Pattern Recognition Letters 80 (2016) 231–237

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A feature weighted penalty based dissimilarity measure for *k*-nearest neighbor classification with missing features☆

Shounak Datta[a], Debaleena Misra[b], Swagatam Das[a,*]

[a] Electronics and Communication Sciences Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700 108, India
[b] Department of Instrumentation and Electronics Engineering, Jadavpur University, Salt Lake campus, Block-LB, Plot No. 8, Sector-III, Kolkata 700 098, India

ARTICLE INFO

Article history:
Received 28 December 2015
Available online 6 July 2016

Keywords:
*k*NN classifier
Missing features
Dissimilarity measure
Penalized dissimilarity

ABSTRACT

The *k*-Nearest Neighbor (*k*NN) classifier is an elegant learning algorithm widely used because of its simple and non-parametric nature. However, like most learning algorithms, *k*NN cannot be directly applied to data plagued by missing features. We make use of the philosophy of a Penalized Dissimilarity Measure (PDM) and incorporate a PDM called the Feature Weighted Penalty based Dissimilarity (FWPD) into *k*NN, forming the *k*NN-FWPD classifier which can be directly applied to datasets with missing features, without any preprocessing (like marginalization or imputation). Extensive experimentation on simulations of four different missing feature mechanisms (using various datasets) suggests that the proposed method can handle the missing feature problem much more effectively compared to some of the popular imputation mechanisms (used in conjunction with *k*NN).

1. Introduction

1.1. Overview

The *k*-Nearest Neighbor (*k*NN) classifier, dating back to [11], is one of the oldest as well as simplest pattern classification techniques. Its continued popularity is owing to its simple yet effective philosophy, competitive performance, ease of implementation, and non-parametric computational basis, i.e. the fact that the *k*NN classifier does not make any prior assumptions about the class distributions. Cover and Hart [7] showed that the 1NN classifier achieves a probability of error less than twice the Bayes probability of error, when the size of the training set tends to infinity. The *k*NN classifier functions by finding the k nearest neighbors of a test point from among a set of training data instances with known class labels. It then assigns the test point to the class corresponding to the majority of the k nearest neighbors, i.e. the class label predicted for the test point is that of the majority of its k nearest neighbors. The only parameter involved is k, which should be chosen so that $k \to \inf$ and $\frac{k}{n_1} \to 0$, as the number of training points $n_1 \to 0$.

Real applications of classification often have to deal with datasets consisting of some instances having one or more unobserved features. This is termed as *missingness*. There can be a variety of reasons behind missingness, such as data input errors, inaccurate measurement, equipment malfunction or limitations, measurement noise, data corruption, etc. This is called unstructured missingness as it does not have any structural implications on the dataset [5,17]. Missingness may also occur if all features are not defined for all the data points in a dataset. Such missingness is referred to as structural missingness or absence of features [6]. For example, credit-card details may not be defined for non-credit clients of a bank. This paper deals with unstructured missingness, hereafter referred to simply as missingness. Little and Rubin [15] proposed a three-fold classification of missingness, viz. Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). Features are said to be MCAR when the likelihood of a feature being unobserved for a particular data instance depends neither on the observed nor on the unobserved features of that instance. For example, in an annual income survey, a citizen is unable to participate owing to reasons unrelated to the survey, such as traffic or schedule problems. Features are MAR when the missingness depends only on the observed features of an instance, and not on the unobserved features. Suppose, college-goers are less likely to report their income than office-goers. But, whether a college-goer will report his/her income is independent of the actual income. MNAR refers to the case where missingness is subject to the unobserved features of an instance. For example, people with lower earnings are less likely to report their incomes in the annual income survey.

There are two main traditional approaches to missing feature handling, namely *marginalization* and *imputation*. Marginalization

☆ This paper has been recommended for acceptance by Prof. A. Marcelli.

* Corresponding author. Tel.: +91 33 2575 2323.
E-mail address: swagatamdas19@yahoo.co.in (S. Das).

http://dx.doi.org/10.1016/j.patrec.2016.06.023
0167-8655/

refers to the practice of excluding data instances with missing features. This leads to loss of additional data, and will be ill-advised in applications where a sizable portion of the data have unobserved values. Therefore, most of the research on missing feature handling has been focused on imputation techniques which aim to estimate the missing features on the basis of the observed features. Common imputation methods [10] involve filling the missing features of data points with zeros (Zero Imputation (ZI)), or the averages of the corresponding features over the entire dataset (Average Imputation (AI)). Class Mean Imputation or Concept Mean Imputation (CMI) is a slight modification of AI where a missing feature is filled with the average of the feature over all instances within the same class as the instance being filled. Another simple yet effective imputation method is *k*-Nearest Neighbor Imputation (*k*NNI) [9], where a missing feature of a data instance is estimated to be the average of corresponding features of its *k* nearest neighbors (on the observed subspace). Rubin [18] suggested Multiple Imputation (MI), a technique where the missing values are imputed by a few (typically 5–10) distinct estimates; the actual number depending on the percentage of missingness. Such a method of repeated imputation is capable of incorporating (in the learning process) the uncertainty inherent in imputation. Some more sophisticated techniques have been developed, especially by the bioinformatics community, which attempt to estimate the missing features by exploiting the correlations between data. Troyanskaya et al. [20] proposed a weighted variant of *k*NNI along with a Singular Value Decomposition based Imputation (SVDI) technique, which performs regression based estimation of the missing values using the *k* most significant Eigenvectors of the dataset. The Least Squares Imputation (LSI) technique [4] is another such technique. Sehgal et al. [19] further combined LSI with Non-Negative LSI (NNLSI) to create the Collateral Missing Value Estimation (CMVE) algorithm.

Despite the simplicity and elegance of the *k*NN classifier, there has been limited research on its ability to handle missingness. Perhaps, this is because the said classifier is known to be very sensitive to noise [2]. Since imputation methods often introduce noise into the dataset due to noisy estimations [15], imputation is not the best approach to adapt the *k*NN classifier to missingness. Acuña and Rodriguez [1] published a comparative study on the effects of some imputation methods on *k*NN classifier and Linear Discriminant Analysis (LDA). The former technique is observed to produces higher error compared to the latter. García-Laencina et al. [12] proposed a mutual information based *k*NN algorithm to simultaneously perform classification and imputation. Ashraf et al. [3] proposed a scheme to iteratively employ 1NNI and *k*NN classification until a desired level of accuracy is achieved (did not provide any guidelines for the selection of an appropriate accuracy threshold). Liu et al. [16] proposed an adaptive imputation scheme, where the *k*NN algorithm is used as the underlying classifier.

1.2. Motivation

Imputation methods introduce noise into the dataset and are also known to be less effective when features are MNAR [13]. Moreover, many of the more sophisticated imputation methods are computationally expensive and do not scale well to large datasets. Hence, a more suitable alternative way of adapting the learning methods to missingness is to modify the underlying *distance* or *dissimilarity measure*, so that the modified dissimilarity measure uses the *common observed features* (features observed for both instances) to approximate the distances between two data instances if they were to be fully observed. Such approaches neither require marginalization nor imputation, while possibly yielding results better than both. Let us look at an example. Let $X_{full} = \{\mathbf{x}_1 = (1, 5), \mathbf{x}_2 = (2, 3), \mathbf{x}_3 = (3, 6)\}$ be a dataset consisting of three points in $\mathbb{R}^2$. Then, $d_E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{5}$ and $d_E(\mathbf{x}_1, \mathbf{x}_3) = \sqrt{5}$

