A new supervised classification technique based on the joint use of K-nearest neighbors and weighted K-means to discover new patterns in the data

Mohammed Sabri, Rosanna Verde, Fabrizio Maturo, Antonio Balzanella, Hamid Tairi and Ali Yahyaouy

Abstract This study deals with supervised classification (K-nearest neighbors classifier) using a multi-step procedure. A weighted k-means clustering is adopted in the first step to discovering new patterns in the training groups to partition each one separately into a predetermined number of subgroups. The clustering algorithm uses an adaptive distance to evaluate the importance of the variables in the clustering process by measuring the distance between the objects and centroids. Therefore, a new objective function is proposed to take into account the homogeneity between the subgroups of each original group. In the second step, a K-NN algorithm is performed using the new labels and weights extracted from the K-means clustering algorithm. The results of applying the proposed classifiers to a real data set have their effectiveness in terms of accuracy and their practical utility as a result of their interpretable nature.

Key words: Supervised classification, Unsupervised classification, Adaptive distance, K-nearest neighbors (KNN)

Mohammed Sabri
Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy.
Department of Informatics, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco, e-mail: mohammed.sabri@usmba.ac.ma

Rosanna Verde, Antonio Balzanella
Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy
e-mail: rosanna.verde@unicampania.it, antonio.irpino@unicampania.it

Fabrizio Maturo
Faculty of Economics, Universitas Mercatorum, Roma, Italy e-mail: fabrizio.maturo@unimercatorum.it

Hamid Tairi, Ali Yahyaouy
Department of Informatics, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco e-mail: htairi@yahoo.fr, ali.yahyaouy@usmba.ac.ma

1 Introduction

Classification of high-dimensional data is driven by a need for methods to deal with the increasing availability of data coming from various sources, such as sensors, imaging devices, and digital platforms. Furthermore, with the explosion of big data, high-dimensional classification has become an essential tool for many real-world applications, ranging from computer vision, speech recognition, and natural language processing to finance healthcare, and marketing [1; 2]. As a result, there is a growing demand for advanced algorithms and techniques that can effectively handle the complexity and variability of high-dimensional data while ensuring robust and accurate predictions. The ongoing research in this area aims to develop new approaches that can overcome the challenges posed by high-dimensional data, such as the curse of dimensionality, overfitting, and the need for large amounts of labeled data, and offer scalable and interpretable solutions for real-world applications. Notably, in this contribution, we focus on supervised classification to learn from high-dimensional data using a new approach.
This study suggests improving the performance of a supervised classifier by proposing a two-phases classification approach [3]. In the first step, an adaptive k-means clustering classification [6] is proposed to discover hidden patterns in the original groups by adopting a new objective function by considering both intracluster compactness and intercluster separation of the subgroups.
Then, supervised classification based on K-nearest neighbors [4] is applied to exploit the additional information on the weights and new labels coming from the first step. The basic idea is to get additional information in the training data to improve the accuracy of the final classifier. Also, an adaptive Euclidean distance is used to measure the similarity among samples and cluster centers by providing weights to each variable. The benefit of using the adaptive distance with the clustering algorithm is to evaluate the clustering process [5]. Therefore, the adaptive distance might be viewed as an automatic variable scaling that assigns new vector weights to each cluster.

2 The two-steps classification technique

The results of the supervised classification investigation revealed that the algorithms encountered challenges in complex and high-dimensional data sets. This highlights the need for a novel approach to address this type of data. Our proposal advocates for the utilization of clustering to uncover novel patterns within the dataset, thereby incorporating a blend of supervised and unsupervised techniques to enhance the efficiency of the classification process.
The initial phase involves proposing a variation of the K-means algorithm by proposing a new objective function for grouping each class within the training dataset into a specified number of subclusters. Thus, this novel objective function incorporates both intracluster compactness and intercluster separation between the

subgroups. In addition, this research proposes a new adaptive distance to measure the similarity between two instances. The latter adaptive distance involves the use of weights with the Euclidean distance to assign weights to each cluster. The use of this adaptive distance with K-means facilitates the automatic calculation of a system of weights for the variables involved in the optimization procedure of the objective function.

