Development of Nearest Neighbour Techniques for Rainfall-Runoff Model

WORD COUNT	2683	TIME SUBMITTED	21-MAY-2021 01:29PM
		PAPER ID	72396783

DEVELOPMENT OF NEARSEST NEIGHBOUR TECHIQUES FOR RAINFALL-RUNOFF MODEL

Dyah IndrianaKusumastuti
Departemen of Civil Engineering
University of Lampung (UNILA)
Jl. Sumantri Brojonegoro No. 1 Bandar Lampung
Phone/fax : 62-721-704947
e-mail : dyindria@hotmail.com

Abstract

Nearest neighbor techniques use a distance function to weight the evidence of a neighbour close to an unclassified observation more heavily than the evidence of another neighbour, that is at a greater distance from the unclassified observation. The Distance-weight nearest neighbour and regression nearest neighbour were two alternative methods developed based on Nearest Neighbour Technique. These two methods were applied as rainfall-runoff mode in a natural catchment. Of the two methods developed for the Nearest Neighbour technique, the Regression method gave the best result.

Keyword: distance-weighted, regression

1. Introduction to Nearest Neighbour

Nearest neighbour techniques are based on an assumption that nearby points are more likely to be given the same classification than distant ones. Suppose we have a set of points in state space that has been given classification. Suppose a new pattern is to be classified. The distance from the new pattern to all the old ones is computed. The new pattern is given the classification of the previously classified points that is closest to it, its nearest neighbour (Anderson, 1995).

Nearest neighbour classification simply takes the learning set {v, k} as a collection of known cases {v,k} and searches for a given pattern v to be recognized for the best match among the precedents vj. The class label k of the nearest neighbour vnearest is forwarded as the result of the classfication.

It must be noted that nearest neighbour classification needs a metric for measuring distances between the reference vectors vj (members of the learning set) and the pattern vector v to be recognized. Obviously the result dapends on the metric chosen. Normally the Euclidean metric is applled, but depending on the situation, any other metric may be applied.

$$|v - v_j|^2 = (v - v_j)^T (v - v_j) = |v|^2 - 2\, v_j^T v + |v_j|^2 \quad (1)$$

Schurmann (1996) considers Nearest neigbour cassification as a special case of multi-reference minimum distance classification, the speciality lying in the fact that the whole learning set is used as a set of reference vectors.

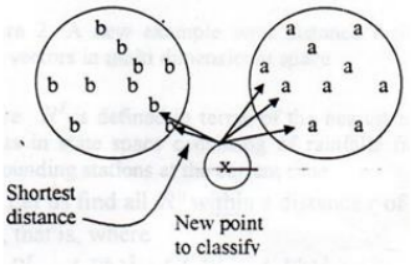

Figure 1. A simple nearest neighbour classifier. The distance from a new point to classified old ones is computed. The new point is given the classification of the nearest old point.

In terms of everyday common sense reasoning, nearest neighbour classification is what we ourselves will do when confronted with unknown situations or an unknown case. We scan our memories for the most similar case and decide in accordance with our past experience. The fact, however, that we need a distance metric for that purpose is not likely to be perceived.

The intuitively appealing concept of weghting the votes of nearby samples more heavily than those farther away from the sample

under consideration was formalized into a distance-weighted k-NN rule by Dudani (1990). He used a distance function to weight the evidence of a neighbour close to an unclassified observation more heavily than the evidence of another neighbour that is at greater distance from the unclassified observation. From his experiment at least for one arbitrary chosen example of a small training set, a lower probability of miss- classification was obtained for the distance weighted k-nearest neighbour rule than for a simple majority k-nearest neighbour. He suggested that the use of the distance- weighted k-nearest neighbour rule with training sample stes of small or moderate (miss-classification).

In a catchment with some rainfall gauges and one or more run off gauges, we have precedent rainfall and runoff records. We also have rainfall records at the current time. Using a nearest neighbour technique, based on these data. Can we predict the run off at current time? This chapter explores nearest neighbour methods for rainfall runoff modelling with real data from a natural catchment.

