Adaptive Learning-Based k-Nearest Neighbor Classifiers With Resilience to Class Imbalance

Sankha Subhra Mullick, Shounak Datta, and Swagatam Das, *Senior Member, IEEE*

***Abstract*— The classification accuracy of a k-nearest neighbor (kNN) classifier is largely dependent on the choice of the number of nearest neighbors denoted by k. However, given a data set, it is a tedious task to optimize the performance of kNN by tuning k. Moreover, the performance of kNN degrades in the presence of class imbalance, a situation characterized by disparate representation from different classes. We aim to address both the issues in this paper and propose a variant of kNN called the Adaptive kNN (Ada-kNN). The Ada-kNN classifier uses the density and distribution of the neighborhood of a test point and learns a suitable point-specific k for it with the help of artificial neural networks. We further improve our proposal by replacing the neural network with a heuristic learning method guided by an indicator of the local density of a test point and using information about its neighboring training points. The proposed heuristic learning algorithm preserves the simplicity of kNN without incurring serious computational burden. We call this method Ada-kNN2. Ada-kNN and Ada-kNN2 perform very competitive when compared with kNN, five of kNN's state-of-the-art variants, and other popular classifiers. Furthermore, we propose a class-based global weighting scheme (Global Imbalance Handling Scheme or GIHS) to compensate for the effect of class imbalance. We perform extensive experiments on a wide variety of data sets to establish the improvement shown by Ada-kNN and Ada-kNN2 using the proposed GIHS, when compared with kNN, and its 12 variants specifically tailored for imbalanced classification.**

***Index Terms*— Heuristic learning, imbalanced classification, k-nearest neighbor (kNN), parameter adaptation, supervised learning.**

I. Introduction

A. Overview

Classification can be posed as the task of predicting a many-to-one mapping $g(.)$ from a set X of D-dimensional data points (thus $X \subset \mathbb{R}^D$, assuming that the categorical features are replaced by suitable real values) to a set of class labels $\mathcal{C} = \{1, 2, \ldots, C\}$. A classifier is designed for the purpose of estimating the properties of the mapping $g : X \rightarrow \mathcal{C}$. First, in the training phase, the classifier is fed with a training set P ($P \subseteq X$ and $|P| = n$) to learn about the characteristics of $g(.)$. In this stage, for all data points $x_i \in P$, where $i = 1, 2, \ldots n$, the value of $g(x_i)$ is available to the classifier. Once trained, the classifier is expected to correctly predict the value of $g(y_i)$ for a new data point $y_i \in Q$ ($Q \subseteq X$, $|Q| = m$, and $i = 1, 2, \ldots, m$). This is called the testing phase. The data point y_i is known as a test or query point, while the set Q of all such points is called a test set.

The k-Nearest Neighbor (kNN) classifier has always been preferred for its methodical simplicity, nonparametric working principle [1], and ease of implementation. The kNN classifier involves tuning of a single parameter k (the number of nearest neighbors to be considered). However, it is not easy to find the value of k for which the algorithm performs optimally on a wide range of data sets (or for all the points in the same data set). Theoretical studies suggest that the number of points in the training set (say n) and the value of k ($1 \leq k \leq n$) both control the performance of the kNN algorithm [2]. Furthermore, if $k = 1$, then the probability of misclassification will be bounded above by twice the risk of the Bayes decision rule as $n \rightarrow \infty$ [2]. However, depending on the data set, choices other than $k = 1$ may be more suitable [3]. Therefore, the theory discussed in [2] and [4] does not help with the choice of k in practical cases. Usually, a global k is chosen, i.e., a single value of k for classifying all test points. Conventional choices of such a global k value are $1, 3, 5, 7$, and 9 [5], [6], but may also be as large as $k = \sqrt{n}$ [1], [7]. To optimize the performance, kNN is commonly run with a number of different k values. Subsequently, several techniques, such as cross validation and probabilistic estimation [8], [9], may be used to choose the best k value among the tested k values. While probabilistic modeling-based algorithms are hard to implement and usually depend on prior assumptions about the data set, the technique of cross validation is computationally rather expensive. Moreover, as the distribution of the classes is not known *a priori*, any choice of global k stands a risk of ignoring the local distribution of the neighborhood of a test point, whereas the consideration of the unique features of the locality of a test point should decrease the chance of misclassification of that point. In this paper, these facts encouraged us to choose a data-point-specific k value using an indicator of the local density and class distributions of its neighborhood.

Besides the difficulty with the selection of k, kNN classification rule also faces a challenge over the data sets with class imbalance, i.e., when all the classes do not have comparable number of representatives [10]. Hence, we also introduce a class-specific global weighting scheme to tackle the issue of class imbalance.

Manuscript received December 28, 2016; revised July 11, 2017 and November 29, 2017; accepted February 27, 2018. Date of publication March 27, 2018; date of current version October 16, 2018. *(Corresponding author: Swagatam Das.)*

The authors are with the Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: sankha_r@isical.ac.in; shounak.jaduniv@gmail.com; swagatam.das@isical.ac.in).

This paper has supplementary downloadable material for online publication only, as provided by the authors. This includes the detailed proof of some theorems stated in the main paper, details of the parameter tuning, experimental results on individual datasets, and explanation of biases of some of the indices used. This material is 260 KB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2812279