization methods. This method does not require a categorical variable as the target variable to serve as the basis for its discretization. This method divides the continuous value interval into several sub-hoses based on user considerations. The considerations taken are subjective in which the user determines the discretization mechanism. The two discretization methods used are Equal Width Interval Binning and K-Means clustering.

2.1.1. Equal Width Interval Binning

The equal width binning interval is the simplest method of discriminating data and has often been applied. This involves sorting the observed values of the continuous feature and dividing the range of observed values for variables into k bin of the same size, where *k* is the parameter specified by the user. If the x variable observed has a value limited by x_{min} and x_{max} then this method calculates the bin width and constructs bin boundaries, or thresholds, at $x_{min} + i\delta$, where $i = 1, \dots, k - 1$. The method is applied to each continuous feature independently [17].

$$\delta = \frac{x_{max} - x_{min}}{k} \tag{1}$$

2.1.2. K-Means Clustering

K-Means is one of the clustering algorithms. The function of this algorithm is to partition the existing data into one or more clusters. In this learning algorithm, computers classify their own data into the input without knowing first the target class. This learning is included in unsupervised learning. The received input is the desired data or object and *k* group (cluster). In each cluster, there is a centroid that represents the cluster. The K-Means clustering method is one of the most attractive methods of grouping. The algorithm was established as a top 10 algorithm at the IEEE International Conference on Data Mining (ICDM) in December 2006. This algorithm is considered one of the most common data mining algorithms in the research society [18]. The stages of the algorithm in the K-Means clustering are as follows [19]:

1. Select *k* centroid point randomly
2. Group the data to form k clusters with the centroid point of each cluster is the centroid point that has been selected previously
3. Update the centroid point value
4. Repeat steps 2 and 3 until the value from the centroid point no longer changes

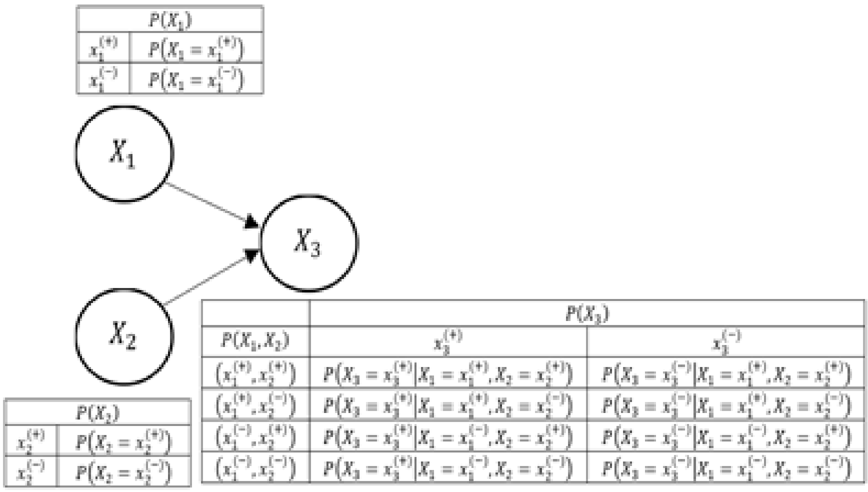
The process of grouping data into a cluster can be done by calculating the closest distance from a data to a centroid point. The formula for calculating distances is [20]:

$$d\left(x_i, x_j\right) = \left(|x_{i1} - x_{i1}|^g + |x_{i2} - x_{i2}|^g + \cdots + \left|x_{ip} - x_{ip}\right|^g\right)^{1/g} \quad (2)$$

where:
$g = 1$, to calculate the Manhattan distance
$g = 2$, to calculate the Euclidean distance
$g = \infty$, to calculate the distance of Chebychev
x_i, x_j are two pieces of data to be calculated
p = the dimension of a data

2.2. Bayesian network

Bayesian network is one of the probabilistic graphical models constructed from probabilistic theory and graph theory. The probabilistic theory is directly connected to the data while the graph theory is directly related to the form of representation to be obtained. Knowledge is represented qualitatively using the graph structure and quantitatively using numerical parameters. Bayesian network method consists of two main parts, namely the Directed Acyclic Graph (DAG) and Conditional Probability Table (CPT). DAG is a directed graph without cycles. DAG consists of nodes and edges. Nodes represent variables and random sides represent direct dependency relationships and can also be interpreted as a direct effect (cause-and-effect) between related variables. No edges indicate a free relationship between variables. Each node in the network structure has a CPT. Quantitative parameters in the network show a probabilistic relationship between the child and parents and are expressed by the conditional probability distribution. Each node is associated with a set of conditional probabilities, $P\left(X_i \middle| \mathbf{pa}(X_i)\right)$ so that X_i is the variable associated with the node and $\mathbf{pa}(X_i)$ is the set of the parent in the graph. In Figure 1 it is seen that each node in the network structure has its own conditional probability table.



