Pattern Recognition Letters 63 (2015) 50–58

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Automated feature weighting in clustering with separable distances and inner product induced norms – A theoretical generalization ☆

Arkajyoti Saha [a], Swagatam Das [b,*]

[a] Stat-math Unit, Indian Statistical Institute, Kolkata 700108, India
[b] Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 2 December 2014
Available online 11 June 2015

Keywords:
Automated feature weights
Separable distance measures
Inner product induced norms
Clustering
Lloyd type algorithm
Alternative optimization

ABSTRACT

For decades practitioners have been using the separable distance and inner product induced norms as the distance measures for *k*-means, Fuzzy C-Means (FCM), hard and fuzzy *k*-modes clustering algorithms. In this paper, we introduce a novel concept of automated feature weighting for general clustering algorithms (including both hard and fuzzy clustering) to amplify the effect of the discriminating features, which play a key role in identifying the naturally occurring groups in data with minimal computational overheads. We derive a Lloyd heuristic and an alternating optimization algorithm for solving the hard and the fuzzy clustering problems respectively. We also investigate the mathematical nature of the problems in sufficient details to guarantee the existence and feasibility of a solution at each iteration of the aforementioned algorithms. We show that majority of the automated feature weighting schemes existing in the literature turn out to be the special cases of this proposed generalization. A brief discussion on practical utility of the proposed generalization is also presented along with indication of the future applications of this new approach.

1. Introduction

Ability to partition a collection of objects into meaningful groups stands out to be a fundamental mode of learning. Clustering algorithms appear as the formal tools for the computer-aided detection of the naturally occurring groups in a collection of objects or dataset. The aim of any clustering algorithm is to evolve a partition matrix representing a possible grouping of the dataset into a number of clusters such that objects belonging to the same group may share maximum amount of similarity in some sense, whereas objects from different groups are as dissimilar as possible with respect to the same sense.

Modified versions of *k*-means and similar hard partitional clustering algorithms [19,26] are very widely used to cluster large datasets owing to their simplicity and ability to handle both numerical as well as categorical attributes. The hard *k*-means algorithm assigns each data point to the cluster, whose cluster centroid is nearest to that data point. An optimal hard clustering can be identified with a Voronoi diagram whose seeds are the centers of the elements of the cluster. Lloyd's heuristic [24] is a popular choice for optimizing the *k*-means objective function. On the other hand, fuzzy clustering does not define subsets in the usual sense but rather model each cluster as a fuzzy set as defined by Zadeh [32]. Perhaps the most widely used popular fuzzy clustering algorithm is the fuzzy ISODATA clustering or Fuzzy C-Means (FCM) algorithm which was first proposed by Dunn [12] and then generalized (by generalizing the value of the fuzzifier) by Bezdek and his co-workers [4,5]. The FCM algorithm assigns a membership (in [0, 1], subject to conditions discussed in (2a)–(2b)) to each data point indicating its belongingness to a particular cluster. The memberships to different clusters are inversely related to the relative distance of that data point from the corresponding cluster centroids. This algorithm uses an alternating optimization (AO) heuristic [3] to locally minimize the criterion function.

Huang [17] extended the conventional *k*-means algorithm for categorical attributes by introducing a simple matching dissimilarity measure and replacing the mean by mode. The resulting algorithm was, thus, named as the *k*-modes algorithm. Later a fuzzy counterpart of *k*-modes was also introduced [18]. In this case the membership upgrading formula closely followed that of the conventional FCM algorithm, with the matching dissimilarity measure.

In practical data clustering situations, all the features that characterize a data point do not bear equal importance. Some features may even affect the partitioning task adversely. Thus, it is very important to select the most discriminating features and at the same time to eliminate the non-discriminative and/or derogatory features prior to clustering. Each feature may be considered to have a relative degree of importance (usually mapped to the interval [0, 1]) which should be

☆ Paper has been recommended for acceptance by Y. Liu.

