JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

Optimization of Flood Prediction Using SVM Algorithm to Determine Flood Prone Areas

Saruni Dwiasnati[1*], Yudo Devianto[2]
[1]Fakultas Ilmu Komputer, Teknik Informatika, Universitas Mercu Buana
[2]Fakultas Ilmu Komputer, Sistem Informasi, Universitas Mercu Buana

[*]*saruni.dwiasnati@mercubuana.ac.id*

Abstract

Flooding is one thing that can slow down the economic pace in the affected area. Bandung is called the city of flowers and the city of fashion because the nickname makes Bandung a city with a variety of fashions growing in multiple places as a starting point for the buying and selling process. Not only did Bandung spawn fashions that became hits every year, but it also had many Meccas of traditional food preparation that were extraordinarily unique and interesting. Because that makes Bandung Regency currently one of the areas that are at risk of flooding. Creating a flood-prone area model can make it easier to provide information for communities in Bandung Prefecture that belong to flood-prone and non-flood-prone areas. The SVM algorithm is a technique that can be used in the case of classification and regression, which is very popular lately. SVM is in a class with Artificial Neural Networks (ANN) in terms of features and conditions of problems that can be solved, and to be able to increase its accuracy it uses what can be optimized with PSO (Particle Swarm Optimization), where the test data is used BNPB official website data, BPS Bandung District and BMKG processed. Data included in the classification criteria for flood-prone areas are rainfall intensity, water runoff, area, rainfall length, and population density. The accuracy rate generated by using the SVM algorithm is 85.71% and the generated AUC is 0.841, while the accuracy rate generated by using the PSM-based SVM algorithm is 97.62%. and AUC produced at 1,000.

Keywords: SVM Algorithm; PSO; Flood; Bandung

1. Introduction

In recent years, flooding is a disaster that often occurs in Indonesia. According to data from the National Disaster Management Agency (BNPB), flooding is the most common disaster in Indonesia, with 464 flood events per year. Floods accompanied by landslides are the sixth most common disaster in Indonesia, with 32 events per year. There are several key factors that cause flooding that are currently unavoidable, including reduction in tree cover, extreme weather conditions, and topographical conditions of the watershed. Finding out which areas are included as flood prone zones in Bandung Regency is necessary for a flood prone area analysis.

In this research, an analysis of flood prone areas is performed by a data mining approach to find out which areas belong to the flood prone areas in Bandung Regency. RapidMiner version 10 is used in the processing of data used in this research. Data mining is an interdisciplinary field based on computer science (database, artificial intelligence, machine learning, graphics and model visualization), statistics and techniques (pattern recognition, neural networks) based. Data mining is also often referred to as Knowledge Discovery in Databases (KDD), an activity that collects historical data and uses it to find regular patterns and relationship patterns in large datasets [1].

SVM is a new technique to make predictions, both in classification and in regression, which is very popular at the moment. SVM is in the same class as ANN in terms of features and problem conditions it typically solves. Both algorithms are included in the supervised learning class. Whether for scientists or practitioners passionate about machine learning, data mining or big data, many have used this technique to solve problems that seem real in everyday life and serve to predict things in the years to come. The machine learning algorithm learns from data, so it is important to properly prepare data to solve a problem [2].

On the other hand [3] it states that in many machine learning algorithms, the standard step before training is to remove the average from the data, known as zero-mean or standardization. The application of unsupervised learning was also performed to cluster historical data from transaction logs of Internet users on Microsoft and MSNBC websites using the rules of soft set theory [4], [5]. While the application of supervised learning has also been carried out for igneous rock classification [6], river water quality classification and prediction [7], image data classification in medical field [8] and attribute selection in the data set, which has a significant impact on the decision-making process [9], the prediction of oil consumption in a certain period of time [10] and the prediction of the occurrence of dengue fever [11]. Confusion matrix is a model that forms a matrix composed of true positive or positive tuples and true negative or negative tuples [12]. The analysis of the highest frequency

Accepted by editor: 26-05-2020 | Final revision: 10-09-2022 | Online publication: 25-09-2022

pattern yields a combination of items that meet the minimum support requirements determined for the formation of the association rule pattern to find the association rules that satisfy the minimum confidence of the highest frequency found [13].