$P(X_1)$	
$x_1^{(+)}$	$P(X_1 = x_1^{(+)})$
$x_1^{(-)}$	$P(X_1 = x_1^{(-)})$

$P(X_2)$	
$x_2^{(+)}$	$P(X_2 = x_2^{(+)})$
$x_2^{(-)}$	$P(X_2 = x_2^{(-)})$

	$P(X_3)$	
$P(X_1, X_2)$	$x_3^{(+)}$	$x_3^{(-)}$
$(x_1^{(+)}, x_2^{(+)})$	$P(X_3 = x_3^{(+)} \mid X_1 = x_1^{(+)}, X_2 = x_2^{(+)})$	$P(X_3 = x_3^{(-)} \mid X_1 = x_1^{(+)}, X_2 = x_2^{(+)})$
$(x_1^{(+)}, x_2^{(-)})$	$P(X_3 = x_3^{(+)} \mid X_1 = x_1^{(+)}, X_2 = x_2^{(-)})$	$P(X_3 = x_3^{(-)} \mid X_1 = x_1^{(+)}, X_2 = x_2^{(-)})$
$(x_1^{(-)}, x_2^{(+)})$	$P(X_3 = x_3^{(+)} \mid X_1 = x_1^{(-)}, X_2 = x_2^{(+)})$	$P(X_3 = x_3^{(-)} \mid X_1 = x_1^{(-)}, X_2 = x_2^{(+)})$
$(x_1^{(-)}, x_2^{(-)})$	$P(X_3 = x_3^{(+)} \mid X_1 = x_1^{(-)}, X_2 = x_2^{(-)})$	$P(X_3 = x_3^{(-)} \mid X_1 = x_1^{(-)}, X_2 = x_2^{(-)})$

Figure 1. BN structure and conditional probability table [14]

Let $U = \{X_1, \cdots, X_n\}$ is a set of variables. If the joint probability is $P(U) = P(X_1, \cdots, X_n)$, then $P(X_i)$ and $P(X_i|e)$ can be calculated, where e is the evidence of several variables in the Bayesian network. Proposition 1 [21] (The general Chain rule). Let $U = \{X_1, \cdots, X_n\}$ be a set of variables. Then for any probability distribution $P(U)$ we have

$$P(U) = P(X_n|X_1, \cdots, X_{n-1})P(X_{n-1}|X_1, \cdots, X_{n-2}) \cdots (X_2|X_1)P(X_1)$$

Theorem 1 [22] (The chain rule for Bayesian networks). Let BN be a Bayesian network over $U = \{X_1, \cdots, X_n\}$. The joint probability distribution $P(U)$ is the multiplication of all conditional probabilities of each node in BN:

$$P(U) = \prod_{i=1}^{n} P\left(X_i \middle| \mathbf{pa}(X_i)\right) \tag{3}$$

where $\mathbf{pa}(X_i)$ are the parents of X_i in BN.

The BN model can be constructed from the data through a learning process that includes structural learning and parameter learning. BN is one of the leading methods in the field of artificial intelligence because of its ability to learn data. Let $P(U)$ be the combined probability distribution, and $\boldsymbol{e}$ is the previous search. The combined probability distribution is obtained from by substituting all entries with X not in state *i* or *j* with a zero value and allowing other entries to remain. This is equivalent to multiplying $P(U)$ by $\boldsymbol{e}$

$$P(U, e) = P(U) \cdot \boldsymbol{e}$$

Theorem 2 [21] Let BN be a Bayesian network over the universe U, and let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m$ be findings. Then

$$P(U, e) = \prod_{X \in U} P\left(X \middle| \mathrm{pa}(X)\right) . \prod_{i=1}^{m} \boldsymbol{e}_i$$

and for $X \in U$

$$P(X|e) = \frac{\sum_{U \setminus \{X\}} P(U,e)}{P(e)} \tag{4}$$

3. Results and Analysis

The object of the study is the building that was damaged by the earthquake in West Sumatra in 2009. The West Sumatra earthquake in 2009 was one of the biggest earthquakes that occurred in Indonesia with a power of 7.9 MW. The quake precisely occurred at 10:16 pm local time on September 30, 2009, Indonesia. The object of interest is the building located in Padang which consists of 11 districts. Data sets are based on building units. Data obtained from the Regional Disaster Management Agency of Padang and the Indonesian Meteorology, Climatology and Geophysics Agency. The data to be used as a case study consists of 61344 building data. The research data consisted of three independent variables (close to faults, slope, and epicenter distance) and four dependent variables (construction, landslide risk, PGA, and damage). The description of variables can be seen in Table 1.

