Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 204 (2022) 180–186

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry Sciences and Computer Science Innovation

River quality classification using different distances in k-nearest neighbors algorithm

Nurnadiah Zamri[a]*, Mohammad Ammar Pairan[b], Wan Nur Amira Wan Azman[c], Siti Sabariah Abas[d], Lazim Abdullah[e], Syibrah Naim[f], Zamali Tarmudi[g], Miaomiao Gao[h]

[a,b,c,d]*Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut Campus, Besut, Terengganu, Malaysia*
[e]*Management Science Research Group, Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia Management Science Research Group,*
[f]*Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom*
[g]*Universiti Technology MARA, Johor Branch, Segamat, Johor, Maaysia*
[h] *Faculty of Biology, Medicine and Health, School of Medical Siences, University of Manchester, Manchester, United Kingdom*

Abstract

The practice of river quality classification usually uses Water Quality Index (WQI) to evaluate the WQI values of the river. However, due to huge data collection on river pollution with uncertain water quality parameter values, need to a different approach to classify the river quality. One of the supervised classification algorithms known as K-Nearest Neighbors (KNN) seems to give new approach for river quality classification where each data points are classified according to the k number or the closest data points neighbors. Therefore, the purpose of this paper is to apply different distances and distance-weighted in KNN for finding the most accurate river quality classification. The accuracy results are compared with Support Vector Machine (SVM) and Decision Tree (DT) algorithms. This KNN algorithm will give a different approach in classify the river quality.

Keywords: Supervised classification; k-nearest neighbors (KNN); river quality classification; accuracy

* Corresponding author. Tel.: +609-6993027; fax: +609-6993215.
E-mail address: nadiahzamri@unisza.edu.my

1877-0509
10.1016/j.procs.2022.08.022

1. Introduction

Water safety and quality are essential to human growth and other well-being. Supplying access to safe water is one of the most valuable instruments in supporting health and decreasing poverty [1]. Most of our drinking water comes from rivers and streams [2]. River quality evaluation is one of the crucial issues of environmental water management [3]. Nowadays, developing countries have confronted crucial problems in managing water quality when struggling to increase safe water supply and sanitation [4]. Markedly, Malaysia has not escaped such depressing water quality management [5]. Water quality management become crucial in Malaysia, and this can be proved by several previous research paper [6],[7],[8]. Managing water quality needs the compilation and assessment of large water quality datasets that can be challenging to evaluate and synthesize [9]. One of the best techniques nowadays is using a supervised classification algorithm. K-Nearest Neighbor (KNN) is one of the most well-established classification algorithms in the group of supervised classification algorithm [10]. It is due to its simplicity and straightforward algorithm implementation [11].

Various authors have discussed KNN in the field of water management. For example, Fathabadi et al., [12] studied the uncertain sediment concentration in water to predict three watersheds in Iran using four different methods; Bayesian Segmented Linear Regression (BSLR), Bayesian Linear Model (BLR), Gaussian Process Regression (GPR) and KNN. Motevalli et al., [13] quantified the effects of point and non-point source nitrate pollution in groundwater using the combination of boosted regression tree and KNN method. Sapin et al., [14] studied daily streamflow and attendant stream temperature at five streams that drain into Lake Shasta using the modified KNN based stochastic simulation methods. As far as our knowledge, none of the papers has discussed KNN in the case of water quality classification for the Terengganu River.

Therefore, the objective of this paper is to offer different distances approach in the KNN algorithm for the classification of the Terengganu River. Three different distances approach includes Euclidean distance, Hamming distance, and Entropy weighted-distance. Results from each algorithm are compared with different supervised algorithms, Support Vector Machines (SVM) and Decision Tree (DT). We obtained Terengganu River classification between 81.48% for Euclidean distance, 90.12% for Hamming distance, and 99.90% for entropy estimator. These three different distances give different dimensions in classifying the river quality assessment.

