International Journal of Engineering & Technology, 7 (4.38) (2018) 1142-1145

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Bayesian Approach to Prediction of Flood Risks

Nur Izzati Mohd Roslin, Aida Mustapha, Noor Azah Samsudin, Nazim Razali

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, 86400, Malaysia
**Corresponding author E-mail:zateyroslin@gmail.com*

Abstract

Flood is a temporary overflow of a dry area due to overflow of excess water, runoff surface waters or undermining of shoreline. In 2014, Malaysia grieved with the catastrophic flood event in Kuala Krai, Kelantan, which sacrificed human lives, public assets and a total of RM 2 billion loss. Due to uncertainties in flooding event, this research is set to compare three variations of Bayesian approaches in classifying the risk of flood into two classes; flood or no flood. The study involved data from Kuala Krai, which serves as the main observation point. The dataset contains six attributes, which are water level, rainfall daily, rainfall monthly, wind, humidity, and temperature. The classification experiment will be conducted using three variants of Bayesian approaches, which are Bayesian Networks (BN), Naïve Bayes (NB), and Tree Augmented Naive Bayes (TAN). The outcomes of this research will show the best algorithm performance in term of accuracy for three Bayesian-based learning prediction algorithms. In the future, this prediction system is hoped to assist related agencies in Malaysia to categorize land areas that face high risk of flood so preventive actions can be planned in place.

Keywords: *Rainfall, Flood, Risk Prediction, Bayesian.*

1. Introduction

Malaysia is a country comprising Peninsular Malaysia, Sabah, and Sarawak. It covers fourteen states that are Perlis, Kedah, Penang, Perak, Selangor, Negeri Sembilan, Pahang, Melaka, Johor, Kelantan, Terengganu, Sabah, and Sarawak. Asian nation additionally has one central consisting of three Territories that are Federal Territory of Malaysian capital, Federal Territory of Labuan and Federal Territory of Putrajaya. Malaysia has two main areas separated by the South China Sea. The northern border is Thailand and the southern border is Singapore. Meanwhile, the border of Indonesia on the south and Brunei on the north. Malaysia is located near the equatorial line at the Latitude 1˚ and North 7˚ and 100˚ and 100˚ East. Malaysia covers 329,960.22 km [1]. Malaysia has hot and humid weather throughout the year. The average daily temperature throughout Malaysia is between 21˚C to 32˚C. Typically, the Malaysian climate is experiencing a strong equator influenced by the north eastern monsoon from November to March and the western monsoon from June to October. The annual rainfall is very high which is 2500 mm in Peninsular Malaysia between 2300 mm in Sarawak, and 3300 mm in Sabah [2].

Due to the high rainfall and river flow, the risk of flood in Malaysia is very high. Flood can be defined as a situation where water flows exceed the carrying capacity of a river resulting in overflows over the river banks [3]. There are several factors that can cause flood such a sudden rise in water levels such as continuous rainfall, land humidity and non-smooth water drainage. One other contributing factor the uncontrollable rapid development. Widespread land clearing and overcutting trees causes water absorption to land to decline and runoff continues to the river more rapidly. For every increase of development rate between 0-40 percent, it will result in a flow rate of 190 percent, hence twice the runoff speed.

In addition, the rate of erosion will increase resulting in increased silt in the river. Shallow river will have a lower capacity, unable to accommodate the increased water and cause the water to flood the cliffs. Not to ignore the river basin, which can also cause flood. The size of a large river basin will have a large run of water when heavy rain. If the river capacity is insufficient, floods will occur. According to [4], Malaysia was shocked by the news of the catastrophic natural disaster that flooded Kelantan especially at Kuala Krai, Kelantan back in 2014.