Table 1. Variables and their Descriptions in BN of Seismic Vulnerability

Code	Variable	Description
C	Construction	If the structure of different buildings then the resistance to the earthquake disaster will also be different.
P	PGA	Peak Ground Acceleration (PGA) is the ground acceleration caused by seismic wave propagation. The magnitude of this PGA can be made a response spectrum that will be used as a material evaluation of the strength of the building against the earthquake. This value indicates the seismic risk (hazard) required for mitigation and design of earthquake resistant structures [23].
E	Epicenter Distance	The epicenter is the point on the Earth's surface directly above a hypocenter or focus, the point where an earthquake or an underground explosion originates. Epicenter distance is the distance between the epicenter and the building.
L	Landslide risk	The effect of a ground shaking is the main earthquake hazard. The ground shaking may also cause landslides.
S	Slope	Slopes is assumed to cause mudslides that cause damages to buildings [24].
F	Close to faults	If the building is closer to the fault, then the risk of damage to the building is greater [25].
D	Damage	The level of damage consists of three levels namely low, medium, and high. (BPBD: Regional disaster office)

Figure 2 explicit spatial modeling process for our approach where GIS, spatial analysis, and BN are used to assess damage rates. Our modeling is done on a data set, consisting of points, spacing, and area. Since the existing variables consist of continuous and discrete variables, we discretize the data. We can extract the data from the data set that has been discretized for further analysis. Each of these data has predictive indicators and target variables. We further established the Bayesian network structure based on expert analysis. In the model, ArcGIS is used to transfer maps to numeric values, and GeNIe is used to modeling and learning Bayesian networks. We use MatLab da R as a machine to support calculations. In system modeling, we have developed some special functions including data conversion, construction, and learning from BN.

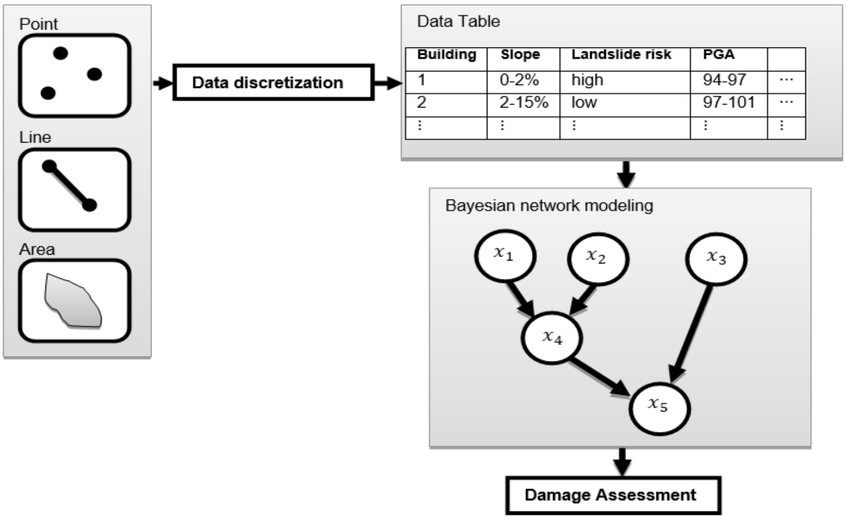


Figure 2. Spatial modeling uses bayesian networks

The data in this study consist of continuous variables (E, P, and F) and discrete variables (C, S, L, and D). Since we use the assumption that all discrete variables, we will discretize data on all continuous variables by means of the K-Means clustering method. Clustering for all three continuous variables looks like in the Figure 3, where each variable E and P consists of four clusters, while the F variable consists of three clusters. After doing the clustering process as in Figure 3, we will get the summary of sample as in Table 2. In this study, a reference opinion of experts in [11] is used.