Classification is a form of data analysis that creates a model to describe important classes of data [14]. ANN found a solution in the form of a local optimum, while SVM found an optimal global solution [15]. Support Vector Machines (SVM) have become a classification and regression method commonly used for linear and nonlinear problems. The advantage of the Support Vector Machines algorithm is that it is able to implement linear separation in high-dimensional non-linear data inputs and this is achieved by using the necessary kernel functions. The effectiveness of support vector machines is strongly influenced by the type of kernel functions that are chosen and implemented based on data properties [16].

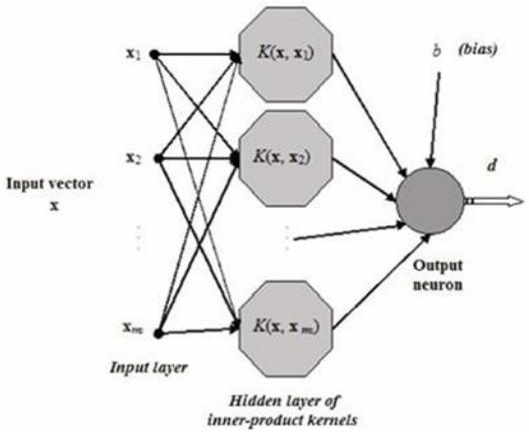

Figure 1. Architecture Support Vector Machine (SVM)

Several algorithms included in the metaheuristic method include Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Firefly Algorithm and many more. There are several problems that are commonly solved using metaheuristic methods, for example NP-hard problems, namely problems that have a non-deterministic solution, this solution runs in polynomial time. Genetic algorithms and Particle Swarm Optimization (PSO) are random search methods that are often used to find parameters of the system of nonlinear equations [17]. The PSO algorithm has been widely used to search for nonlinear model parameters. According to research [18,19] using PSO to determine the flood model parameters. While in other studies [20] the PSO algorithm was implemented as part of the artificial intelligence algorithm. Researchers have also used PSO in precipitation research models [21]. The population in this study is from a data set published by the official website of BNPB, BPS Bandung District and BMKG. The data collected ranged from 2016 to 2017 and the region that was the center of the study was Bandung Regency. Criteria for variables used in the current floodplain classification study are rainfall intensity, water runoff, area, rainfall length, and population density.

This study aims to help provide information to the people of the Regency area of Bandung no matter whether the area they live in belongs to the flood prone zone or non-flood prone area. The information can be obtained from the official BNPB page, BPS Bandung District and BMKG. The method used in this study is the Support Vector Machine (SVM) with additional feature selection, namely Particle Swarm Optimization (PSO). The variables used in this study include: precipitation, water runoff, area, number of rainy days, and population density. Precipitation itself has a definition of the amount of rainwater that collects in a shallow, non-evaporative, non-absorptive, and non-flowing place. The rainfall unit is 1 (one) millimeter, which means that in an area of one square meter in a flat place there is water up to a millimeter high or a liter of water. Then the definition of water runoff is a measure of the amount of water that can flow through a location or be accommodated in a location per unit time. Water runoff is an important part of managing a watershed.

Water delivery has a unit of volume per time or liter/second, ml/second, m/sec, liter/hour, m/hour and others. While the definition of territorial area is an area covered by a state's territorial authority, both land and sea area, where the jurisdiction of the state contained in the area applies, namely longitude and latitude. Understanding the rainy day is a unit of parameter duration as well as a weather phenomenon. While the definition of population density is the ratio of population to area, population density shows the average population per km2. Previous studies on optimizing classification algorithms use Nave Bayes algorithm and information gain function selection optimization, which increases the accuracy of analyst opinion on movie reviews by only 0.85% [4]. It is hoped that this research will use the Support Machine Vector (SVM) algorithm. and optimizing this feature selection can increase the accuracy of the results, even though the subject of investigation is different.