Studies on flood prediction has been very active especially in the recent movement on awareness of climate change especially using Bayesian approaches due to its ability to deal with uncertainties. Most recently in 2018, [5] proposed a dynamic flood assessment and discovered that urban underground facilities tend to be prone to flood due to breaking of a dam or a barrier, or a flash flood when exceptional degree of rain occurred. Rapid and dynamic assessment of underground flood evolution method vital for safety evacuation and reduce disaster. The research proposed an integrated framework using Bayesian Networks to rapidly and dynamically access the flood evolution method and consequence in underground areas. In the networks, 17 nods represented the flood disaster drivers, flood disaster bearers, flood mitigation action, and additionally on the spot feedback data. The results showed that the projected framework was especially helpful for dynamically evaluating underground flood evolution method and to spot the crucial influencing factors.

When dealing with dynamic systems, [6] focused on flood data to test three different dynamic algorithms with different tools, which were Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Dynamic Evolving Spiking Neural Network (deSNN). The proposed algorithm, deSNN, achieved best accuracy rate in predicting flood when fed with Spatio/Spectro Temporal Data Modelling (SSTD). SSTD data is not supported with existing data mining tool such as WEKA. The analysis of the data was based on the analysis of space and time. As a comparative experiment,

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

conventional machine learning methods such as MLP and SVM are used as a baseline performance and accuracy measures.

In an alpine catchment, [7] used different precipitation data for flood prediction to accurately predict such events, accurate and representative precipitation data required. In the study, three value of precipitation datasets commonly used in hydrological studies were investigated. The datasets include station network precipitation (SNP), interpolated grid precipitation (IGP), and radar-based precipitation (RBP). They performed a Bayesian uncertainty analysis with an improved description of model systematic errors to quantify their effects on runoff simulations.

Monthly precipitation forecast by [8] used Bayesian approach technique for monthly mean precipitation prediction at twenty-one stations in Assam, India. The interstation precipitation dependencies and independencies are delineating mistreatment Bayesian Networks (BN) structure and five atmospherically variables including temperature, relative humidity, wind speed, overcast, and southern oscillation index were used as predictors. The research aimed to match between two different structural learning rules in Bayesian Networks, which were K2 and Markov Chain Monte Carlo (MCMC) algorithm. 13 different models are developed with different combinations from 5 predictors. At the end of this experiments, K2 algorithms outperformed MCMC algorithms for all combinations.

According to [9], rainfall thresholds are primarily based flood warning. Therefore, it is important to derive the likelihood of providing flood warnings at given water course sections based on comparison of quantitative precipitation forecast with important precipitation threshold values although this was not necessarily the requirement of real time statement system. The proposed resulted in an especially simplified alert system employed by non-technical stakeholders and may be used additionally to support the normal flood statement system just in case of system failures.

This paper presents Bayesian approaches in predicting flood risk into flood or no flood. The study involved data from Kuala Krai, Kelantan, which serve as the main observation point. The dataset contains six attributes, which are water level, rainfall daily, rainfall monthly, wind, humidity, and temperature. This research aims to develop and compare between three variations algorithms which are a Bayesian Network, Naive Bayes and Tree Augmented Naive Bayes for flood prediction.

The remaining of this paper is organized as follows. Section 2 presents the CRISP-DM methodology used to perform the data mining task along with the dataset and the evaluation metrics. Section 3 presents the results and finally Section 4 concludes with some direction for future work.

2. Materials and Methods

The flood risk prediction model in this paper will be developed using the Cross Industry Standard Process for Data Mining (CRISP-DM) approach. CRISP-DM is an abstract, high-level model for data processing and it is additionally general enough to be used for different information analysis desires [10]. In general, it describes the process of data mining in six phases as visualized in Fig. 1.

The process of CRISP-DM begins with phase of business understanding, data understanding, data preparation, modelling, evaluation and deployment. In the figure, the arrows represent the most important dependencies between phases. The large outer circle indicates the iterative nature of this framework which going back and forth between steps is often needed, as findings along the way trigger new questions [11].