The main point is that each cluster has a distinct distance associated with it for comparing clusters and their representatives, which change at each iteration,

This study considers the case of data that originally belong to two starting groups, designated as A and B. The membership matrix for group A is represented by U_A, a binary $g_A \times K_A$ matrix, where $u_{ip} = 1$ indicates that object i has been assigned to subgroup p of the group A. Conversely, a value of 0 indicates non-assignment. The centroids of k_A clusters in group A are represented by a set of vectors, $Z = \{Z_1, Z_2, \ldots, Z_{k_A}\}$, and the corresponding centroids for k_B clusters in group B are represented by another set of vectors, $Z' = \{Z'_1, Z'_2, \ldots, Z'_{kB}\}$. Additionally, the weights of the k_A clusters in group A are represented by the matrix $W = \{W_1, W_2, \ldots, W_{k_A}\}$, where $W_p = \{w_{p1}, w_{p2}, \ldots, w_{pm}\}$ represents the weights of the m features in cluster p. Analogously, the subgroup weights in group B are represented by $W' = \{W'_1, W'_2, \ldots, W'_{kB}\}$.

We change the objective function to combine intracluster compactness and intercluster separation as indicated in eq. (1). Our technique uses the K-means algorithm to minimize the following objective function,

$$
\begin{aligned}
P(U,W,Z,U',W',Z') = \sum_{j=1}^{m}\Big(\sum_{p=1}^{k_A}\frac{w_{pj}^2\sum_{i=1}^{n_A}u_{ip}(x_{ij}-z_{pj})^2}{g_A(z_{pj}-z'_{Bj})^2} + \sum_{q=1}^{k_B}\frac{w'^2_{qj}\sum_{l=1}^{n_B}u'_{lq}(x'_{lj}-z'_{qj})^2}{g_B(z'_{qj}-z_{Aj})^2}\Big) \\
\text{s.t.} \quad u_{ip}, u'_{i'q} \in \{0,1\}, \ \sum_{p=1}^{n_A} u_{ip} = 1 \ \ \sum_{q=1}^{n_B} u'_{lq} = 1 \\
\sum_{j=1}^{m} w_{pj} = 1 \qquad \sum_{j=1}^{m} w'_{pj} = 1
\end{aligned}
\tag{1}
$$

z_{Aj} is the jth feature of the global centroid z_A of the group A.

We calculate z_{Aj} as

$$z_{Aj} = \frac{\sum_{i=1}^{n_{g1}} x_{ij}}{n_A}$$

z'_{Bj} is the jth feature of the global centroid z_B of the group B.

We calculate z_{Bj} as

$$z_{Bj} = \frac{\sum_{l=1}^{n_{g2}} x'_{lj}}{n_B}$$

To solve eq. (1), we first initialize the parameters for the two groups. Then, we consider the case that the allocation in group A. So the minimization problem is reduced to solving

$$P(U,W,Z) = \sum_{p=1}^{k_A}\sum_{i=1}^{n_A} u_{ip} \sum_{j=1}^{m} w_{pj}^2 \frac{(x_{ij}-z_{pj})^2}{n_{g1}(z_{pj}-z'_{Bj})^2}$$
$$\text{s.t.} \quad u_{ip} \in \{0,1\}, \ \sum_{p=1}^{k_A} u_{ip} = 1 \tag{2}$$
$$\sum_{j=1}^{m} w_{pj} = 1$$

We can minimize eq. (2) by solving the following problems iteratively:

1. Problem $P1$: fix $Z = \hat{Z}$, $W = \hat{W}$ and solve the reduced problem $P(U,\hat{Z},\hat{W})$
2. Problem $P2$: fix $U = \hat{U}$, $W = \hat{W}$ and solve the reduced problem $P(\hat{U},Z,\hat{W})$
3. Problem $P3$: fix $U = \hat{U}$, $Z = \hat{Z}$ and solve the reduced problem $P(\hat{U},\hat{Z},W)$

The problem $P1$ is solved by

$$u_{ip} = \begin{cases} 1 & \text{if } \sum_{q=1}^{k_B}\sum_{j=1}^{m} w_{pj}^2 \frac{(x_{ij}-z_{pj})^2}{n_{g1}(z_{pj}-z'_{Bj})^2} \leq \sum_{q=1}^{k_B}\sum_{j=1}^{m} w_{pj}^2 \frac{(x_{ij}-z_{rj})^2}{n_{g1}(z_{pj}-z'_{Bj})^2} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $1 \leq r \leq k$, $r \neq p$.

To solve the problem $P2$ we derive the gradient of P with respect to $z_{p,j}$ as

$$z_{pj} = \frac{\sum_{i=1}^{n_A} u_{ip} x_{ij} (x_{ij} - z'_{Bj})}{\sum_{i=1}^{n_A} u_{ip} (x_{ij} - z'_{Bj})} \tag{4}$$

The problem $P3$ is solved by setting up a Lagrangian equation to $P(\hat{U},\hat{Z},W)$ with multiplier λ.

$$w_{pj} = \frac{1}{D_{pj} \sum_{j=1}^{m} \frac{1}{D_{pj}}} \tag{5}$$

where

$$D_{pj} = \sum_{i=1}^{n_A}\sum_{q=1}^{k_B} u_{ip} \frac{(x_{ij}-z_{pj})^2}{n_A (z_{pj}-z'_{Bj})^2} \tag{6}$$

Following the same procedure, we represent the parameters U', Z' and W' for group B.
The entire process of the new k-means clustering algorithm based on the adaptive distance is described in Algorithm 1.