* Corresponding author. Tel/fax.: +91 033 2575 2323.
E-mail address: swagatamdas19@yahoo.co.in, swagatam.das@isical.ac.in (S. Das).

http://dx.doi.org/10.1016/j.patrec.2015.06.001
0167-8655/

called the *feature weight*. Feature weighting can be thought of as an extension of the conventional feature selection that can have weight values either 0 or 1.

Existing literature on cluster analysis comprises a good volume of works on both hard and fuzzy clustering coupled with the feature weighting schemes. The first known approach of integrating variable weighting as a part of the clustering process dates back to the works of Sneath and Sokal [30] and Lumelsky [25]. Sneath et al. first used the term *attribute weighting* in this context. Subsequently in 1984, Desarbo et al. [11] used a two-stage SYNCLUS process to estimate the optimal weights of the features in a k-means clustering framework. Subsequently De Soete [10] proposed a method to derive the most suitable variable weights for ultrametric and additive tree fitting. This method was used in a hierarchical clustering framework. Makarenkov and Legendre [27] extended De Soete's method to optimal feature weighting for the k-means clustering. However, use of the Polak–Ribiere optimization procedure to minimize a squared cost function involving the weights, considerably reduced the computation speed of their method. Modha and Spangler (2003) proposed a method to optimize attribute weights for obtaining the best clustering through minimization of the ratio of the mean within-cluster distortion over the mean between-cluster distortion, referred to as the generalized Fisher ratio Q. Chan et al. [7] developed a new procedure to generate a weight for each of the attributes from each cluster for the k-means type algorithms, in context to both numerical and categorical data. Variable weighting has also been integrated with several fuzzy clustering schemes. Keller and Klawonn [21] introduced an automatically determined influence parameter for each single data variable for each cluster. Frigui and Nasraoui [14] modified the FCM objective function by incorporating the feature weights and used an AO heuristic to find the optimal feature weights besides refining the cluster centroids. Huang et al. [16] introduced a new step into the k-means algorithm for estimating the locally optimal feature weights based on the current grouping of the data. They also analytically investigated the convergence of the weighted k-means algorithm. This particular automated weighting concept was extended to the FCM algorithm in [28]. The scheme was also generalized for the Minkowski metric [9]. Some recent works in the same direction can be found in [1,8,13,20].

Most of the previously mentioned clustering algorithms used squared Euclidean distance and introduced the idea of automated feature weighting in that context. A clustering algorithm can be designed if the proper distance metric is combined with automatic learning of the cluster centroids, membership values, as well as the feature weights. The present paper is a humble attempt to generalize the automated feature weighting scheme for all separable distance measures and some non-separable distances like the inner product induced norms (IPINs) and the polynomial distances. We develop a Lloyd heuristic and an AO algorithm for solving the hard and fuzzy clustering problems respectively. We investigate the existence of the optimal points for the optimization tasks under consideration. We show that the introduced weighting scheme and the solution to the clustering problems coincide with that of the automated feature weighted clustering algorithms discussed earlier ([21]; [7,9,14,28]). We also discussed the advantages of this generalization and its practical utility.

The rest of the paper is organized in the following way. Section 2 provides a detailed description of the generalized weighting scheme. Section 3 outlines the algorithms to solve the hard and fuzzy clustering problem with the generalized weighting scheme. Section 4 discusses the mathematical properties of optimization tasks in the two algorithms developed in Section 3. In Section 5, we show that the weighting scheme and the corresponding algorithms developed in some of the aforementioned papers are only special cases of the newly proposed generalized weighting scheme and the corresponding general algorithm. In Section 6, we present a discussion on how this concept of automated feature weighting can be extended to the non-separable distance measures. Finally the paper is concluded in Section 7 with a brief discussion on the advantages of this generalization along with its practical utility and future scope of applications.