The experiments were carried out using the WEKA tool [12] with 10-fold validation method for training and testing. Cross-validation is a statistical method that can be used to evaluate the performance of the algorithm where the data is separated into two subsets: learning process data and validation or evaluation data. Algorithms are trained by a subset of learning and validated by a subset of validation. Furthermore, cross-validation selection can be based on the dataset size. Cross-validation method with k-fold is often used because it can reduce computational time while maintaining accurate estimates.

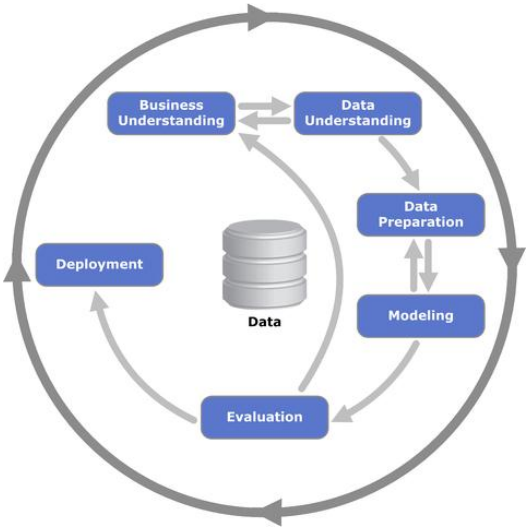


Fig. 1: CRISP-DM Process Model for Data Mining [13].

3.1. Dataset

In order to test the flood risk prediction model developed, the model will be tested using the flood data in Kuala Krai, Kelantan. The data was recorded from 1st January to 31th December between 2012 and 2016 extracted from [14] and [15]. The dataset consists of 1,828 instances and each is described by rainfall monthly (RF Month), rainfall daily (RF Daily), water level (in cm), humidity, wind and temperature. The features correspond to predict a binary class of flood and no flood. The excerpt of the dataset is shown in Fig. 2.

Date	Level(cm)	RF Month(mm)	RF Daily(mm)	Temperature	Humidity	Wind (m/s)	class
1/1/2012	1871	1057	45	24.2	92.8	0.7	NOFLOOD
2/1/2012	1911	1058	1	24.1	92.8	0.6	NOFLOOD
3/1/2012	1799	1064	6	24.7	91.2	0.7	NOFLOOD
4/1/2012	1763	1064	0	25	82.8	0.9	NOFLOOD
5/1/2012	1738	1064	0	24.3	83.7	1	NOFLOOD
6/1/2012	1721	1064	0	24.6	80.5	0.9	NOFLOOD
7/1/2012	1711	1064	0	24.3	82.1	1.1	NOFLOOD
8/1/2012	1703	1071	7	24.6	85.5	0.6	NOFLOOD
9/1/2012	1703	1071	0	24.2	89	0.5	NOFLOOD
10/1/2012	1755	1078	7	23.9	92.1	1.1	NOFLOOD
11/1/2012	1818	1084	6	24.5	89.9	0.9	NOFLOOD
12/1/2012	2082	1103	18	24.4	92.5	1.4	NOFLOOD
13/1/2012	2501	1143	33	23.6	95.9	0.4	FLOOD
14/1/2012	2543	1153	9	24	94.8	0.6	FLOOD
15/1/2012	2239	1153	0	26.2	86.5	0.6	FLOOD
16/1/2012	1955	1156	3	26.4	87.4	1	NOFLOOD
17/1/2012	1863	1156	0	26.6	84.6	0.9	NOFLOOD
18/1/2012	1923	1168	12	25.7	88.8	0.8	NOFLOOD
19/1/2012	1916	1174	6	26.2	88.1	1	NOFLOOD
20/1/2012	1991	1275	101	25.4	88.3	0.7	NOFLOOD
21/1/2012	1904	1275	0	26.1	84.9	1	NOFLOOD
22/1/2012	1867	1275	0	26	84.1	1	NOFLOOD

Fig. 2: Kuala Krai Flood Dataset.

3.2. Algorithms

In investigating the Bayesian approach to flood risk prediction, three algorithms will be used, which are the Bayesian Networks [16], Naive Bayes [17], and Tree Augmented Naive Bayes [18] with oversampling technique called MOTE: Synthetic Minority Oversampling Technique (SMOTE) and without use oversampling technique (Normal). Oversampling is needed considering the imbalanced nature of flood risk classes between flood and no flood. Bayesian Networks is a probabilistic-based data modelling method that represents a variable and conditional interdependencies through a DAG (Directed Acyclic Graph). By applying Markov Chain-Rule, the joint probability distribution of the nodes in Bayesian Network can be decomposed as shown in Equation 1.

$$P_B\left(X_{1,\ \ldots,}X_n\right) = \prod_{i-1}^{n} P(X_i\,|Pa_i) \quad (1)$$

where Pa_i represents the set of parents of X_i in the networks. Fig. 3 shows a graphical model of Bayesian Networks. The class implementation of Bayesian Networks in Weka is available at http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/BayesNet.html.

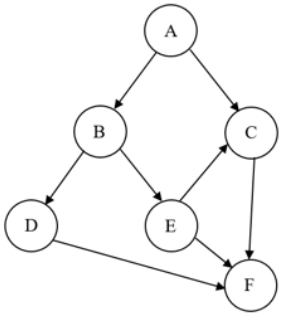


Fig. 3: Graphical Model of Bayesian Networks.

Naive Bayes could be a straightforward probabilistic classifier that calculates a collection of chances by forward the frequency and combos of values from the given datasets. The algorithm uses the Bayes theorem and assumes all the independent or non-interdependent attributes given by the value of the class variable [17]. Naive Bayes is based on a simplified assumption that attribute values are conditional on each other free of charge if given output value. In other words, given the output value, the probability of collectively observing is the product of the individual probability [19].

Naive Bayes often works much better in most complex real-world situations than expected [20] because the algorithm is based on posterior probability that combines previous experience and likelihood of event. According to the Bayes theorem, Equation 2 shows on how to calculate posterior probability,

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (2)$$

where Pa_i represents the set of parents of X_i in the networks. Fig. 4 shows a graphical model of Naïve Bayes while its class implementation in Weka is available from http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/NaiveBayes.html.

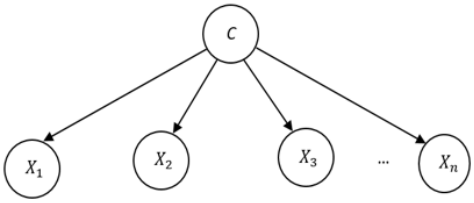


Fig. 4: Graphical Model of Naïve Bayes.

Tree Augmented Naive Bayes (TAN) is related to Naive Bayes classifier because it is a continuation of the Naive Bayes classifier. Naive Bayes classifier is obtained by learning D training data by determining the probability of each attribute X_i when given the class C variable. This is because Naive Bayes does not realistic to be applied to real data, so there is a Naive Bayes fix called Augmented Naive Bayes. Developing Augmented Naive Bayes classifier equivalents by finding a good Bayesian Network with class C variable as root [18]. Because of intensive computing, an efficient solution to finding the Bayesian Network is the ability to influence each other between variables.

Fig. 5 shows a graphical model of Tree Augmented Naïve Bayes and its implementation available from http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/net/search/global/TAN.html.

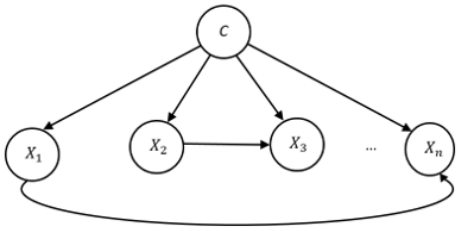


Fig. 5: Graphical Model of Tree Augmented Naïve Bayes.

3.3 Evaluation Metrics

The evaluation metrics used in the experiments are accuracy, precision, recall, and *f*-measure.

- **Accuracy.** Accuracy is total number of samples correctly classified to the total number of samples classified. The formula for calculating accuracy is shown in Equation 3, where TP is True Positive, TN is True Negative, and FN is False Negative.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (3)$$

- **Precision.** Precision the number of samples is categorized positively classed correctly divided by total samples are classified as positive samples. The formula for calculating precision is shown in Equation 4, where TP is True Positive, and FP is False Positive.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4)$$

- **Recall.** Recall is the number of samples is classified as positive divided by the total sample in the testing set positive category. The formula for calculating recall is shown in Equation 5, where TP is True Positive, and FN is False Negative.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (5)$$

- ***f*-Measure.** *f*-Measure is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. The formula for calculating f1 score is shown in Equation 6.

$$F - \text{Measure} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (6)$$

4. Results and Discussion

The purpose of this experiments is to compare the performance of Naive Bayes (NB), Tree Augmented Naïve Bayes (TAN) and Bayesian Networks (BN) algorithms with oversampling technique (SMOTE) and without the oversampling technique (Normal) when classifying the Kuala Krai flood data into risks of flood or no flood as shown in Fug. 2. Oversampling and under sampling in data analysis are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes/categories

represented). In classification, the distribution of training data can greatly influence the generalization ability of a classifier. In this experiment, the WEKA tools has been used to get the results.

In an oversampling process such as using the SMOTE technique, the first step is to determine the number of its nearest neighbors which is five. This is based on the consideration that the value of the attribute on synthetic data formed from the nearest neighbor is five. The nearest number set to five neighbors is also frequently used in experimental methods that apply SMOTE such as by [21].

As a comparison in the performing tests, sampling methods used will include random oversampling in WEKA, known as resample. This experiment evaluated training models by 10-fold cross validation technique. That means, applying the algorithm 10 times, each time 9 of the folds are used for training and 1-fold is used for testing.

Table 1 shows the results in terms of accuracy with oversampling technique (SMOTE) and without oversampling (Normal).

Table 1: Experimental Results.

Algorithm	Accuracy (%)		Precision (%)		Recall (%)		*f*-Measure (%)	
	SMOTE	Normal	SMOTE	Normal	SMOTE	Normal	SMOTE	Normal
Naive Bayes	98.290	97.920	0.984	0.990	0.983	0.979	0.983	0.983
Tree Augmented Naive Bayes	100.000	99.450	1.000	0.999	1.000	0.999	1.000	0.999
Bayesian Networks	99.880	100.000	0.999	1.000	0.999	1.000	0.999	1.000

Based on Table 1, TAN is the most efficient classifier with accuracy 100% for classifying flood risk datasets into risk of flood or no flood. However, without oversampling, BN algorithm has been the best accuracy with 100%. In terms of precision, oversampling SMOTE with TAN achieved the higher precision, which is 1.0%. Without oversampling technique, BN algorithm has the higher precision with 1.0%.

In terms of recall with SMOTE oversampling, TAN produced higher recall of 1.0%. Without oversampling, BN algorithm has the higher recall of 1.0%. Finally, the results in terms of *f*-measure, which is a combination of recall and precision values as general evaluation for imbalance data, the result shows that SMOTE oversampling with (TAN) has the best *f*-measure of 1.0%. Meanwhile, without oversampling, BN algorithm has the best *f*-measure with 1.0%.

Overall, prediction model of flood risks will perform better with oversampling such as using he SMOTE algorithm in exception of BNs because BN has better generalization capabilities even when dealing with imbalanced classes as compared to variations of naïve Bayes algorithm such as the NB and TAN.

5. Conclusions

In conclusion, this paper presented a Bayesian approach to classify flood data in Kuala Krai, Kelantan to predict flood or no flood. It also explored the used of Synthetic Minority Oversampling (SMOTE) to treat the imbalanced nature of the flood dataset. By using SMOTE it is able to handle the problem imbalance of the flood dataset with its performance value results. The result has shown that overall with by treating imbalanced using Synthetic Minority Oversampling (SMOTE), Tree Augmented Naive Bayes (TAN) has the best algorithms compare to other algorithms. This can be attributed to the actual fact that, combining all the datasets resulted in larger training set that the model may well be trained well. This research paper currently only focused on imbalanced dataset. In the future this research proposes to use dynamic Bayesian network to treat the flood dataset as time series data.

Acknowledgement

This project is sponsored by the Ministry of Education Malaysia under the Fundamental Research Grant Scheme under Vot 1609.

References

[1] Jabatan Penerangan Malaysia (2018), Geografi, available online: http://pmr.penerangan.gov.my/index.php/profil-malaysia/4-geografi.html
[2] Tan BC. (1995), Seratus Negara Asia Tenggara 1, Prisma Sdn. Bhd.
[3] Goh KC (1981), *Geografi Fizikal*. Longman, Kuala Lumpur.
[4] Hussin WNTW, Zakaria NH, Ahmad MA (2015), Knowledge Sharing and Lesson Learned from Flood Disaster: A case in Kelantan, *Journal of Information System Research and Innovation*.
[5] Wu J, Fang W, Hu, Z, Hong B (2018), Application of Bayesian Approach to Dynamic Assessment of Flood in Urban Underground Spaces. *Water*, 10(9), 1112.
[6] Rashid NAMA, Othman M (2017), Predicting Flood Risk Using Spiking Neural Network: A Framework. Dissertation, Faculty Computer Science and Information Technology, University Tun Hussein Onn Malaysia.
[7] Sikorska AE, Seibert J (2016), Value of different precipitation data for flood prediction in an alpine catchment: A Bayesian approach. *Journal of Hydrology*.
[8] Sharma A, Goyal MK (2016), Bayesian network for monthly rainfall forecast: a comparison of K2 and MCMC algorithm, *International Journal of Computers and Applications*, 38(4), 199-206.
[9] Martina MLV, Todini E, Libralon A (2005), A Bayesian decision approach to rainfall thresholds based flood warning, *Hydrology and Earth System Sciences Discussions*, 2(6), 2663-2706.
[10] Chapman P, Clinton J, Khabaza T, Reinartz T, Wirth R (2000), The CRISP-DM Process Model.
[11] Shearer C (2000), The CRISP-DM Model: The New Blueprint for Data Mining, *Journal of data warehousing*, 13-22.
[12] Singhal S, Jena M (2013), A study on WEKA tool for data preprocessing, classification and clustering, *International Journal of Innovative Technology and Exploring Engineering*, 2(6), 250-253.
[13] Wirth R, Hipp J (2000), CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pp. 29-39.
[14] InfoBanjir Portal, available online: http://infobanjirwater gov.my, 2016. 22/2/2027.
[15] Meteorologi Portal, available online: http://www.met.gov.my/, 2016. 26/11/2014.
[16] Pham DT, Ruz GA (2009), Unsupervised Training of Bayesian Networks for Data Clustering. In *Proceedings of the Royal Society A-Mathematical Physical and Engineering Sciences*, 5, 2927-2948.
[17] Patil TR, Sherekar MS (2013), Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, *International Journal of Computer Science and Applications*, 6(2), 256-261.
[18] Friedman N (1997), Bayesian Network Classifier. *Machine Learning*, 29, 131–161.
[19] Ridwan M, Suyono H, Sarosa M (2013), Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier, *Jurnal EECCIS*, 1(7). 59-64.
[20] Pattekari SA, Parveen A (2012), Prediction System for Heart Disease Using Naive Bayes, *International Journal of Advanced Computer and Mathematical Sciences*, 3(3), 290-294.
[21] Machado EL, Ladeira (2007), Dealing with Rare Cases and Avoiding Overfitting: Combining Cluster Based Oversampling and SMOTE, Department of Computer Science. Brazil.