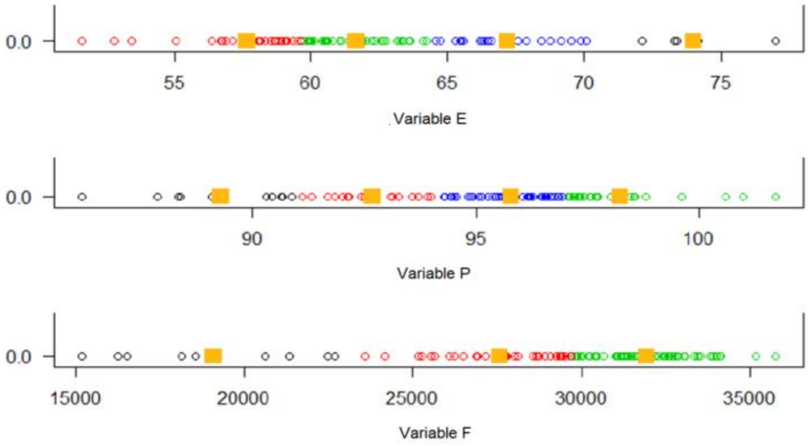


Figure 3. K-Means clustering of continuous variables

After the Bayesian network structure in Figure 4 is formed, the next step is to calculate the probability of each node. Neither the probability of unaffected nodes (exogenous variables) nor the conditional probabilities for affected nodes (endogenous variables). Furthermore, we perform significance tests for each arc using the arc.strength function in R. The arc.strength function is used to measure the strength of the relationship between each arc. The value of power disclosed is the resulting p-value. From Table 3, we can see that all arcs have p-values smaller than 0.05. This means the structure in Figure 4 is significant and we can advance the next process.

Table 2. The Summary of Sample

Variabel	State	Description of State	Number	Probability
Construction (C)	1	Wood	1566	0.03
	2	Semi Permanent	5912	0.10
	3	Permanent	53866	0.88
PGA (P)	1	86.19-90.89	1991	0.03
	2	91.11-93.99	18687	0.30
	3	94.27-96.94	25661	0.42
	4	97.07-101.71	15005	0.24
Epicenter distance (E)	1	51.62-59.62	15005	0.24
	2	59.78-64.22	25661	0.42
	3	64.56-70.09	18687	0.30
	4	72.14-77.02	1991	0.03
Landslide risk (L)	1	Low	841	0.01
	2	Medium	60503	0.99
Slope (S)	1	0-2%	54493	0.89
	2	2-15%	713	0.01
	3	15-40%	2209	0.04
	4	>40%	3929	0.06
Close to faults (F)	1	15164.33-22683.49	4219	0.07
	2	23574.32-29712.09	26178	0.43
	3	29813.73-35780.49	30947	0.50
Damage (D)	1	Slight	22564	0.37
	2	Medium	21504	0.35
	3	Weight	17276	0.28

Table 3. Result of Significance

No.	From	To	Strength
1	E	P	0.00E+00
2	F	L	0.00E+00
3	S	L	0.00E+00
4	P	D	9.40E-171
5	F	D	1.32E-238
6	L	D	9.01E-09
7	C	D	8.53E-25

The next step determines the probability of nodes based on the structure of the BN formed. Based on the structure of BN in Figure 4 the complete joint probability distribution of BN is:

$$P(D,P,E,L,F,S,C) = P(D|P,L,F,C).P(P|E).P(E).P(L|F,S).P(F).P(S).P(C)$$

we do marginalization to get *P(D)*

$$P(D) = \sum_{P,E,L,F,S,C} P(D|P,L,F,C).P(P|E).P(E).P(L|F,S).P(F).P(S).P(C)$$

The number of BN parameters can be calculated with the nparams function and the value is 178, as expected from the set of parameters of the local distribution. Marginal probability distribution for each value of D is:

$$P(D=1)=\sum_{P,E,L,F,S,C} P(D=1|P,L,F,C).P(P|E).P(E).P(L|F,S).P(F).P(S).P(C)=0.3546$$

$$P(D=2)=\sum_{P,E,L,F,S,C} P(D=2|P,L,F,C).P(P|E).P(E).P(L|F,S).P(F).P(S).P(C)=0.3514$$

and,

$$P(D=3)=\sum_{P,E,L,F,S,C} P(D=3|P,L,F,C).P(P|E).P(E).P(L|F,S).P(F).P(S).P(C)=0.2940$$

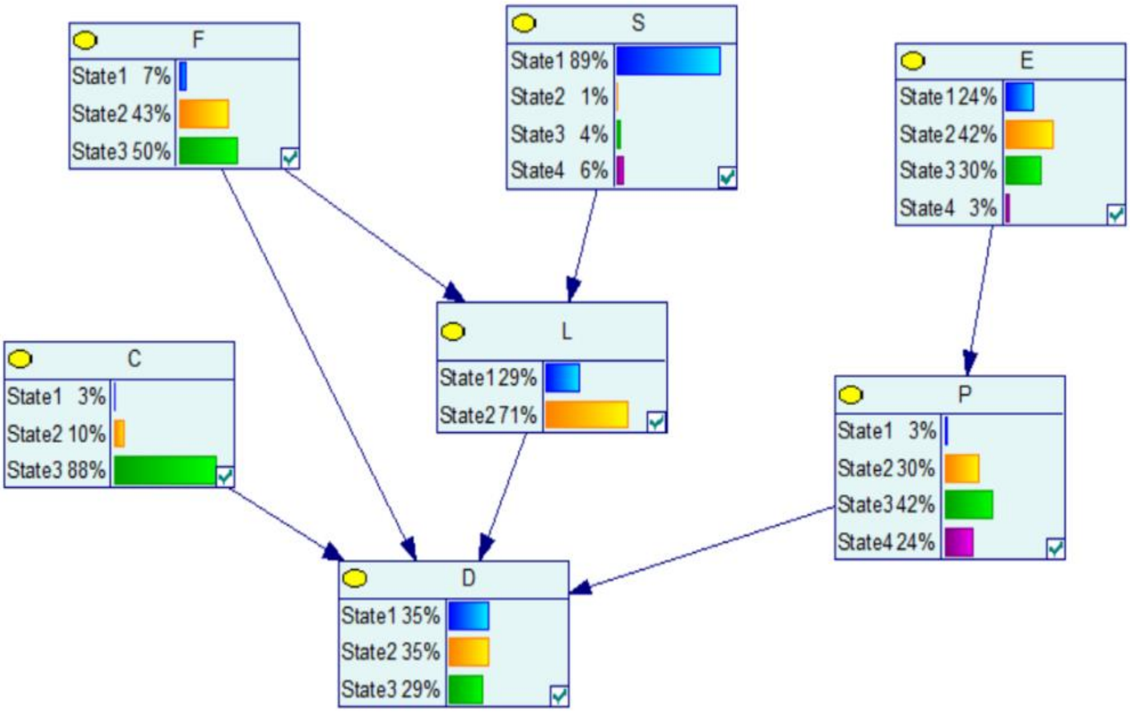


Figure 4. Predicted model of damage level with Bayesian network

From the model on Figure 4, the probability level for the level of damage to buildings slight, medium, and high are 35.46%, 35.14%, and 29.4% respectively. The level of accuracy is the comparison of predicted results with the BN model and the real conditions. The level of accuracy of building damage to the Padang city is

$$\text{Level of accuracy}=\frac{43018}{61344}=70\%$$

If we look at the level of accuracy of building damage per region, the highest accuracy is achieved by the Padang Timur subdistrict, where the accuracy rate reaches 89%.

Classification results using K-Means clustering and expert analysis have 70% accuracy rate. In this case, we also try to compare by using equal width interval binning method. The virtual can be seen in Table 4. From Table 4 we can see that K-Means clustering method provides a better degree of accuracy than the equal width interval binning method.

Table 4. Comparison of the Level of Accuracy Between K-Means Clustering and Equal Width Interval Binning Method

Damage state	K-Means clustering	Equal Width Interval Binning
Slight	0.756072	0.34165
Medium	0.763207	0.742513
Weight	0.552558	0.307189
Level of accuracy	0.701258	0.472467

4. Conclusion

The significant contribution of this model is the development of K-Means clustering discrimination method and Bayesian network especially for earthquake damage case in Indonesia by using individual data of the building. Although the use of K-Means clustering can increase the level of accuracy but further research is needed so that the level of accuracy of damage to the building of the Padang city increases. To improve accuracy, we can use primary data so that all the variables that play a role can provide information.The improvement of the results is expected to be applied in determining the amount of premium and insurance claims of building damage due to the earthquake. In addition, the results of this study assist the government in reserve for recovery funds for disaster-prone areas.

References

[1] King AC. Pathways to Probability: History of the Mathematics of Certainty and Chance. New York: Holt, Rinehart and Winston. 1963.

[2] Cheng J. Efficient Stochastic Sampling Algorithms for Bayesian Networks. Dissertation. Pitsburgh: University of Pitsburgh; 2001.