2. Variable weighting scheme with separable distances and inner product induced norms

2.1. General clustering framework

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in D \subseteq \mathbb{R}^d$ be the finite set of patterns (also synonymously called objects, data-points, observations, or feature vectors) under consideration. To partition the patterns into c groups with $2 \le c \le n$, the following mathematical program is considered:

$$P: \textit{minimize } f_m(\mathbf{U}, \mathcal{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{c} {u_{ij}}^m d(\mathbf{z_j}, \mathbf{x_i}). \tag{1}$$

subject to

$$\sum_{j=1}^{c} u_{ij} = 1, \quad \forall i = 1, 2, \ldots, n, \tag{2a}$$

$$0 < \sum_{i=1}^{n} u_{ij} < n; \quad \forall j = 1, 2, \ldots, c. \tag{2b}$$

For hard clustering algorithm we have

$$m = 1, u_{ij} \in \{0, 1\}, \quad \forall i = 1, 2, \ldots, n; \quad \forall j = 1, 2, \ldots, c; \tag{2c}$$

and for fuzzy clustering algorithm:

$$m > 1, \ u_{ij} \in [0, 1], \quad \forall i = 1, 2, \ldots, n; \quad \forall j = 1, 2, \ldots, c; \tag{2d}$$

where

$\mathbf{U} = [u_{ij}]$ is the membership matrix,

$\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_c\}$, $\mathbf{z}_j \in S$, $\forall j = 1, 2, \ldots, c$; is the set of centers of the clusters, $S \subseteq \mathbb{R}^d$; $S = (S_1 \times S_2 \times \ldots \times S_d)$, $S_l \subseteq \mathbb{R}$, with $l = 1, 2, \ldots, d$.

For the Euclidean distance measure, $S_l = \mathbb{R}$; $S = \mathbb{R}^d$, but for some general distance measures it can happen that $S_l \subset \mathbb{R}$. For example, if the used divergence measure is f divergence [2], $S_l = \mathbb{R}_+$, $l = 1, 2, \ldots, d$; $S = {\mathbb{R}^d}_+$. In what follows for sake of notational simplicity, we consider $S = \mathbb{R}^d$.

Now $d(\mathbf{z}_j, \mathbf{x}_i)$ is the distance measure between the jth cluster center and the ith pattern. Note that

(a) If separable distance measures are used

$$d(\mathbf{z}_j, \mathbf{x}_i) = \sum\nolimits_{l=1}^{d} d(z_{jl}, \mathrm{x}_{il}).$$

(b) If IPIN is used, then

$$d(\mathbf{z}_j, \mathbf{x}_i) = (\mathbf{z}_j - \mathbf{x}_i)^T \mathbf{A} (\mathbf{z}_j - \mathbf{x}_i),$$

where $\mathbf{A}$ is any positive definite matrix.

Now, we introduce the weighting scheme in the following way.

Let $\mathbf{w}_j = [w_{j1}, w_{j2}, \ldots, w_{jd}]^{\boldsymbol{T}}$ be the vector of weights of the d variables of the pattern corresponding to the jth cluster and the weight matrix be $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_c]^{\boldsymbol{T}}$. Also let $\beta = (\beta_1, \beta_2, \ldots, \beta_c)$ be the corresponding vector of exponent parameters for the attribute weights, where β_j corresponds to the jth cluster. Then the optimization problem under consideration in (1) can be modified as follows:

$$WP: \textit{minimize } f_{m,\beta}(\mathbf{U}, \mathcal{Z}, \mathbf{W}) = \sum_{i=1}^{n}\sum_{j=1}^{c} {u_{ij}}^m d_{\mathbf{W}}(\mathbf{z}_j, \mathbf{x}_i). \tag{3}$$

Here $d_{\mathbf{W}}(\mathbf{z}_j, \mathbf{x_i})$ is the weighted distance measure between the jth cluster center and the ith pattern, defined in the following way: