Data & Knowledge Engineering 103 (2016) 44–59

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

A novel and powerful hybrid classifier method: Development and testing of heuristic k-nn algorithm with fuzzy distance metric

Hamdi Tolga Kahraman

Karadeniz Technical University, Software Engineering Department, Turkey

ARTICLE INFO

Article history:
Received 23 August 2012
Received in revised form 13 February 2016
Accepted 17 February 2016
Available online 2 March 2016

Keywords:
Classification
Data mining methods and algorithms
Artificial bee colony optimization
Fuzzy distance metric
Heuristic weight-tuning method
Hybrid k-nearest neighbor classifier

ABSTRACT

Weight-tuning methods and distance metrics have a significant impact on the k-nearest neighbor-based classification. A major challenge is the issue of how to explore the optimal weight values of the features and how to measure distances between the neighbors affecting the classification accuracy of the k-nn. In this paper, a powerful similarity measurement method, which is called the fuzzy distance metric, is explained and extended to measure the distances between the test and training observations. Depending on the fuzzy metric, similarity arrays can be produced more efficiently than the classic and other weighted distance measurements. Finally, the weighting methods are combined with the fuzzy metric-based similarity measurement and the k-nearest neighbor algorithm to increase the classification accuracy of the proposed algorithm. The effectiveness of the proposed approaches is proven by comparing their performances with the performances of the classic and the population-based heuristic methods on the well-known, real-world classification problems obtained from the UCI machine-learning benchmark repository. The experimental results show that the proposed hybrid algorithms significantly explore more optimal weight vectors significantly and provide more accurate classification results than the powerful and well-known instance-based intuitive and heuristic classification algorithms and classic approaches over real datasets.

1. Introduction

There are many research and development efforts in the fields of population-based intuitive and heuristic searching and optimization techniques [1–7], classification algorithms [8–10] and similarity measurements [11–17]. Genetic algorithm (GA) is one of the most popular optimization and searching algorithms and has also been used to develop various hybrid algorithms and provide better solutions [4–5,9]. One of the modern heuristic optimization techniques is the artificial bee colony (ABC) algorithm [18]. It is a swarm based meta-heuristic algorithm for optimizing numerical problems [1,2]. It was inspired by the intelligent foraging behavior of honey bee swarm. There are numerous optimization studies reporting success of the GA and ABC-based hybrid approaches [4–5,19–20].

While the GA and ABC-based methods are the most popular optimization and searching algorithms, the k-nearest neighbor is one of the most commonly used classification algorithms [4,21]. In the classic k-nn-based classification algorithms, in order to find out the 'k'-nearest neighbor observations (samples) for a test observation, all the similarities/distances between the training and test observations must be calculated. Each of the sample observations is equally important to measure the distances [16–17,21–22]. In other words, in the classic k-nn algorithm, there is no weight difference among the samples. It assigns equal weight values for each neighbor. This mostly causes misclassifications when there are some irrelevant data in the training observations set. One of the solutions is

E-mail address: htolgakahraman@yahoo.com.

http://dx.doi.org/10.1016/j.datak.2016.02.002

to weight the effect of the training samples (each one of the k-neighbors) [4,10,15,21–22]. It is also known as the "sample/observation weighting". In the distance metrics, the feature values of the observations or the samples are used to measure the similarities among them. Another disadvantage of the classic approaches is that they consider all the features equally to belong to observations in the similarity measurements [4,23]. In other words, the features have equal weight values in distance metrics. The effects (weight values) of the features on the class of an observation should be weighting [4,23]. It is also known as the "feature weighting" [4,10,24]. While the lower weights are assigned to the less relevant features, the higher weights are assigned to more relevant features. The weighting of features also optimizes the classification results [10,17,21–22]. The weighting of the observation features/attributes is not new. Although there are many successful optimization methods to weigh the observation features and create a vector of real-valued weights, the number of the misclassified observations may still be high. In other words, weighing the observation features has a significant impact on the similarity measurements and the k-nn-based classification, but it is not enough for the classification of the new observations in a perfect form. Additionally, the results of the similarity measurements or the selection of the distance metrics is an important factor in k-nn based classification process [4,15,17]. In other words, an effective distance metrics or similarity measurement method and a weight-tuning method should be developed and/or combined with the k-nn algorithm to produce perfect classification results.

The main purpose of this paper is to increase the classification accuracy and the stability of nearest neighbor-based and well-known classic and intuitive classification methods. For this purpose, the paper offers solutions to the challenges of the effective and powerful weighing and similarity measurements. A major challenge is the issue of how to explore the optimal weight values of the features/attributes that belong to the observations or samples in a classification problem or a dataset. Another challenge is how to measure the distances or similarities among observations more efficiently. Firstly, a novel and efficient weight-tuning procedure by applying the ABC-based heuristic searching approach to adjust the weights of the features and an overview of the previous genetic-based weight-tuning procedures is introduced. Secondly, a novel and powerful similarity measurement method, which is called the fuzzy distance metrics, is extended to measure the similarities between the test and the training observations. It has combined the weight-tuning procedure and the automatic indexing method with the fuzzy metrics. Depending on the modified fuzzy distance metrics, the similarities are measured more accurately than the classic and the other weighted distance metrics. The distance arrays or the similarities are used to predict the class of the test observations in the classification process. Thirdly, a fuzzy distance metric-based heuristic classification approach is introduced. The fuzzy logic-based distance metric and population-based search procedures are successfully combined with the k-nn-based classification algorithm in this stage. The effectiveness of the proposed methods is proven by comparing their performances with the performances of the classic and population-based heuristic methods on the well-known, real-world classification problems obtained from the UCI machine-learning benchmark repository. Experimental results show that the proposed *Heuristic and Intuitive Classifier with Fuzzy Metric* explores optimum weight vectors for the *n*-dimensional space of the features and the different k-values. Moreover, the proposed algorithm improves the classification results of the classic and intuitive k-nearest neighbor-based methods over the real datasets.

The paper is organized in the following order. The related study is given in Section 2. In Section 3, a simple presentation of the features/attributes of a classification problem is mathematically given in the form of similarity array. The proposed fuzzy logic distance metric-based heuristic classifier approach is introduced in Section 4. There are many things given in this Section such as the fuzzy distance metric-based heuristic classifier process, the main components of heuristic classifier algorithm, the artificial bee colony-based weight-tuning procedure, the classic and weighted distance metrics, application of the fuzzy distance metrics to weighted feature set, a fitness function (the k-nn classifier with fuzzy distance metric). Besides, there are many things given in Section 5 such as the parameters of genetic algorithm, artificial bee colony algorithm and the k-nn classifier and the results of the experimental studies. The application areas, attributes and characteristics of the data for experimental studies are introduced in this section. Empirical comparison of the developed and adopted classification approaches is also presented in Section 4. Finally, the discussion of the proposed approach and the conclusions of the study are given in Section 6.

2. The related study

The k-nn is a widely used classification and estimation algorithm in pattern recognition and data mining. The studies about the k-nn, which is more of a nonparametric approach and instance-based learning (lazy learning) method, focus on the classification problems but, some researchers have used this algorithm to produce real valued decisions [4,25–27]. Various hybrid models have been developed in order to increase the performance and the accuracy of the k-nn models and applications. Aksoy et al. [4], developed an intuitive k-nn estimator (IKE) model asphalt mixtures data. IKE algorithm explored the impacts of the features on the target values of the problem and gained the ability of computing and estimating for real-valued target outputs. Instead of the *majority vote* method, which is used to determine the class of a test sample in the k-nn classification process, the *root average sum of the squares* method was used to predict the numerical target values.

The similarity measurements, distance metrics, weighting procedures, determining the best of the k-value are commonly used solutions in hybrid k-nn approaches. The GA-based k-nn method is one of the most popular and widely used instance-based hybrid classification algorithms [4,10,21,28]. Suguna and Thanushkodi [21], introduced a version of hybrid k-nn method using GA-based weighting method to improve its classification performance. According to this approach, initially the reduced feature set is constructed from the training samples using the rough set-based bee colony optimization, and then by using the GA, the k-number of the samples is chosen for similarity measurements between the test and the training samples. This intuitive approach is especially useful in text classification tasks; however, the dimensionality reduction for the classification problems in *n*-dimensional space of samples is not a new method [29]. It specifically decreases the response time of the classic k-nn approach, rather than improving the classification accuracy of it. Weight-tuning is one of the most widely used and practical methods to increase