2. Theoretical background

Machine Learning (ML) is a concept to make machine learn the things from the given input and present the accurate output without human intervention. ML can be distributed into three subgroups: Supervised ML, unsupervised ML and semi-supervised ML. Supervised ML uses a labeled dataset to train algorithms to predict outcomes or classify data. The training dataset can be divided into two variables; input variables usually stated as X and response variables usually stated as Y. Prediction can be retrieved from the response variables (Y) for a new dataset called testing data. This testing data can later be used to verify the model's accuracy. Two categories in the supervised learnings includes regression and classification algorithm. Classification mainly focused on predicting which class a data point is part of, usually a discrete value. The top 5 classification algorithms in ML are Logistic Regression (LR), Naïve Beyes (NB), DT, SVM, and KNN. This section introduces the basic definitions relating to KNN, DT, SVM, and accuracy. In this study, we focused on KNN, DT, and SVM.

2.1. K-nearest neighbors (KNN)

KNN is one of the nonparametric classification methods. KNN is become famous due to its most broad and easiest algorithms. It can store all the available problems or cases and classify them to the new clusters based on their similarity measure. KNN uses the concept of finding the nearest neighbor and classifying them based on their similarity. Commonly, KNN uses Euclidean distance to find the best similar data to the group.

2.2. Decision tree (DT)

DT resembles the graph that uses a branching concept to describe the result of a decision in which each attribute destruction is calculated and get the information then a new branch starts at position where that has more information/ In this wat, it creates new branches like a tree and get the information of all attributes. In DT, there are two types of values: discrete set values and continuous values. Classification trees will take a discrete set of values as the target variable. Meanwhile, regression trees will take continuous values as the target variable [15]. All the sequence of simple tests in this DT model can be grouped together to partition the data and fit a prediction model with this partition. Results of this DT can be described in graphic as a decision tree [16].

2.3. Support vector machine (SVM)

Apart from KNN and DT, SVM is also one of the supervised learning methods. Mainly, SVM is also used to focus on classification and regression. Compared to KNN and DT, SVM has its advantages. These advantages include outlier detection, efficiency in high dimensional spaces, efficiency in handling a greater number of dimensions, and the possibility of using a subset of training points in the decision function due to memory efficiency. Based on Vapnik [17], SVM can tackle design acknowledge the issue, where it can measure the learning hypothesis for factual example acknowledge under the state of a limited measure of preparing tests [17]. Thus, it can highlight the modest quantity of tests, that have worldwide streamlined and have great speculation. SVM can work when we have a set of training examples is marked to one of two categories. From these categories, SVM trains this algorithm to build a model that can assign to the new category. Besides, SVM trains examples to points in space to find the best maximize width between the gap of two categories.

2.4. Accuracy

The rate of accuracy, specificity and sensitivity is portrayed in the algorithm's performance. The overall effect of the algorithm depicts the algorithm. The way to find the best prediction in the data set is using the accuracy where the total number of correctly classified points can be divided by the total number of data points. The overall effect of the algorithm depicts the accuracy and can be calculated as

$$Accuracy = \frac{\alpha+\beta}{\beta+\alpha+\mu+\pi} \; x\,100 \tag{1}$$

where true positive (β) is when observation is positive and is predicted to be positive; false negative (π) is when the observation is positive but is predicted to be negative; true negative (α) is when the observation is negative and is predicted to be negative; false positive (μ) is when the observation is negative, but is predicted positive.

These theoretical backgrounds are being used in proposing the KNN algorithm, finding the best accuracy, and ultimately constructing the main method in this paper.

3. The proposed method

The main target of the KNN model is to predict the target class label from the most often nearest and similar neighbor based on the given query point. In other words, KNN is the model to find the majority voting from the class label considered as the *k* most similar training labels for a given query point. This section focuses on the KNN algorithm with different distances and distance-weighted for finding the most accurate river quality classification. Next subsection explains all the steps of KNN with different distances:

3.1. KNN algorithm with different distances

Assume we have a target function $g(r) = s$ that assigns a class label $s \in \{1, \cdots, t\}$ to a training example,

$$g: \mathbb{R}^{>} \rightarrow \{1, \cdots, t\} \tag{2}$$

Assuming we identified the KNN ($Q_{||} \subseteq Q$) of a query point $r^{[q]}$,

$$Q_k = \{< r^{[1]}, g(r^{[1]}) >, \cdots, < r^{[k]}, g(r^{[k]}) >\}, \tag{3}$$

we can define the KNN hypothesis as

$$h(r^{[q]}) = \arg \max_{s \in \{1,\cdots,t\}} \sum_{i=1}^{k} \delta(s, h(r^{[i]})) \tag{4}$$