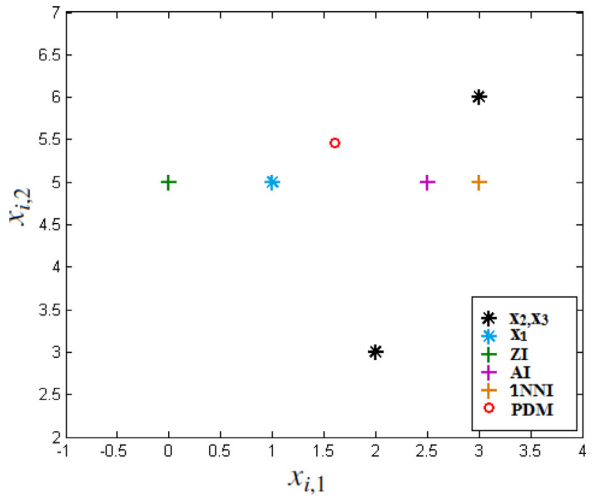

Fig. 1. Comparison of various techniques for handling missing features.

(where $d_E(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between the fully observed points $\mathbf{x}_i$ and $\mathbf{x}_j$ in X_{full}). Now, let the 1st coordinate of the point (1, 5) be unobserved. Then, the resulting dataset, on which learning must be undertaken, is $X = \{\mathbf{x'}_1 = (*, 5), \mathbf{x}_2 = (2, 3), \mathbf{x}_3 = (3, 6)\}$, where '$*$' denotes the unobserved value. Then the filled in datasets X_{ZI}, X_{AI}, and X_{1NNI} obtained respectively using ZI, AI and 1NNI (*k*NNI with $k = 1$) are

$$X_{ZI} = \{\hat{\mathbf{x}_1} = (0, 5), \mathbf{x}_2 = (2, 3), \mathbf{x}_3 = (3, 6)\},$$
$$X_{AI} = \{\hat{\mathbf{x}_1} = (2.5, 5), \mathbf{x}_2 = (2, 3), \mathbf{x}_3 = (3, 6)\},$$
$$\text{and } X_{1NNI} = \{\hat{\mathbf{x}_1} = (3, 5), \mathbf{x}_2 = (2, 3), \mathbf{x}_3 = (3, 6)\},$$

where $\hat{\mathbf{x}_1}$ denotes an estimate of $\mathbf{x}_1$. Since the observed distance between two data instances is a lower bound on the fully observed distance between them, adding a suitable penalty to this lower bound can yield a reasonable approximation of the actual distance. We call this a Penalized Dissimilarity Measure (PDM). Let the penalty between $\mathbf{x'}_1$ and some other $\mathbf{x}_i \in X$ be given by the ratio of the number of features which are unobserved for at least one of the two data instances and the total number of features in the dataset. Then, the dissimilarity $\delta'(\mathbf{x}_1, \mathbf{x}_i)$ between $\mathbf{x'}_1$ and some other $\mathbf{x}_i$ is

$$\delta'(\mathbf{x}_1, \mathbf{x}_i) = \sqrt{(x_{1,2} - x_{i,2})^2} + \frac{1}{2},$$

where the 1 in the numerator of the penalty term is due to the fact that the 1st feature of $\mathbf{x'}_1$ is unobserved. Therefore,

$$\delta'(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(5-3)^2} + \frac{1}{2} = 2.5,$$
$$\text{and } \delta'(\mathbf{x}_1, \mathbf{x}_3) = \sqrt{(5-6)^2} + \frac{1}{2} = 1.5.$$

The estimates obtained by the different methods are illustrated in Fig. 1. The reader should notice that while the points estimated using ZI, AI and 1NNI exist in the same 2-D Cartesian space to which X_{full} is native, the point estimated by the PDM exists in an abstract space (likely distinct from the native 2-D space). However, for the sake of easy comparison, we illustrate all the estimates together by superimposing this abstract space on the native 2-D space so as to coincide at the points $\mathbf{x}_2$ and $\mathbf{x}_3$. It is seen that the approach based on the PDM is better able to preserve the relationship between the points. Based on this knowledge, we deduce that the *k*NN classifier can easily be adapted to problems with missing features, using a PDM as the underlying dissimilarity.