A distance-weighted nearest neighbour, which considers the weight as a function of distance, and a regression nearest, neighbour are investigated. In the former method, the calculation is done by ignoring the catchment area is included by multiplying it by the appropriate rainfall to generate a discharge.

2. Development of distance-weghted nearest neighbour

It is reasonable to assume that observations which are close together (according to some appropriate metric) will have the same classification. Furthermore, it is also rasonable to say that one might wish to weight the evidence of a neighbour close to an unclassified observation more heavily than the evidence of another neighbour that is at a great distance from the unclassified observation. Therefore, one would like to have a weighting function which varies with the distance between the sample and the considered neighbour in such a maner that the value decreases with the increasing sample-to-neighbour distance.

The Nearest Neighbour algorithm may be stated briefly as follows. Training set patterns are fisrt plotted in multi-dimensional feature space, and then test patterns are taken one at a time and classified according to which training set pattern is the nearest in feature space.

The discussion about distance-weighted nearest neighbour includes whether the method considers the catcment area or not. The algorithm for distance-weight nearest neighbour for both methods is the same, which is explained below, where for the method neglecting ctachment area, variable A(for area) is not considered.

Suppose we have made a state space reconstruction in dimension n+m with data vectors

$$R^{t_i} \equiv (A_1P_1^{t_i}, A_2P_2^{t_i}, \ldots, A_nP_n^{t_i}, Q_1^{t_i}, Q_2^{t_i}, \ldots, Q_m^{t_i}) \quad (2)$$

Where R^{t_i} is defined as nearest old points in state space, which consists of rainfalls from surrounding stations at previous times. We have a new vector

$$R^{T} \equiv (A_1P_1^{T}, A_2P_2^{T}, \ldots, A_nP_n^{T}, Q_1^{T}, Q_2^{T}, \ldots, Q_m^{T}) \quad (3)$$

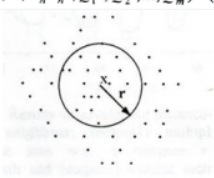

Figure 2. A new example with distance r from data vectore in mukti dimensional space.

Where R^T is defined in terms of the nearest new points in state space consisting of rainfalls from surrounding stations at the current time.

Lets us find all R^{Ti} within a distance r of R^T, that is, where

$$\left[\left(A_1{P_1}^T - A_1{P_1}^t\right)^2 + \left(A_2{P_2}^T - A_2{P_2}^t\right)^2 + \left({Q_1}^T - {Q_1}^{t_i}\right)^2 + \cdots + \left({Q_m}^T - {Q_{1m}}^{t_i}\right)^2\right] \le r^2$$

From this, we can calculate Q^T by summing up the Q^{t_i} and multiplying by weights :

$$Q^T = \sum Q^{t_1} w_1$$

Where w_r is the weight of the values of ${Q_1}^{t_i}$ according to distance,

$$w_i = 1 \text{ if } r_i = 0$$
$$w_i = 0 \text{ if } r_i = 1$$

Take a value of r_i and exclude all vectors if one of the condition below is true

$$\left(A_2{P_2}^T - A_2{P_2}^t\right)^2 > r^2$$
$$\left(A_1{P_1}^T - A_1{P_1}^t\right)^2 > r^2$$
$$(A_nP_n^T - A_nP_n^{t_i})^2 > r^2 \quad \text{or}$$
$$(Q_1^T - Q_1^{t_i}) > r^2 \quad \text{or}$$
$$\vdots$$
$$(Q_m^T - Q_m^{t_i}) > r^2$$

Exclude all vectors for which

$$\sum_{a=1,n}\sum_{i}(A_n P_n^T - A_n P_n^{t_i})^2 \geq r^2 \quad \text{or}$$
$$\sum_{a=1,m}\sum_{i}(Q_n^T - Q_n^{t_i})^2 \geq r^2 .$$

With the new set of I vectors. calculate

$$\text{where } w_i = \frac{\frac{r - r_i}{r}}{\sum(\frac{r - r_i}{r})}$$

2.1.Case Study

The study area is located in Tulang Bawang basin, Province Lampung which is discussed in author's paper : "Artifical Neural Network: An Experience in Filling in Missing Data". The monthly rainfall and average monthly flow data with 20 year period are used in setting up the rainfall-runoff model.