Here, δ denotes the Kronecker Delta function

$$\delta(a,b) = \begin{cases} 1, & if\ a = b, \\ 0, & if\ a \neq b. \end{cases} \tag{5}$$

Or, in simple notation, if you remember the "mode" from introductory statistics classes:

$$h(r^{[t]}) = \text{mode}\ (\{g(r^{[1]}), \cdots, f(r^{[k]})\}). \tag{6}$$

KNN algorithm with Euclidean distance

Next, identify the KNN D_k is the Euclidean distance measure,

$$D(r^{[a]}, r^{[b]}) = \sqrt{\sum_{j=1}^{m} (r_j^{[a]} - r_j^{[b]})^2}, \tag{7}$$

which is a pairwise distance metric that computes the distance between two data point $r^{[a]}$ and $r^{[b]}$ over the m input features.

KNN algorithm with Hamming distance

Next, identify the KNN H_k is the Hamming distance measure,

$$H(r^{[a]}, r^{[b]}) = \sum_{j=1}^{m} (|r_j^{[a]} - r_j^{[b]}|) \tag{8}$$

which is a distance between two data point $r^{[a]}$ and $r^{[b]}$ over the m input features.

KNN algorithm with Entropy distance-weighted

A key goal of entropy minimization analysis is to determine the quantify of information in a given data set. The entropy of a probability distribution is a measure of the uncertainty of the distribution. This measure compares contents and the prior probability of data. The higher the prior estimate of the probability for an outcome to occur, the lower the information will be by observing it to occur. The entropy on a set of possible outcomes of a trial where one and only one outcome is true is defined by the summation of probability and the probability algorithm for all outcomes. In other words, the entropy is the expected value of information.

Entropy weight is a parameter that describes how much different alternatives approach one another in respect to a certain attribute [18]. Conversely, low information entropy is a sign of a highly organized system. In information theory, the entropy value can be calculated as Eq. 9:

$$e_j = -\sum_{j=1}^{M} p_{ij} \cdot \ln p_{ij} \tag{9}$$

where, e_j is the level of entropy, p_j is the probability of occurrence of event.

Further, we can modify distances metrics by adding an entropy weight to each feature dimension, which is equivalent to feature scaling. In the case of the Euclidean distance, this would look as follows:

$$E(\mathbf{r}^{[a]},\mathbf{r}^{[b]}) = \sqrt{\sum_{j=1}^{m} e_j (r_j^{[a]} - r_j^{[b]})^2}, \tag{10}$$

where $w_j = (1 - e_j)/\sum_{j=1}^{M}(1 - e_j)$; $e_j = -\sum_{j=1}^{M} p_{ij} \cdot \ln p_{ij}$

In this section, we give different distances to be combine with the KNN algorithm. In the next section, we will apply this algorithm to classify river quality.

4. Results

The datasets were used to present measurement of river pollutions in the Terengganu River, Malaysia [19]. These datasets contain 405 samples with 27 features for five different levels of river pollutions. Five different levels of river pollutions include Very Clean, Clean, Slightly Polluted, Polluted, and Highly Polluted. All data are in positive and negative integers where each value represents the characteristics of the river pollutions level that allow the learning process from supervised ML algorithms. For the experiments, the information is divided into two parts. The first part matches the features (X); includes all the river quality parameters, and the second part matches the classes (Y); includes all the river pollution hotspots. The features compose a matrix of size pxq, and the classes are a vector of size $qx1$, where p is the number of river pollution hotspots area, and q is the number of river quality parameters. Next, using the same 405 river pollution datasets, we subdivided them into two subsets: 80% training and 20% validation. The training dataset is utilized to calibrate the supervised ML algorithm, and the validation dataset is used to hyperparameter tuning and measure the accuracy. Standard Scaler's preprocessing is executed using preprocessing modules and decomposition from Python scikit-learns [20]. All KNN algorithms were executed using Python programming language and scikit-learn libraries.