Algorithm 1 K-means with adaptive Euclidean distance

1: **Input:** $X = \{X_1, X_2, \ldots, X_n\}$, k1, k2
2: **Output :** U, Z, W, U', Z', W'.
3: Split the dataset into two original groups
4: Randomly initialize $Z^0 = \{Z_1, Z_2, \ldots, Z_{k1}\}$, $Z'^0 = \{Z'_1, Z'_2, \ldots, Z'_{k2}\}$, and weights W^0 and W'^0
5: **repeat**
6: Fixed $\hat{W}, \hat{Z}, \hat{W}', \hat{Z}'$ solve the membership matrix U, U' with eq. (3);
7: Fixed $\hat{U}, \hat{W}, \hat{U}', \hat{W}'$ solve the membership matrix Z, Z' with eq. (4);
8: Fixed $\hat{U}, \hat{Z}, \hat{U}', \hat{Z}'$ solve the membership matrix W, W' with eq. (5);
9: **until** convergence

Afterward, in the second phase, the features serve as inputs to the KNN algorithm for the purpose of training and evaluating the classification model. However, the KNN classifier trains on newly assigned labels, and the weights are calculated via the K-means clustering algorithm. It should be noted that objects belonging to the same cluster possess identical vector weights.

3 Application

We apply the method described in the previous section to the well-known Breast Cancer Wisconsin (Diagnostic) dataset[1] formatted in 1998 by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasaria. Features are taken as digitized images of a fine needle aspiration (FNA) of breast mass from the UCI machine learning repository. The characteristics of the multivariate data set depict the cell nuclei features values within the image. Our objective is to create a model to classify the data into "benign" or malignant".
The performance of the new supervised classification approach with the breast cancer datasets is depicted in Figure 1. This algorithm employs adaptive K-means in the first phase of unsupervised classification, where the benign group is divided into four subgroups and the malignant group into three subgroups to uncover the hidden layer of the original dataset. The proposed approach, which includes the augmented labels, demonstrates a noticeable improvement in accuracy compared to the classical KNN, as indicated in the plot. Specifically, a 98.25% accuracy rate is achieved with the use of the 9-Nearest Neighbor.

	Number of subgroups				
Group 1	2	2	3	4	4
Group 2	3	2	2	2	3
Accuracy	96.49	94.15	95.91	96.49	98.83

Table 1 Accuracy of the proposed approach according to different combinations in the number of subgroups

[1] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

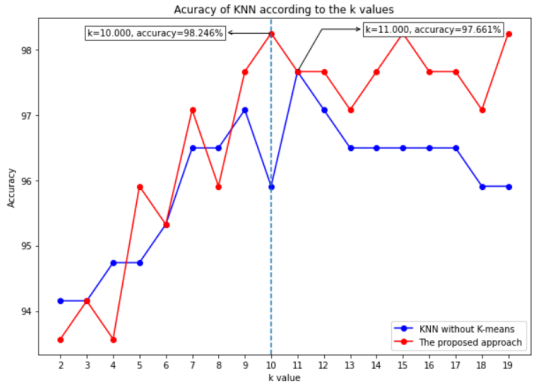

Fig. 1 New subgroups discovered in diseased subgroup

Table 1 shows the effectiveness of the proposed two-step classification approach based on the adaptive distance with different combinations of subgroups numbers. A maximum accuracy of 98.83% is obtained using RF and K-means with the new distance.

References

[1] Bhadani, A. K., Jothimani, D. . Big data: challenges, opportunities, and realities. Effective big data management and opportunities for implementation (2016): 1-24

[2] Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(**3**), p.160 (2021)

[3] Maturo, F. and Verde, R.: Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers. Computational Statistics, 1-32 (2022)

[4] Peterson, L. E.: K-nearest neighbor. Scholarpedia 4.2, 1883 (2009)

[5] Diday, E.: Classification automatique avec distances adaptatives. RAIRO Informatique Computer Science, 4(**11**), 329-349 (1977)

[6] Irpino, A., Verde, R. and De Carvalho: Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Systems with Applications, 41(**7**), 3351-3366 (2014)