Expert Systems With Applications 88 (2017) 258–269

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Evolutionary-modified fuzzy nearest-neighbor rule for pattern classification

Peyman Hosseinzadeh Kassani, Andrew Beng Jin Teoh, Euntai Kim*

School of Electrical and Electronic Engineering, Yonsei University, 120-749, Seoul, South Korea

ARTICLE INFO

Article history:
Received 18 May 2016
Revised 10 July 2017
Accepted 11 July 2017
Available online 12 July 2017

Keywords:
Pattern classification
Fuzzy nearest-neighbor rule
Multi-objective genetic algorithm
Graphical-processing unit

ABSTRACT

This paper presents an improved version of the well-established k nearest neighbor (k-NN) and fuzzy NN (FNN), termed the multi-objective genetic-algorithm-modified FNN (MOGA-MFNN). The MFNN design problem is converted into a multi-modal objective maximization problem constrained by four objective functions. Thereafter, the associated parameter set of the MFNN and the feature attributes can be determined optimally and automatically via the non-dominated sorting genetic algorithm II. We introduce two new objective functions termed the Margin-I and Margin-II, which are used to improve the generalization capability of the MFNN for the unknown data, along with two existing performance functions: the geometric mean and the area under the receiver-operated characteristic curve for the training accuracy. Moreover, we proposed a novel data-dependent weight-assignment technique for local class membership functions of the MFNN. The technique enables the MFNN to determine its local neighbors adaptively through the MOGA algorithm. To expedite the classification, the MOGA-MFNN is implemented on a graphical processing unit (GPU), which significantly increases the computation speed. Furthermore, the local class-membership function of the MFNN can be computed in advance, rather than delaying it to the classification stage. This again can improve the classification speed. The MOGA-MFNN is evaluated on 20 datasets obtained from the repository of the University of California, Irvine (UCI). The experiments with rigorous statistical significance tests demonstrate that the proposed method performs competitively with the existing methods.

1. Introduction

Pattern classification is one of the fundamental components in intelligence systems, which is used to predict unknown samples using a learned classifier. Some of the popular classifiers include support vector machine (SVM) (Cortes and Vapnik, 1995), neural networks (Haykin, 1999), decision trees (Quinlan, 1987), k-nearest neighbor (k-NN) (Cover & Hart, 1967), and linear discriminant analysis (LDA) (Fisher, 1936).

The k-NN is a type of instance-based learning, or lazy learning, wherein the function is only approximated regionally and all the computation is delayed until the classification stage. For classification using the k-NN, the output is a class label. A query data class is determined by a majority vote of its neighbors with the query data assigned to the class most common among its k nearest neighbors. k is discrete and often small. If $k=1$, the query data is assigned to the class of that single nearest neighbor. It has been proved that the error of the k-NN when $k=1$ is asymptotically bounded by twice the Bayes error rate (Duda and Hart, 1973). This implies that 1-NN has an asymptotic error rate that is less than twice that of the Bayes classifier. The complexity of the k-NN is $O(HN)$ for N training data points with H dimensions. The complexity increases with the increase in the value of k. Different values of k can alter the classification result, and thus, choosing k is a non-trivial task.

The k-NN has gained popularity since 1967 (Cover and Hart) and it is still one of the highly favorable methods for pattern classification thus far (Chen, et al., 2011; Cheng & Hoang, 2015; Derrac, et al., 2014; Kundu & Mitra, 2015; Xia, Mita, & Shibata, 2015). In fact, the k-NN is considered one of the top ten algorithms according to the mining community (Chapman and Hall, 2009).

Section 2 discusses the related studies. Section 3 introduces the fuzzy nearest-neighbor (FNN) algorithm and the multi-objective genetic algorithm (MOGA). Section 4 elaborates the proposed multi-objective genetic-algorithm-modified FNN (MOGA-MFNN). Section 5 presents the experiments of the proposed method by considering the interaction of four objective functions, comparison with competing methods, and the associated statistical significance tests. The conclusions of this study are given in Section 6.

* Corresponding author.
E-mail addresses: bjteoh@yonsei.ac.kr (A.B.J. Teoh), etkim@yonsei.ac.kr (E. Kim).

http://dx.doi.org/10.1016/j.eswa.2017.07.013
0957-4174/

2. Related studies

Despite the simplicity and favorable accuracy performance of the conventional k-NN, equal weights are assigned to all the selected neighbors regardless of their distances from the test point. Because the *k*-NN finds the nearest neighbors based on the distances, assigning appropriate weights to the neighbors would likely enhance the performance of the *k*-NN. Accordingly, the fuzzy version of the *k*-NN (FNN) (Keller, Gray, & Givens, 1985) was proposed. This makes the FNN a competitive classifier compared to other well-known classifiers (Kassani, Hyun, & Kim, 2015; Liu, Chen, Yang, Li, & Liu, 2012; Mahmoud, El Hadad, Mousa, & Hassanien, 2015; Ramentol et al., 2015; Tomasev, Radovanović, Mladenić, & Ivanović, 2014). The fuzzification of the *k*-NN is indeed a long-standing research topic in pattern recognition. Interested readers should refer to a comprehensive survey paper for numerous descriptions on different types of FNNs (Derrac et al., 2014).

In this section, we focus on several different types of FNNs relevant to our proposed method. They can be categorized into three types: FNN design via optimization, weight-assignment strategy for FNN, and fast FNN. Several factors play an important role in the performance of the FNN such as the data form used for the distance matrix in the FNN. The FNN contains two dominating parameters: the neighboring size k and the fuzzy strength m. The parameter-searching problem in the FNN can be formulated as a multi-modal optimization problem, and thus, can be solved using meta-heuristic techniques. Cheng and Hoang (2015) and Chen et al. (2011) conducted studies pertaining to the same. In particular, the latter uses a parallel single-objective particle swarm optimization (PSO) algorithm for parameter tuning to classify credit-scoring data.

Yu, Backer, and Scheunders (2002) applied a genetic-algorithm-based feature-selection method for a set of NN classifiers on high-dimensional data. In this work, a composite architecture of the *k*-NN is used in the form of a decision-tree structure. The use of the FNN in this composite architecture improves the results compared to the conventional *k*-NN. The experiments are conducted on airborne visible/infrared imaging spectrometer (AVIRIS) data.

Many studies have attempted to improve the FNN by designing a better weight-assignment scheme for the local class membership function. Han and Kim (1999) proposed a variant of the FNN, termed the variance-weighted fuzzy *k*-NN (VWk-NN). The crux of this method is that the weight of the neighbors is assigned based on the standard deviation of their class-membership values that reflect the value of a discriminant function. Unlike the conventional *k*-NN and FNN, the VWk-NN yields better accuracy performance under various conditions. However, the performance of the VWk-NN may degrade when the number of neighbors increases. This could be because of using a small number of samples compared to the dimension of the feature space.

Derrac, Chiclana, Garcia, and Herrera (2015) introduced a new approach based on the interval-valued fuzzy sets, namely the interval-valued *k*-NN (IV-kNN). The interval values facilitate the membership of the instances and the computation of the votes in a more flexible way than that in the FNN, thus improving its adaptability to different supervised learning problems. This approach reduces the sensitivity of the FNN to the k and m parameters. The change in the local class-membership function of the FNN can be seen as a weight-assignment method to the FNN. An experimental study, wherein nonparametric statistical procedures were applied, was conducted and the superiority of the IV-kNN over the *k*-NN, FNN, and other fuzzy NN classifiers was demonstrated.

Rhee and Hwang (2003) employed a fuzzy logic type-2 to the FNN wherein the membership values for each pattern vector were estimated via interval type-2 fuzzy membership functions by assigning uncertainties to the type-1 membership functions where the FNN is based upon. This alteration can be perceived as a new weight-assignment strategy. Thus, the classification result was found to be more reasonable compared to the FNN.

Ezghari, Zahi, and Zenkouar (2017) proposed a new method called the fuzzy-analogy-based classification (FABC). In their work, the fuzzy linguistic modeling and approximate reasoning are exploited to make the FABC tolerable to the imprecision. This study claims that the FNN fails to handle the imprecision in the feature measurement and the uncertainty due to the choice of the distance-measurement value and the number of neighbors in the decision rule. Hence, to solve this problem, the local class membership function of the FNN is replaced by fuzzy linguistic variables.

In the *k*-NN, the distances between the query data and all the training data should be computed, followed by a partial sorting to find the k NNs; thus, this procedure is computationally expensive. A rapid-sorting function such as the insertion sort (Oren, Shechtman, & Irani, 2008) can be applied to increase the computation speed. Garcia, Debreuve, and Barlaud (2008) demonstrated that the computation speed in the GPU-based *k*-NN could be improved to one or two orders of magnitude with compute unified device architecture (CUDA) and insertion sort compared to the CPU-based *k*-NN. Kassani et al. (2015) employed the same strategy and proposed a modified version of the FNN with a single-objective function, which was applied to a traffic-sign detection problem.

As the calculation process of the FNN membership is delayed until the classification stage, the run-time complexity increases with respect to the training-data size. Taneja, Gupta, Aggarwal, and Jindal (2015) outlined a modified-fuzzy-based *k*-NN (MFZ-kNN) to address this problem. In the MFZ-kNN, the fuzzy clusters are sought at the preprocessing step and are given the centroid of the clusters, and the membership between the centroids and the training data points was then computed. The query data were then classified using the pre-computed membership matrix and *k*-NN. This significantly reduces the time complexity. The results show that the MFZ-kNN outperforms the *k*-NN and FNN both in terms of the accuracy and time complexity.

2.1. Motivation and contributions

The above studies reveal that the performance of the FNNs can be improved by formulating the FNN-design problem as a mathematical optimization problem or by re-designing the weight-assignment strategy while its limitation such as the run-time complexity can be rectified. In this study, we combine the three aspects aforementioned to develop a better FNN, termed the MOGA-MFNN. The existing studies largely neglected *hard-to-learn* data i.e., query data that reside on the border of the classes in the fuzzy class membership space, which could jeopardize the generalization performance of the FNNs. To address the limitation of the conventional FNNs wherein the effect of the distances between the local neighbors was neglected, we outline a new weight-assignment method for the local class-membership function, through which weights can be assigned to the local neighbors based on their distances.

In summary, the contributions of this study are as follows.

1. We propose two new objective functions to optimize the MFNN along with two existing performance functions (G-mean and AUC). The two new objective functions, namely Margin-I and Margin-II, are developed based on the notion of margin in the fuzzy membership space. The former is used to maximize the distance between the degrees of the global fuzzy class memberships for all the samples whereas the latter is used to maximize the number of samples that are outside a specific range in the fuzzy class membership space. Both the new functions are employed to address the generalization capability of the