Article 10

Comparison between Clustering Algorithm for Rainfall Analysis in Kelantan

Wan Nurshazelin Wan Shahidan, SitiNurasikin Abdullah
Faculty of Computer and Mathematical Sciences,
University of Technology Mara Perlis Branch, Malaysia

Abstract
Analysis of rainfall behaviour has become important in many regions because it is related to many factors such as agricultural sector, water resource management, and flood disaster and landslide occurrence. The weather in Malaysia is characterised by two monsoon regimes called as Southwest Monsoon and Northeast Monsoon. Heavy rainfall will cause water level of river to reach its maximum level that may lead to flood disaster. Floods become more serious when people start losing the life of beloved ones and property. Although natural disasters are caused by nature and there is nothing that we can do to prevent them from happening, but yet being aware of its impact is a much required process that should be looked into thoroughly. The goal of this study is to analyse the rainfall analysis in Kota Bharu, Kelantan in order to overcome any bad consequences in future. Three types of clustering algorithm were used in this study, namely K-Means clustering, density based clustering and expectation maximization (EM) clustering algorithm. Comparisons between the clustering algorithms were conducted in this study to identify which clustering algorithm is the most suitable and simple for rainfall distribution. So, in this study clustering algorithm on rainfall distribution dataset is done using WEKA 3.8 software. The results found that K-Means clustering was the suitable and simple clustering algorithm based on time taken to build model.

Keywords: *clustering algorithm, K-mean clustering, density based clustering, expectation maximization clustering, rainfall analysis*

Introduction

Analysis of rainfall behaviour has become important in many regions because it is related to many factors such as agricultural sector, water resource management, and flood disaster and landslide occurrence. The weather in Malaysia is characterised by two monsoon regimes called as Southwest Monsoon and Northeast Monsoon. Southwest Monsoon occurs during late May to September while for Northeast Monsoon it occurs during November to March (Ismail, 2015). Contribution of heavy rainfall will lead to natural disaster such as flood. According to Toriman et al., (2014) flood is also defined as the domination of high water flow in river system. There are several factors that cause flood disaster to happen such as natural phenomenon which is La Nina, heavy rainfall, drainage system failure, rapid economic development, and illegal logging activity. La Nina is the natural phenomena that will affect rainfall pattern in Malaysia to change until flood disaster occurs. Flood disaster will give bad impact to human life, damage to property, destruction of crops, problem in health condition and loss of livestock.

In 2014, the massive flood hit Malaysia and more than 200,000 Malaysians were affected by the flood while 21 people were killed due to floods and the floods in 2014 was the worst in decades. Kelantan is one of the areas in Peninsular Malaysia that often vulnerable with a big flood and it also affected all local mains road and housing areas. Kelantan was the worst state in Malaysia that was destroyed and damaged during the occurrence of flood in 2014.Although natural disasters are caused by nature and there is nothing that we can do to prevent them from happening, but yet being aware of its impact is a much required process that should be looked

into thoroughly. The main objective of this study is to find out which clustering algorithm will be the most suitable and simple for rainfall distribution and show the comparison of different clustering algorithm using WEKA 3.8 software. It is important to study the rainfall distribution since heavy rainfall will lead to natural disaster such as floods. Flood disaster may lead to property and life loss.

Data Mining Technique

Generally, data mining involves transformation of raw data into relevant patterns. Primarily, data mining is predominantly used for data pattern identification to achieve a firm understanding in data generating process as well as assisting useful predictions. In addition, there are two types of data mining namely direct and indirect data mining. In order to identify pattern and relationship in large data set, methods such as statistical model, mathematical algorithm and machine learning could be used (Phyu, 2009). Phyu also highlighted that besides collecting and managing data, data mining was also useful in analysing and predicting data.

Clustering Algorithm

Clustering, cluster analysis, segmentation analysis, taxonomy analysis or unsupervised classification is a method of creating similar groups of object, or cluster, in such a way that objects in one cluster are very similar and objects in different cluster are quite distinct (Jain et al,. 2010). The purposes of using clustering were because clustering had ability to deal with noisy data, had scalability to deal with large dataset and had ability to deal with different kind of attributes. Several clustering techniques could be applied such as k-means clustering, density based clustering and expectation–maximization (EM) algorithm.

Various Clustering Algorithm

i. K-Means Clustering

K-means clustering is a partitioning method. The function partitions data into k mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation K-means clustering introduced by J.B. McQueen in 1967 (Dhakshinamoorthy and Kalaiselvan, 2013). K-means clustering is the most common clustering that groups data with similar characteristics. The basic step of k-means is simple. First, the number of cluster k which is assumed as the centroid or centre of the clusters should be determined. Any random number objects can be taken as the initial centroid. Then, the k-means will do the three steps below until convergence.

Step 1: Determine the centroid coordinate

Step 2: Determine the distance of each object to the centroids.

Step 3: Group the object based on the minimum distance.

ii. Density Based Clustering

Density based clustering was proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes (Sharma et al., 2012). Most of scientific literature cited about density based clustering and it is one of the most common clustering algorithms. According to Raviya and Dhinoja (2013) clusters for density based clustering are identified by looking at the density of points. The advantages of density

based clustering are the shaped of clusters can find arbitrarily and can find the clusters completely surrounded by different clusters. It is robust to noise and do not need any priori k deterministic. Density based finds a number of clusters starting from the estimated density distribution of corresponding nodes. Density based clustering algorithm is an important part of clustering technique which is mainly used in scientific literature. Density is measured by the number of objects which are nearest the cluster.

iii. Expectation Maximization (EM) Clustering

Expectation maximization is a type of model based clustering method. Umale and Nilav (2014) declared that expectation maximization algorithm assigns objects to cluster according to parameters of probabilistic clusters or the current fuzzy clustering. The expectation maximization algorithm gains its name by Arthur Dempster, Nan Laird, and Donald Rubin in a classic 1977 paper. Besides that, expectation maximization algorithm is an iterative method in finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models. . Expectation maximization algorithm consists of two key stages. The two stages were showed below.

Stage 1: E (expectation) was the calculation of the cluster probabilities. In this stage, assume that we know the values of all the model parameters.

Stage 2: M (maximization) was the calculation of the model parameters. In this stage, the process was aimed to maximize the likelihood of the model given the available data.

These stages are repeated until the algorithm starts to converge.

Waikato Environment for Knowledge Analysis (WEKA)

Waikato Environment for Knowledge Analysis (WEKA) is one of the software or machine learning for data mining. WEKA is freely available for download and offers many powerful features. Besides that, WEKA provides the extensive support for the whole process of experimental data mining. WEKA implements algorithms for data pre-processing, classification, regression, clustering and association rules and also includes visualization tools. Jain et al., (2010) conclude that WEKA has made an outstanding contribution to the data mining.

Methodology

There were four processes in this study such as data collection, data pre-processing, data clustering and comparison results by clustering algorithm. The rainfall distribution data was analysed one by one in WEKA using the three clustering algorithm. This study uses the secondary rainfall distribution data from Kota Bharu, Kelantan Station. The rainfall distribution data is from January 2008 to December 2014. The data recorded is daily rainfall distribution for 365 or 366 days per year. The average rainfall distribution also consists in this study. There were three attributes from the data such as wind, radiation and humidity. After conducted k-means, density based and expectation maximization clustering algorithm thus, compared the result from the output in WEKA. The comparison based on the time taken to build the model, values of log likelihood and the number of clusters. The significant of this comparison is to see the suitable and simple clustering algorithm for rainfall distribution data.

Results Analysis

K-means clustering, density based clustering and expectation maximization clustering was conducted one by one in WEKA in order to find the results and made comparison table. The comparisons of clustering algorithm were based on the number of cluster, time taken to build model and the values of log likelihood. From WEKA we found results using the entire clustering algorithm as shown in Table 1.

Table 1: Result of Comparison of Three Clustering Algorithms

Clustering Algorithm	Number of Clusters	Clustered Instances				Time Taken to Build Model (Seconds)	Log Likelihood
		0	1	2	3		
K-means	2	26%	74%			0	-
Density Based	2	26%	74%			0.02	-6.4439
Expectation Maximization	4	39%	21%	26%	13%	0.23	-5.4829

Table 1 compares the results for three different types of clustering algorithm such as k-means, density based and expectation maximization. Density based clustering was better than the expectation maximization clustering because the time taken to build the model is 0.02 seconds which is less than expectation maximization clustering. The lower the time taken to build the model means that the model was simple. Density based clustering take less time to build a cluster but it does not better than the k-means clustering because density based clustering has high log likelihood value, if the value of log likelihood is high which is negative value than it does not make good cluster. Log likelihood measured the goodness of the clustering. K-means clustering has the lowest time taken to build the model which is 0 seconds and smaller number of clusters. As a conclusion, k-means clustering was the suitable and simple clustering algorithm in this study based on the time taken to build model. K-means clustering algorithm is the simplest clustering algorithm as compared to other clustering algorithms.

Conclusion

This study wants to obtain the most suitable and simple clustering algorithm by using rainfall distribution data in Kota Bharu, Kelantan. The selection of appropriate clustering algorithm is really important in order to gain better analysis for rainfall distribution data. Besides, the main challenges while conducting this study is to use WEKA 3.8 software in order to come out with the best clustering and able to build the clustering model which it is an important issue that had been discuss by the researchers. Results shows that the model that had been developed and conduct in WEKA 3.8 software were able to identify the best and simple clustering algorithm based on rainfall distribution data from years 2008 until 2014. K-means clustering give the lowest time taken to build this model in WEKA 3.8 software. It is simply means that k-means clustering is simple clustering algorithm. This study recommends continuous study in future to make comparison with the other clustering algorithm or make some adjustment for clustering algorithm in order get the best clusters. Besides, this study also recommends a study in future to apply using other open sources software that provides for data mining. Furthermore, this study can be used as the reference for future studies.

References

Dhakshinamoorthy, P., &Kalaiselvan, T. (2013).Crime pattern detection using data mining.*International Journal of Advanced Research in Computer Science and Applications, 1*(1), 46-51.

Ismail, N. A. (2015). A Study on social influences during annual flood occurance in Kota Bharu, Kelantan: the positive sides of disaster. *Advances in Environmental Biology*, *9*(27), 456-460.

Jain, S., Aalam, M. A., &Doja, M. N. (2010). K-means clustering using weka interface. *Computing for Nation Development.*

Phyu, T. N. (2009). Survey of classification techniques in data mining.*InternationaLMUlticonference of Engineers and Computer Scientists,*18–20.

Raviya, K. H., &Dhinoja, K. (2013).An empirical comparison of k-means and DBSCAN clustering algorithm.*Indian Journal of Research, 2*(4), 153-155.

Sharma, M., Bajpai, A., &Litoriya, R. (2012).Comparison the various clustering algorithms of weka tools.*International Journal of Emerging Technology and Advanced Engineering, 2*(5), 73-80.

The Star. (2014). *Floods in Kelantan, Terengganu Worsen*. Retrieved 14 June 2017 from http://www.thestar.com.my/news/nation/2014/12/23/floods-kelantan-terengganu/

Toriman, M. E., Abdullahi, M. G., D/iya, S. G., &Gasim, M. B. (2014). Floods in Malaysia historical reviews, causes, effects and mitigations approach. *International Journal of Interdisciplinary Research and Innovations, 2*(4), 59-65.

Umale, B., &Nilav, M. (2014).Overview of k-means and expectation maximization algorithm for document clustering.*International Journal of Computer Applications,* 5-8.