The KNN algorithm depends on the neighbor points, where it uses the shortest distance to find the shortest path for neighbor points. Hence, the choice of the best distance method is important in finding the best path. Three different distances and weighted-distance algorithms are used in the KNN algorithm. They are modified version of KNN and each of them were embedded differently into the KNN method. Euclidean distance and Hamming distance were embedded in the distance measure part, where Entropy estimates was embedded in the weighting of the distance measure part. Each of them was repeated the same process starting from fit a KNN classifier for each value of K, scaling of the feature axes, choice of the distance and measure and next, weighting of the distance measure. Lastly finding the nearest k by representing the accuracy results. Once the results of each accuracy have retrieved, DT and SVM algorithms were used to compare the accuracy results. DT and SVM were choose due to the same nature which also include in the supervised group and able finding the best accuracy.

Table 1. Accuracy for each of the algorithms

Algorithms	Accuracy
KNN with Euclidean distance	81.48%
KNN with Hamming distance	90.12%
KNN with entropy estimator	99.90%
DT	97.53%
SVM	83.95%

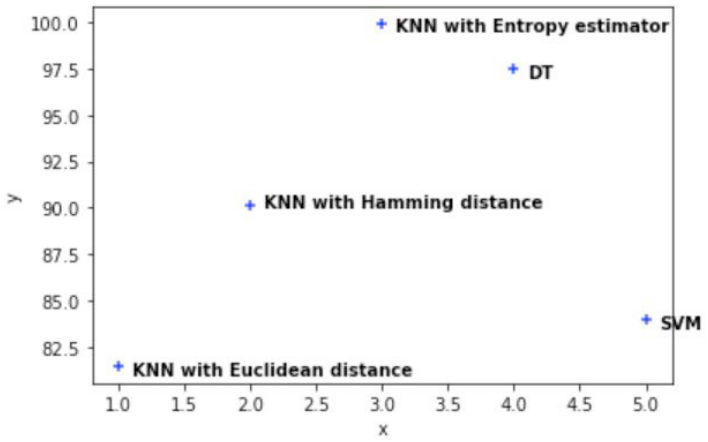

Fig. 1. Performance comparison of algorithms

Three different distances with the KNN algorithm in Section 3 are adapted to classify Terengganu River's river quality. The accuracies between three distances with KNN algorithms, DT, and SVM are illustrated in Table 1 and Fig. 1. Results show that KNN with entropy estimator achieved higher accuracy with 99.90%, DT achieved 97.53% and KNN with hamming distance achieved 90.12%. Results seem comparable between the other two methods: DT and SVM. Our finding demonstrates three different alternative ways to classify river quality. From the results, we can conclude that the performance of KNN algorithms depends significantly on the distance used. The results showed slightly different gaps between the performances of different distances. The slightly different maybe occur because they come from different family. For example, Euclidean distance measures two points of distance using a straight line in Euclidean space. Whereas Hamming distance used a metric for comparing two binary data strings. The way that Hamming distance do is it looks at the whole data and finds the most similar points with each other. Whereas the Entropy estimator is commonly used to measure value dispersion. The greater the degree of dispersion, the greater the degree of differentiation, and more information can be derived [21]. Besides, it can be merged with subjective influences and attained results can be compared for better achievement of the objectives [22].

5. Conclusion

Safe and clean water is essential in our daily lives to sustain a healthy present and future ecosystems, clean environments, and public health. WQI is a well-known method to assess and evaluate each clean river in enhancing safe water supply and sanitation. However, due to the huge data collection on river pollution with uncertain water quality, variable values need a different approach to classify the river quality. Therefore, a supervised ML, KNN algorithm with different distances, was offered as different approaches to classify Terengganu river quality. This paper offered three different distances: Euclidean, Hamming distance and entropy estimator. Besides, DT and SVM were also calculated to compare these three algorithms. Results show that the KNN with entropy estimator retrieved the highest accuracy with 99.90%. Follows by DT with 97.53% and KNN with hamming distance 90.12%. These three different distances give different dimensions in classifying the river quality assessment. Besides, in the future, hybrid or modified supervised ML, tuning the parameters is possible to obtain more highest accuracy in river quality classification.

Acknowledgements

This study was funded by the Malaysian Ministry of Higher Education (FRGS-RACER: RACER/1/2019/STG06/UNISZA//).

References

[1] World Health Organization (WHO). (2021). "Water safety and quality" https://www.who.int/teams/environment-climate-change-and-health/water-sanitation-and-health/water-safety-and-quality

[2] National Geographic. (2011) "Source", https://www.nationalgeographic.org/encyclopedia/source/.

[3] Than, N. H., Ly, C. D., and Tata, P. V. (2021) "The performance of classification and forecasting Dong Nai River water quality of sustainable water resources management using neural network techniques", *Journal of Hydrology* **596**: 126099.

[4] Ortega, D.J.P., Pérez, D.A., Américo, J.H.P., De Carvalho, S.L., and Segovia, J.A. (2016) "Development of index of resilience for surface water in watersheds**",** *J. Urban Environ. Eng*. **10**: 72-82.

[5] Ng, C. K-C., Ooi, P. A-C., Wong, W-L., and Khoo, G. (2020) "First development of the Malaysia river integrity index (MyRII) based on biological, chemical and physical multi-metrics", *Journal of Environment Management* **255**: 109829.

[6] Elfikrie, N., Ho, Y. B., Juahir, H., and Tan, E. S. S. (2020). "Occurrence of pesticides in surface water, pesticides removal efficiency in drinking water treatment plant and potential health risk to consumers in Tengi River Basin, Malaysia". *Science of the Total Environment* **712**: 136540.

[7] Koki, I. B., Low, K. H., Zain, S. M., Juahir, H., Bayero, A. S., Azid, A., and Zali, M. A. (2020). "Spatial variability in surface water quality of lakes and ex-ming ponds in Malacca, Malaysia: the geochemical influence". *Desalination and Water Treatment* **197**: 319-327.

[8] Masthurah, A., Juahir, H., and Mohd Zanuri, N. B. (2021). Case study Malaysia: Spatial water quality assessment of Juru, "Kuantan and Johor River Basins using environmetric techniques". *Journal of Survey in Fisheries Sciences* **7**(**2**): 19-40.

[9] Uddin, M. G., Nash, S., and Olbert, A. I. (2021). "A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators* **122**: 107218.

[10] Duda, R.O., Hart, P.E., and Stork, D.G. (2001) "Pattern Classification", 2nd, John Wiley & Sons, New York, NY.

[11] Gallego, A. J., Rico-Juan, J. R., Valero-Mas, J. J. (2022) "Efficient k-nearest neighbor search based on clustering and adaptive k values. *Pattern recognition* **122**: 108356.

[12] Fathabadi, A., Seyedian, S. M., and Malekian, A. (2021) "Comparison of Bayesian, k-Nearest Neighbor and Gaussian process regression methods for quantifying of suspended sediment concentration prediction", *Science of the Total Environment* (Article in Press).

[13] Motevalli, A., Naghibi, S. A., Hashemi, H., Berndtsson, R., Pradhan, B., and Gholami, V. (2019). "Inverse method using boosted regression tree and k-nearest neighbot to quantify effects of point and non-point source nitrate pollution in groundwater". *Journal of Cleaner Production* **288**: 1248-1263.

[14] Sapin, J., Rafagopalan, B., Saito, L., and Caldwell, R. J. (2017). "A K-nearest neighbor based stochastic multisite flow and stream temperature generation technique", *Environmental Modelling and Software* **91**, 87-94.

[15] Tan, P.N. (2006), "Introduction to data mining". Boston, Pearson Addison Wesley.

[16] Alpaydin E. (2010), "Introduction to machine learning". second ed. Cambridge, Mass: MIT Press. Cambridge, Mass.

[17] Cortes, C., and Vapnik, V. N. (1995). "Support-vector networks", *Machine Learning* **20** (**3**): 273-297.

[18] Liu, H., and Kong, F. (2005). "A new MADM algorithm based on fuzzy subjective and objective integrated weights*"*, *International Journal of Information and Systems Sciences*, 1, 420-427.

[19] Jabatan Pengairan dan Saliran Terengganu. (2021) "Water Pollution Statistics."

[20] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E. (2011) "Scikit-learn: machine learning in python." *Journal of Machine Learning Research* **12**: 2825–2830.

[21] Zhu, Y., Tian, D, and Yan, F. (2020). "Effectiveness of entropy weight method in decision making". *Mathematical Problems in Engineering*, **2020**: 3564832.

[22] Kumar, R., Singh, S., Bilga, P. S., Jatin, Singh J., Singh S., Scutaru, M-L., and Pruncu, C. I. (2021). "Revealing the benefits of entropy weights method for multi-objective optimization in machining operations: A critical review". *Journal of Materials Research and Technology*, **10**: 1471-1492.