Advanced Engineering Informatics 33 (2017) 89–95

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

A novel approach for precipitation forecast via improved *K*-nearest neighbor algorithm

Mingming Huang [a,*], Runsheng Lin [a], Shuai Huang [b,*], Tengfei Xing [c]

[a] Beijing Meteorological Information Center, 44 Zizhu Road, Haidian District, Beijing 100089, PR China
[b] Institute of Crustal Dynamics, China Earthquake Administration, 1 Anningzhuang Road, Haidian District, Beijing 100085, PR China
[c] Tencent Beijing, 66 Zhongguncun East Road, Beijing 100080, PR China

ARTICLE INFO

Article history:
Received 12 December 2016
Received in revised form 5 April 2017
Accepted 8 May 2017
Available online 29 May 2017

Keywords:
KNN algorithm
Improved KNN algorithm
Precipitation
Precipitation forecast

ABSTRACT

The prediction method plays crucial roles in accurate precipitation forecasts. Recently, machine learning has been widely used for forecasting precipitation, and the *K*-nearest neighbor (KNN) algorithm, one of machine learning techniques, showed good performance. In this paper, we propose an improved KNN algorithm, which offers robustness against different choices of the neighborhood size *k*, particularly in the case of the irregular class distribution of the precipitation dataset. Then, based our improved KNN algorithm, a new precipitation forecast approach is put forward. Extensive experimental results demonstrate that the effectiveness of our proposed precipitation forecast approach based on improved KNN algorithm.

1. Introduction

Precipitation plays crucial roles in climate because it not only is vital for agriculture, forestry and the energy industry but also provides stable habitats for great varieties of species [1–4]. Nevertheless, heavy rains in a short time usually result in natural disasters such as flash floods, mud-rock flows, urban waterlogging and landslides, which causes tremendous losses in lives and properties of people [5,6]. And for this reason, reliable precipitation forecast is highly important and essentially needed. Unfortunately, the precipitation forecast is a major challenge in meteorology due to formation mechanism of precipitation, not completely understood so far, involves a rather complex physics [7–9]. And currently the precipitation forecast is far from being satisfactory [10]. Accordingly, the interest in precipitation forecasts has grown significantly in recent years [11–14].

There exist two possible approaches to forecast precipitation in the literature [6,15].The first is the physically-based approach, which means that the underlying physical laws of precipitation are modeled by studying the rainfall processes. The other approach is the pattern recognition methodology based on the use of machine learning techniques to implement which means that precipitation patterns are attempted to recognize based on their features using machine learning techniques. However, it is a general notion that the physically-based approach may not be feasible, since significant variables to reason precipitation are interconnected in an extremely complicated way, and the volume of precipitation calculations requires sophisticated mathematical tools [16,17]. And the pattern recognition methodology based on the use of machine learning techniques to implement has been proven to be very effective in precipitation forecasting [18,19,17]. In recent years, much effort has been invested in developing precipitation forecasts using various machine learning techniques that yield good forecast.

Ramírez et al. [10] established a nonlinear mapping between meteorological variables and surface rainfall data to form an artificial neural network (ANN) model trained by the resilient propagation learning algorithm, which can generate specific quantitative forecasts of daily rainfall in the São Paulo State, Brazil. Hong [17] employed support vector machine (SVM) to implement rainfall forecasting in Northern Taiwan and the empirical results revealed that this approach yields well forecasting performance. Zeng et al. [20] proposed an improved *K*-nearest neighbor (KNN) algorithm to forecast precipitation, in which one optimal *k* is chosen according to the number of exiting weather event k^+ and the number of no weather event k^-, which are computed to match different weather events owing diverse probability based on the crossing verification method in searching *k*-nearest neighbor process. In addition, Chen et al. [7] developed a novel weighted *k*-nearest neighbor algorithm for forecasting precipitation in Nanjing city.

* Corresponding authors.
E-mail addresses: mmhuang1205@163.com (M. Huang), huangshuai3395@163.com (S. Huang).

http://dx.doi.org/10.1016/j.aei.2017.05.003
1474-0346/

It is worth noting that K-nearest neighbor algorithm [21], which is one of the most well-known algorithms in pattern recognition for supervised non-parametric classification, has been widely used in precipitation forecasts. However, there exists one major outstanding issue in the KNN algorithm that it is sensitive to the choice of the neighborhood size k [22]. And the classification performance of many KNN-based methods will degrade dramatically due to the problem above, particularly in the small sample size cases and the data with an uneven distribution [23,24]. Unfortunately, historical rainfall data is imbalanced, which means that the number of training samples in some classes is much larger than that of the others. For example, the number of torrential rainfall events, as compared to the number of dry days, is significantly few because the torrential rainfall events rarely occurred in a year. As a consequence, although the KNN algorithm has been proven to be very effective in making forecast of rainfall, the forecast performance is often limited by the influence of the sensitiveness to the selection of the neighborhood k when the precipitation data with an uneven distribution. To address the issue of sensitivity of different choices of k, there exists one kind of main works, namely distance-based vote weighting schemes for the KNN classifier [25]. Dudani [26] proposed a weighted voting method named the distance-weighted k-nearest neighbor (WKNN) rule, which is the first distance-based vote weighting schemes. In this approach, the farthest neighbor is weighted with 0, the closest with 1 and the others are scaled between by a linear mapping. Gou et al. [27] presented a dual weighted k-nearest neighbor (DWKNN) rule that extended the linear mapping of Dudani, in which the closest and the farthest neighbors are weighted the same way as the linear mapping, but those between them are assigned smaller values.

In this paper, we propose an improved KNN algorithm, which offers robustness against different choices of the neighborhood size k, particularly in the case of the irregular class distribution of the precipitation dataset. Then, based our improved KNN algorithm, a new precipitation forecasting model is put forward. Extensive experimental results show that the proposed precipitation forecasting paradigm achieves much better forecast performance.

The reminder of this paper is organized as follows. In Section 2, we give a brief overview of many algorithms, including KNN, WKNN and DWKNN. Section 3 presents details of our improved KNN algorithm. The construction of our proposed precipitation forecasting model is described in Section 4. Experiment results are discussed in Section 5. Finally, we conclude this paper in Section 6.

2. Related work

Before presenting in detail our improved KNN algorithm, we briefly review of KNN, WKNN and DWKNN that form the basis for our work.

2.1. KNN

The k-nearest neighbor algorithm is a powerful nonparametric classifier which assigns an unclassified pattern to the class represented by a majority of its k nearest neighbors. In the general classification problem, let $T = \{\boldsymbol{x}_n \in R^d\}_{n=1}^N$ denote a training set with M classes consisting of N training samples in d-dimensional feature space, and the class label of one sample $\boldsymbol{x}_n$ is c_n. Given a query point $\boldsymbol{x}$, the KNN rule is carried out as follows.

- Find k nearest neighbors from the set T for the unknown query point $\boldsymbol{x}$, and let $\overline{T} = \{(x_i^{NN}, c_i^{NN})\}_{i=1}^k$ indicate the set of k nearest neighbors for $\boldsymbol{x}$. The distance between $\boldsymbol{x}$ and the neighbor x_i^{NN} is measured by the Euclidean distance metric

$$d(x, x_i^{NN}) = \sqrt{(x - x_i^{NN})^{\mathrm{T}}(x - x_i^{NN})} \tag{1}$$

- The class label of the query point $\boldsymbol{x}$ is predicted by the majority voting of its neighbors

$$c' = \arg\max_c \sum_{(x_i^{NN}, c_i^{NN}) \in \overline{T}} \delta(c = c_i^{NN}) \tag{2}$$

where c is a class label and c_i^{NN} denotes the class label for the i-th nearest neighbor among its k nearest neighbors. $\delta(c = c_i^{NN})$, an indicator function, takes a value of one if the class c_i^{NN} of the neighbor x_i^{NN} is the same as the class c and zero otherwise.

2.2. WKNN and DWKNN

To overcome the negative effect of the neighborhood k, the WKNN assigning weights to the neighbors according to their distance from the unclassified sample. Formally, the weighted function of WKNN is represented as

$$\omega_i = \begin{cases} \frac{d_k - d_i}{d_k - d_1}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases} \qquad d_i = \sqrt{(x - x_i^{NN})^T (x - x_i^{NN})} \tag{3}$$

where x_i^{NN} denotes the i-th nearest neighbor among its k nearest neighbors $\overline{T} = \{(x_i^{NN}, c_i^{NN})\}_{i=1}^k$ sorted in an increasing order according to their corresponding Euclidean distance to the query point $\boldsymbol{x}$.

Accordingly, the class of the query object predicted by the majority weighted voting defined as

$$c' = \arg\max_c \sum_{(x_i^{NN}, c_i^{NN}) \in \overline{T}} \omega_i \cdot \delta(c = c_i^{NN}) \tag{4}$$

To further overcome the negative effect of the neighborhood k, the dual distance-weighted function of DWKNN extends the linear mapping of WKNN. The dual distance-weighted function of DWKNN can be formulated as

$$\omega_i = \begin{cases} \frac{d_k - d_i}{d_k - d_1} \cdot \frac{d_k + d_1}{d_k + d_i}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases} \tag{5}$$

3. Our improved KNN algorithm

Although WKNN and DWKNN algorithms perform well in comparisons with the traditional KNN approach, the sensitivity of the classification performance to the choices of the neighborhood size k still exits, especially in the data imbalance situations. And we noticed that the exponential of some distance, which is chosen as the weighting scheme, exhibits better classification accuracy and lower variance [28]. Inspired by the effectiveness of the exponential of some distance for classification, we believe that it should be a better candidate which is chosen as the weighting scheme. In this paper, we design an improved KNN algorithm in order to further mainly conquer the influence of the neighborhood k in the case of the precipitation dataset with an uneven distribution. Suppose a training set $T = \left\{\boldsymbol{x}_n \in R^d\right\}_{n=1}^N$ with M classes, where N is the sample numbers of T, and d is the feature dimension. In our proposed improved KNN algorithm, the class label of a query point $\boldsymbol{x}$ is yielded by the following steps.