[3] Flores MJ, Nicholson AE, Brunskill A, Korb KB, Mascaro S. Incorporating Expert Knowledge When Learning Bayesian Network Structure: A Medical Case Study. *Artificial Intelligence in Medicine*. 2011; 53(3): 181– 204.

[4] Sari DP, Rosadi D, Effendie AR, Danardono. *Designing Bayesian Network Structure of Data*. Proceedings of the 18th National Conference of Math. Pekanbaru. 2016: 84-90.

[5] Zhang Q, Zheng X, Zhang Q, Zhou C. Inferring Gene Regulatory Network from Bayesian Network Model Based on Re-Sampling. *Telkomnika*. 2013; 11(1): 215-222.

[6] Neil M, Marquez D, Fenton N. Using Bayesian networks to Model the Operational Risk to Information Technology Infrastructure in Financial Institutions. *Journal of Financial Transformation*. 2008; 22: 131–138.

[7] Sembiring J, Sipayung JP, Arman AA. Application Development Risk Assessment Model Based on Bayesian Network. *Telkomnika.* 2018; 16(3): 1376-1385.

[8] Zhang SZ, Yang NH, Wang XK. *Construction and Application of Bayesian Networks in Flood Decision Supporting System*. Proceedings of the 1st Intl. Conference on Machine Learning and Cybernetics. Beijing. 2002; 2: 718-722.

[9] Blaser L, Ohrnberger M, Riggelsen C, Babeyko A, & Scherbaum F. Bayesian Networks for Tsunami Early Warning. *Geophysical Journal International*. 2011; 185(3): 1431–1443.

[10] Yadav RBS, Tsapanos TM, Tripathi JN, Chopra S. An evaluation of tsunami hazard using Bayesian approach in the Indian Ocean. *Tectonophysics*. 2013; 593:172–82.

[11] Bayraktarli YY, Baker JW, Faber MH. Uncertainty treatment in earthquake modeling using Bayesian probabilistic networks. *Georisk*. 2011; 5(1): 44-58.

[12] Li L, Wang JF, Leung H. Using Spatial Analysis and Bayesian Network to Model the Vulnerability and Make Insurance Pricing of Catastrophic Risk. *International Journal of Geographical Information Science*. 2010; 24(12): 1759-1784.

[13] Li L, Wang JF, Leung H, Zhao S. A Bayesian Method to Mine Spatial Data Sets to Evaluate the Vulnerability of Human Beings to Catastrophic Risk. *Risk Anal*. 2012; 32(6): 1072–92.

[14] Sari DP, Rosadi D, Effendie AR, Danardono. *Application of Bayesian Network Model in Determining the Risk of Building Damage Caused by Earthquakes*. Proceedings of the 1st Intl. Conference on Information and Communications Technology (ICOIACT). Yogyakarta. 2018: 131-135.

[15] Urrutia JD, Bautista LA, Baccay EB. *Mathematical Models for Estimating Earthquake Casualties and Damage Cost Throught Regression Analysis Using Matrices*. Journal of Physics: Conference Series. 2014; 495(1): article id. 012024.

[16] Gupta A, Mehrotra KG, Mohan C. A Clustering-Based Discretization for Supervised Learning. *Statistics & Probability Letters*. 2010; 80(9–10): 816-824.

[17] Dougherty J, Kohavi R, Sahami M. *Supervised and Unsupervised Discretization of Continuous Features*. Proceedings of the 12th Intl. Conference on Machine Learning. California. 1995: 194-202.

[18] Wu X, Quinlan VKR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, et al. Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*. 2008; 14(1): 1-37.

[19] Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. First Edition. Boston: Addison-Wesley Longman. 2005.

[20] Maimon O, Rokach L. *Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications*. Singapore: World Scientific. 2005.

[21] Jensen FV, Nielsen TD. Bayesian Network and Decision Graph. New York: Springer Science & Business Media. 2007.

[22] Zhou Y, Fenton N, Neil M. Bayesian Network Approach to Multinomial Parameter Learning Using Data and Expert Judgments. *International Journal of Approximate Reasoning*. 2014; 55(5): 1252–1268.

[23] Ahmadi M, Nasrollahnejad A, Faraji A. *Prediction of Peak Ground Acceleration for Earthquakes by Using Intelligent Methods*. Proceedings of the 5th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS). Tehran. 2017: 7-12.

[24] Jang J, Lee C, YCA Preliminary Study of Earthquake Building Damage and Life Loss Due to the Chi-Chi Earthquake. *Journal of the Chinese Institute of Engineers*. 2002; 25(5): 567–576.

[25] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag. 2001.