Based on the availability of flow gauges, Tulang Bawang rivers basin is divided into 2 large sub basin/catchment for which the outlets or downstream catchment are Pakuan Ratu and Gunung Katun. Both large sub catchments are divided into upstream and intermediate catchments. In the Pakuan Ratu basin, Sukajaya, Rantau Jangkung, Tanjung Agung, and Rantau Temiang are upstream catchments. Sumberjaya, Banjarmasin and Negeri Batin are intermediate catchments. In the Gunung Katin basin, Ogan Enam is the upstream catchments, while Kotabumi is the intermediate catchment.

2.2. Performance of the results

Before doing the experiment with the nearest neighbour method, an investigation was done to define how many years data to use for training in order to give good verification results. the length data should be not too short so that the right insight about the pattern may be obtained. Based on this investigation, it was found that training data with 15 years of data gives a good verification result. so, the first fifteen years of data, from 1974 to 1988, were used to feed information to classify the pattern.A particular input pattern was built from some vector data consisting of corresponding antecedent rainfall and runoff. the last five years of data, from 1989 to 1993, were considered as nearest new point in state space consisting of corresponding rainfall and runoff at the curret time. these data were used to verify the model, hence the output result can be obtained and compared with the observed one. there is no training output result for those data like in training with neural network, rather the verification willl only give the output results.

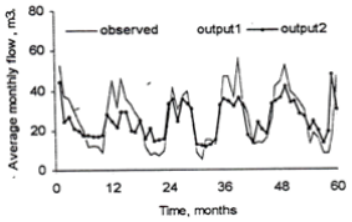

Figure 3. Results of developing distance-weighted nearest neighbour; (output1) multiplying by catchment area with a distance r= 600 mm^3 /month and (output 2) without including the area and with a distance r= 300 mm/month, for Rantau Jangkung, one of the upstream catcments.The coefficent of effeciency for the case of including and neglecting the area are 0.7765 and 0.7714 respectively.

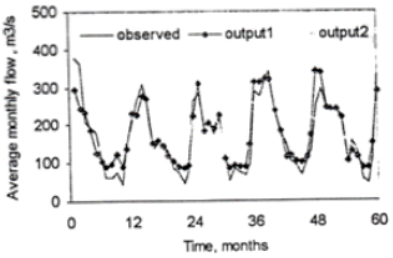

Figure 4. Result of developing the distance-weighted nearest neighbour (output 2) without including the catchments area with a distance r= 300 mm/month and (output 1) multiplying by the catchment area with distance r = 1700 m^3/month, for Pakuan Ratu, one of two bottom outlet, catchment. The coefficients of efficiency for both excluding and including the area 0.6125 and 0.9329 respectively.

For the upstream catchment, teh testing reslut show there is not much different between these two metods. But for the downstream catchment, testing result are very different, and the method including the area gives A much better result. This can be explained as follows. For upstream catchment, the input consist of rainfall only, without any flow information from the same all inputs. however, when the upstream. so, the dimension (mm/month) is thea same for all inputs. however, when dealing with the downstream catchment the inputs are not only rainfall, but also some runoff from the upstream giving contribution. the gives mixed dimensions (some in mm/month and some in m3/month) in input elements which results in creating inadequate data vectors. the second method, considering thea area, converts each of input elements to the same dimension, so that the dimensions is the data vectors are consistent.

from the results above it can be seen that distance-weighted method gives good results; however, not all verification results are stisfactory. some results show that the peak value is underestimated. this is due to not finding the right distance. given a small value of r, we cannot include some peak values that usually have a local distance, ri, bigger than the considered in computing Q^t . however, with a larger value r, many vectors are included in the calculation. the weight, Wt, is relatively small, hence, Q^t will be smaller than expected.

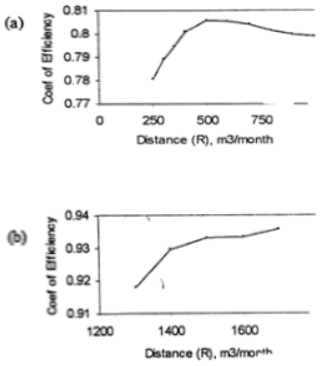

Figure 5. These graphs display the relation between distance, R, and coefficent of efficiency

for two flow gauges: (a) Sukajaya and (b) Pakuan Ratu. Both results come from the distance-weighted method considering area.

The value of the distance is not the same for all sub-catchment, depending on the location of the flow gauge and the area represented. The curve shape did not necessarily have an optimum point. these two graph are typically for all sub catchments.

3. Development of Regression Nearest Neighbour Method

Practical pattern classification usually makes a critical assumption about the statistical structure of the world. Suppose we describe an input pattern as a point in state space, that is, as a set of input element activities and we know pr are told the classification of this point. it is often assumed, consciously or unconsciously, the neareby patterns are likely to have the same classification. such an assuption about similarity suggest a rule that point close in state space are likely to have the same classification.

Based on this assumption, thea idea about a regression nearest neighbour method emerged. The nearby pattern in state space is measured by distance. Less distance means more similarity.

suppose we have made a state space recontruction in dimension n+m with data vectors:

$$R^{t_i} \equiv (A_1 P_1^{t_i}, A_2 P_2^{t_i}, \ldots, A_n P_n^{t_i}, Q_1^{t_i}, Q_2^{t_i}, \ldots Q_m^{t_i})$$

where R^T defined as the nearest old points in state space which consist of rainfalls from surronding stations at previous time. we have a new vector :

$$R^T \equiv (A_1 P_1^T, A_2 P_2^T, \ldots A_n P_n^T, Q_1^T, Q_2^T, \ldots Q_m^T)$$

where R^T defined as nearest new points in state space consist of rainfall from surrounding stations at current time.

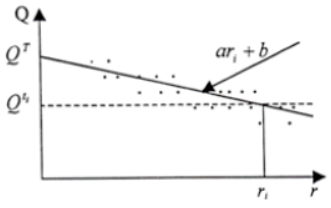

Figure 6. Plot of Qt vs distance

The flow at the current timr Q^T IS difined where the regression line touches Q-axis wich means there is no distance at all ($r_i = 0$). The distance between one single precedent vector to a new example is:

$$r = [(A_1P^T_1 - A_1P_1^{t1})^2 + (A_2P_2^T - A_2P_2^T)^2 + ..+$$

$$(Q_1^T - Q_1^{ti})^2 + ... + (Q_m^T - Q_m^{ti)2}]^{1/2}$$

For every data vector R^{ti} we know the value of Q^{ri}, that is, the precedent runoff. With the new vector R^T, We want to predict Q^T, That is, the runoff (Q^{t1}) and plot the data point, we can draw regression line. Runoff at the curren time, Q^T, Is defined when $Q^T = Q^{t1}$. The value of Q^T is computed by minimizing $\sum(ar_i + b - Q^{t1})_i^t$. For every new example we will have a different line and therefore a different equation.

3.1. Performance of the results

The regression nearest neighbor method developed for this study is tested on all subcatments ini the tuang bawang river basin. As before, the first 15 years of monthly rainfall and runoff data (1974-1988) were used as a testing set, and the remaining five years montly raintfall and runoff data (1989-1993) were used for verification. The test was performed on all eleven sub-catchments, and the performance of the method for two particular flow gauges are shown in figure 7 and 8.

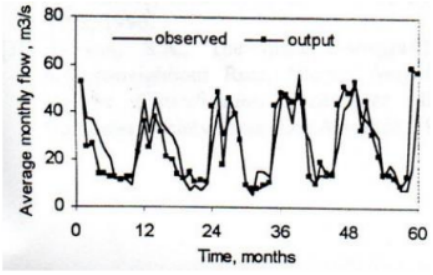

Figure 7. resuts of applying regression method of nearest to sub catchment rantau jangkung.

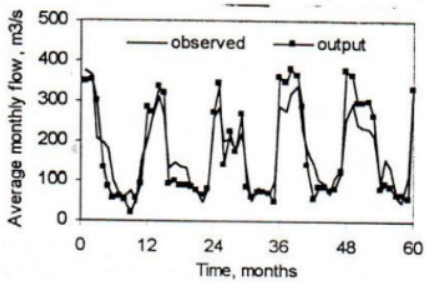

Figure 8. results off apllying regression method of neaghbour technique to sub catchment pakuan ratu.

The verification results of the regression method in an upstream catchmen, rantau jakung, and outlet catctment, pakuan Ratu, are displayed in figure 7 and 8. The coefficients of efficiency are 0.8126 and 0.9322 respectively. The verification resuls from this mrthod snows very good agreement. Moreover, this method solve the problem about catching up the peaks due to the ability of regression method to include thr small sitaces. The complete verification results from the distance-wighted, both for considering and neglecting the area, and the regression nearerst neighbor are snow in table 1. It snow that the regression method gives the best results.

Tabel 1. The comparison of the results

Flow gauge	Neglect area	Consider Area	Regression Method
Sukajaya	0.7788	0.8058	0.8063
Tanjung A	0.7643	0.7701	0.8013
Rantau J	0.7714	0.7765	0.8126
Rantau T	0.7679	0.7726	0.7889
Sumberjaya	0.7593	0.8915	0.9016
Banjarmasin	0.6776	0.6925	0.7939
Negribarin	0.7961	0.7607	0.7722
Pakuan Ratu	0.6125	0.9329	0.9322
Ogan Enam	0.7210	0.6636	0.7557
Kotabumi	0.8103	0.8400	0.8504
Gunung Katun	0.8137	0.9041	0.8567

4. conclusions

The following observation are made from the experimental results:

1. The distance –weighted nearest neigbour method considering the appropriate catchment area shows a better performace in testing; that is, it was a higher coefficient of efficiency than the method neglecting the area. This snows how important it is to have consistent dimensiond in the nearest neighhour pattern : al the inpute elemens should have the same dimension.
2. Some results from distance-wighted method snow that the peak value is underestimated. This is due to the fact that the right distance cannot be found. Given a small value of r, we cannot include some peak values which usually have a local distance, ri, bigger thn the distance r. therefore, some paeak values cannot be considered in computing QT will be smaller than expected.
3. The regression method solve the problem about catching up thr peaks due to the ability of regression method neigbour snoe the best performance.
4. Among the techniques used for this study, the regression nearest shows the best performance.

References

[1]. Anderson, J.A, *An introducsion to neural networks*, MIT Press, Cambridge, Massachussetts, 1995.

[2]. Schurmann, J., *Pattern Classification : A Unified View of Statistical and Neural Approaches*, John Wiley & Sons, Inc., New York, 1996

[3]. Dudani, S.A., The distance-weighted k-nearest-neighbour Rule, *Nearest Neighbour Pattern Classification Techniques*, IEEE Computer Society Press, Los Alamitos, 1990.

Development of Nearest Neighbour Techniques for Rainfall-Runoff Model

ORIGINALITY REPORT

8%
SIMILARITY INDEX

PRIMARY SOURCES

1	Price, Roland. "The growth and significance of hydroinformatics", River Basin Modelling for Flood Risk Mitigation, 2005. Crossref	59 words — 2%
2	L. Vefghi, D.A. Linkens. "Internal representation in neural networks used for classification of patient anaesthetic states and dosage", Computer Methods and Programs in Biomedicine, 1999 Crossref	49 words — 2%
3	E.R. Davies. "Training sets and a priori probabilities with the nearest neighbour method of pattern recognition", Pattern Recognition Letters, 1988 Crossref	40 words — 2%
4	Sudheer, G., and A. Suseelatha. "Short term load forecasting using wavelet transform combined with Holt–Winters and weighted nearest neighbor models", International Journal of Electrical Power & Energy Systems, 2015. Crossref	26 words — 1%
5	Atkinson, P.M.. "Spatially weighted supervised classification for remote sensing", International Journal of Applied Earth Observations and Geoinformation, 200410	14 words — 1%

Crossref

6 www.worldagroforestry.org
Internet
12 words — < 1%

7 es.slideshare.net
Internet
10 words — < 1%

EXCLUDE QUOTES OFF
EXCLUDE BIBLIOGRAPHY ON
EXCLUDE MATCHES OFF