This investigation will compare whether the SVM method alone can increase the accuracy value with the SVM using feature selection, and also compare the AUC (area under curve) value for the two methods. The selection of the SVM algorithm is based on the number of studies using the algorithm to identify flood prone areas, and the results of the algorithm are optimized to be accurate.

2. Method

Machine learning is currently being discussed very intensely by many people, because its ability is almost the same as the process that humans perform, it can perform human skills through a machine. Machine learning, a branch of computer science related to artificial intelligence or artificial intelligence, focuses on the manufacture/development and study of a system to be able to learn from the data obtained from the intended object. According to [22], machine learning is a research area that gives computer programs the ability to learn without being explicitly programmed.

Then the record passed to the machine consists of 2 types, namely [23]:

a. Training data. Namely a set of data used to train machines to figure out and implement the algorithm used. This dataset is useful for finding good models of the data to use for data testing.
b. Data test. Testing data is a collection of data used as a reference, evaluating the models found by the machine. Does the existing model have a good accuracy score or not?

This research is experimental research. This section explains the research methods used, namely Support Vector Machine (SVM) theory and performance measurement algorithm metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Confusion Matrix.

2.1 Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD) is an activity that involves the selection of data that can be stored in various formats such as flat files, spreadsheets or relational tables, and in many places can occupy centralized or distributed data stores. Data pre-processing is a step to transform raw input data into a format suitable for further analysis. The steps to preprocess data include combining data from different sources, cleaning data to remove noise and duplicate observations, and selecting datasets and features relevant to data mining work. modeling of datasets.

The results of data mining are often integrated into a decision support system (DSS). Such integration requires a post-processing step that guarantees that only valid and useful results are combined with DSS. One of the tasks and post-processes is a visualization that allows analysts to examine data and results from data mining from different angles. Compare the two modeling results in terms of accuracy, precision, recall and AUC between the model with feature selection (particle swarm optimization) or not.

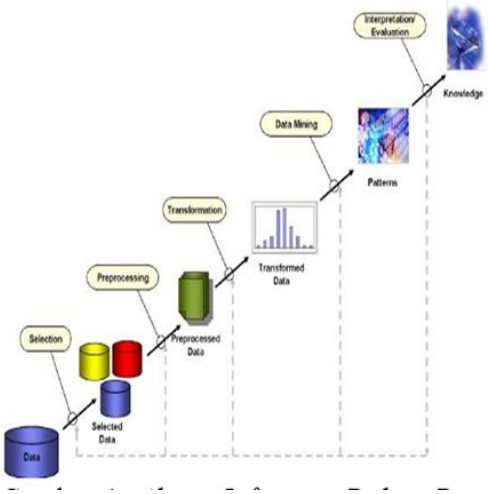

Figure 2. Flow of Information in Data Mining

2.2 Feature Selection (Particle Swarm Optimization)

The problem that often arises in special classifications and in machine learning in general is to find ways to reduce the n dimensions of the feature space F to overcome the risk of overfitting, where the overfitting itself is a value caused by the lack of training data compared to the number of ns as attributes [24] Then we need a technique to avoid the risk of overfitting, use feature selection. Particle swarm optimization is one of the methods used to increase the accuracy score of the developed machine learning model.

2.3 Framework of thinking

In a study, you need a framework for the research to proceed as planned. Furthermore, the state of mind is the flow of logic in a study. Whether a study or not can be seen in the context of thinking. It is also conceivable whether the research carried out makes sense, if viewed from the frame of thinking there is already something that is not included in reason, then it can be determined that the research will probably not be successful. The framework in this study is shown in Figure 3. In the framework of thinking, it appears that the optimization used in the flood forecast in this study was placed before the SVM algorithm.

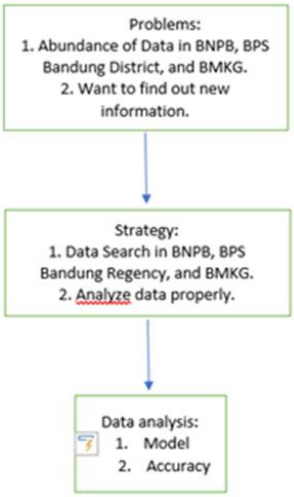

Figure 3. Framework of Thinking

a. Data Collection. Data collection for this research is included in the data understanding process. The retrieval takes place on the official website of BNPB, BPS Bandung Regency and BMKG.
b. Data Cleansing. Before the dataset is included in the model, the data cleansing process is performed, which includes filling in missing values (empty data), smoothing noisy data, identifying or eliminating outliers, and removing inconsistencies. Data must be cleaned before being processed using data mining techniques. Data obtained from real cases (real worlds) are usually not ready-to-use in the sense of faulty data. This can be due to faulty device errors, human errors or transmission errors. After the data has been found good for data processing, it is saved in the form of CSV.
c. Transformation Data that has been saved, then the data is imported to be able to create the desired model. The resulting model uses tools that are RapidMiner version 10. After the model is obtained, the accuracy and ROC results are also obtained.
d. Data Mining In this phase, a 10-fold cross-validation is performed, in which the dataset is split into 10 parts, with one of the other parts becoming the test data and the others becoming the training data. It is then fed into the Support Vector Machine (SVM) algorithm model. This is done alternately in each data item to get the best value from this model.
e. Interpretation / Evaluation After completing the data mining phase, the next step is the evaluation of the modeling results. Compare the two modeling results in terms of Accuracy, Precision, Recall and AUC between the model with Feature Selection (Particle Swarm Optimization) or those using only the SVM algorithm. The numbers are centered. The font used in the title of the image is 8pt. Figures are to be referenced in the text.

3. Result and Discussion

This chapter explains the process carried out in this study. There are several steps used in this research.

3.1 Data retrieval

Obtaining data in this study using official website data from parties involved in the subject matter of this study. Using data miner is very easy for researchers who want to get data from a web page. The services provided by Data Miner also allow the data retrievals to be exported to the desired file format.

3.2 Data cleansing

After collecting data from the website, the data cannot be directly input into the flood disaster forecasting processing. Then proceed to the data cleaning phase.

3.3 Function Selection (Particle Swarm Optimization)

For the next step, the Particle Swarm Optimization operator is used using a quality reference to continuously calculate the probable solution. This algorithm optimizes problems by moving particles/possible solutions in the problem space using specific functions on the position and velocity of the known particles. The motion of particles is affected by the best solutions of those particles, and the best solutions are generally obtained from other particles. This group of particles is called a swarm, and eventually this swarm moves toward the best solution.

3.4 Data Mining

The modeling for this study is shown in full in Figure 4. This is the overall operator used when researching the RapidMiner software

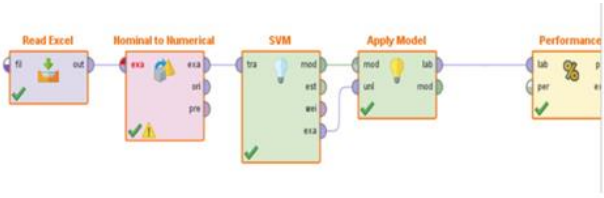

Figure 4. Modeling SVM

In the above modeling, it is known that the use of the SVM algorithm is done without optimization. Figure 5 is an SVM algorithm with optimization, i.e. particle swarm optimization.

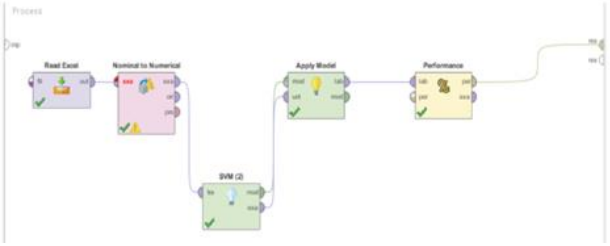

Figure 5. Modelling SVM with PSO

3.5 Interpretation / Evaluation

To evaluate this study based on accuracy and AUC (area under curve). The results of SVM with and without feature selection are compared to measure how much improvement is achieved. To obtain the value to be evaluated, accuracy SVM (Figure 6).

accuracy: 85.71%

	true Tidak Banjir	true Banjir	class precision
pred. Tidak Banjir	35	6	85.37%
pred. Banjir	0	1	100.00%
class recall	100.00%	14.29%	

Figure 6. Accuracy SVM

SVM is known to get only 85.71% from Accuracy with no feature selection:

Figure 7. AUC SVM

DOI 10.29207/joseit.v1i2.1995
Journal of Systems Engineering and Information Technology (JOSEIT) Volume. 1 No. 2 (2022) 40-46

And the AUC value obtained without feature selection is only 0.841:

accuracy: 97.62%

	true Tidak Banjir	true Banjir	class precision
pred. Tidak Banjir	34	0	100.00%
pred. Banjir	1	7	87.50%
class recall	97.14%	100.00%	

Figure 8. Accuracy With PSO

Accuracy SVM is known to get 97.62% with feature selection:

Figure 9. AUC SVM With PSO

And the AUC value obtained by trait selection is 1,000.

4. Conclusion

After researching the SVM algorithm model without feature selection and comparing it to the SVM model with feature selection, we can conclude that using "Weight by Correlation" feature selection can increase the value of accuracy and AUC. The gain obtained is very significant before the SVM model without feature selection produced only 85.71% and an AUC value of 0.841, after using feature selection was added to 97.62% for accuracy and an AUC value of 1.000. This gives a very large increase in the value difference for accuracy of 11.91 and 0.159 for AUC. From this it can be concluded that using the Weight by Correlation selection function is very good when used in the Support Vector Machine (SVM) algorithm. For further research, it may be recommended to use other weight selection functions besides the \"weight by correlation\" selection function. Because there are several operators for feature selection examples such as B. the feature selection Weight by SVM, which calculates the weighting according to the SVM method. Then the data retrieval can also be done with more numbers to get better accuracy values.

Acknowledgements

Many thanks to the Universitas Mercu Buana, Jakarta, to which the author devoted about 4 years. We would also like to thank the Faculty of Informatics, the Degree Program in Informatics Engineering and the Research Center of the Universitas Mercu Buana for funding all expenses related to our research

References

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012.
[2] Agus Ambarwari, Qadhli Jafar Adrian, Yeni Herdiyeni. 2020. Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learninguntuk Identifikasi Tanaman. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). Vol. 4 No. 1. 117–122. ISSN Media Elektronik: 2580-0760.

[3] Y. Tang and I. Sutskever, "Data normalization in the learning of restricted Boltzmann machines,"in Department of Computer Science, University of Toronto, Technical Report UTML-TR-11-2, 2011
[4] E. Sutoyo, I. T. R. Yanto, R. R. Saedudin, and T. Herawan, "A soft set-based co-occurrence for clustering web user transactions," Telkomnika (Telecommunication Comput. Electron. Control., vol. 15, no. 3, 2017.
[5] E. Sutoyo, I. T. R. Yanto, Y. Saadi, H. Chiroma, S. Hamid, and T. Herawan, "A Framework for Clustering of Web Users Transaction Based on Soft Set Theory," in Springer, 2019, pp. 307–314.
[6] I. T. R. Yanto, E. Sutoyo, A. Apriani, and O. Verdiansyah, "Fuzzy Soft Set for Rock Igneous Clasification," in 2018 International Symposium on Advanced Intelligent Informatics (SAIN), 2018, pp. 199–203.
[7] E. Sutoyo, R. R. Saedudin, I. T. R. Yanto, and A. Apriani, "Application of adaptive neuro-fuzzy inference system and chicken swarm optimization for classifying river water quality," in Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference on, 2017, pp. 118–122.
[8] M.-L. Antonie, O. R. Zaiane, and A. Coman, "Application of data mining techniques for medical image classification," in Proceedings of the Second International Conference on Multimedia Data Mining, 2001, pp. 94–101.
[9] R. R. Saedudin, E. Sutoyo, S. Kasim, H. Mahdin, and I. T. R. Yanto, "Attribute selection on student performance dataset using maximum dependency attribute," in Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference on, 2017, pp. 176–179.
[10] H. Chiroma et al., "An intelligent modeling of oil consumption," Adv. Intell. Syst. Comput., vol. 320, 2015.
[11] A. R. Muhajir, E. Sutoyo, and I. Darmawan, "Forecasting Model Penyakit Demam Berdarah Dengue Di Provinsi DKI Jakarta Menggunakan Algoritma Regresi Linier Untuk Mengetahui Kecenderungan Nilai Variabel Prediktor Terhadap Peningkatan Kasus," Fountain Informatics J., vol. 4, no. 2, pp. 33–40, Nov. 2019.
[12] N. Iriadi and N. Nuraeni, "Kajian Penerapan Metode Klasifikasi Data Mining Algoritma C4.5 Untuk Prediksi Kelayakan Kredit Pada Bank Mayapada Jakarta," J. Tek. Komput. AMIK BSI, vol. 2, 201
[13] R Riszky, M Sadikin. 2019. Data Mining Menggunakan Algoritma Apriori untuk Rekomendasi Produk bagi Pelanggan. Jurnal Teknologi dan Sistem Komputer. 103-108.
[14] RA Pangestu, S Rudiarto, D Fitrianah. 2018. Aplikasi Web berbasis Algoritma K-NEAREST NEIGHBOUR untuk Menentukan Klasifikasi Barang STUDI KASUS: PERUM PERURI. Jurnal Ilmu Teknik dan Komputer. Vol. 2 No. 1 Januari. ISSN 2548-740X E-ISSN 2621-1491.
[15] Lukman.2016. Penerapan Algoritma Support Vector Machine (SVM) dalam Pemilihan Beasiswa: STUDI KASUS SMK YAPIMDA. Faktor Exacta 9(1): 49-57, 2016 ISSN: 1979-276X.
[16] Haddi, E., Liu, X., & Shi, Y., 2013. The Role of Text Pre-processing in Sentiment Analysis. First International Conference on Information Technology and Quantitative Management, 17, 26–32. https://doi.org/10.1016/j.procs.2013.05.05
[17] Nahriyatunnur Hidayatus Solihah1), Muliadi1) , Arie Antasari Kushadiwijayanto2*). 2018. Estimasi Parameter Model Curah Hujan Menggunakan Particle Swarm Optimization (PSO): Studi Kasus Ketapang dan Melawi. Jurnal Fisika FLUX. 13-19. Volume 15, Nomor 1. ISSN : 2514-1713.
[18] Mauliana, P., 2016, Prediksi Banjir Sungai Citarum dengan Logika Fuzzy Hasil Algoritma Particle Swarm Optimization. INFORMATIKA, 3, 269-276.
[19] Ary, M., 2017. Aplikasi Prediksi Banjir Metode Fuzzy Logic, Hasil Algoritma Spade dan Algoritma PSO. In: Konferensi Nasional Ilmu Sosial & Teknologi (KNiST), 342-348.
[20] Nurmahaludin., 2013. Perancangan Algoritma Belajar Jaringan Syaraf Tiruan Menggunakan Particle Swarm Optimization (PSO). Jurnal POROS TEKNIK, 5(1),18-23.
[21] Factmawati, M., Widodo, B., and Wahyuningsih, N., 2014. Estimasi Autoregressive Integrated Average (ARIMA) Menggunakan Algoritma Particle Swarm Optimization (Studi Kasus: Peramalan Curah Hujan DAS Brangkal, Mojokerto). Surabaya: Skripsi ITS.
[22] Fikriya, Zulfa Afiq; Irawan, Mohammad Isa; Soetrisno, 2017. "Implementasi Extreme Learning Machine untuk Pengenalan Objek Citra Digital", Jurnal Sains dan Seni ITS, Vol.6, No. 1. 2337-3520
[23] Rahmansyah A., Dewi O., Andini P., Hastuti PN, Triana and Eka Suryana, Muhammad. 2016, Membandingkan Pengaruh Feature Selection Terhadap Algoritma Naïve Bayes dan Support Vector Machine. Seminar Nasional Aplikasi Teknologi Informasi (SNATi) , 2018 p. A1 - A7.
[24] Guyon, I., Weston, J., and Barnhill, S. (2002), Machine Learning, Gene Selection for Cancer Classification using Support Vector Machines, Netherland , Kluwer Academic Publishers.