Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems ••• (••••) •••–•••

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy weighted C-ordered means clustering algorithm

Krzysztof Siminski

Institute of Informatics, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland

Received 21 April 2016; received in revised form 11 November 2016; accepted 7 January 2017

Abstract

In real life data sets some attributes may have lower importance or even may be completely noninformative. The subspace clustering algorithms have been proposed to handle this. The soft subspace algorithms are vulnerable to noise and outliers. The paper presents a novel algorithm that handles both various importance of attributes and outliers. The proposed Fuzzy Weighted C-Ordered Mean (FWCOM) clustering algorithm elaborates clusters in soft subspaces. In each cluster each attribute is assigned a weight from interval [0, 1]. Each attribute has its individual weight (importance) in each cluster. The proposed algorithm applies the ordering technique to effectively reduce the influence of outliers and noise. The paper is accompanied by numerical experiments.

Keywords: Fuzzy clustering; Subspace clustering; Ordered weighted averaging

1. Introduction

The real life data are far from being ideal. Data granules may be embedded in some subspaces of the task space. Some of dimensions may have lower importance or may even be totally noninformative and superfluous. The global reduction of task dimensionality by feature transformation (e.g. Principal Component Analysis or Singular Value Decomposition) may lead to problems with interpretation of elaborated results. The global approach may not be satisfactory because noninformative dimensions in one data granule may be of high importance in the other one. This leads to subspace clustering [7,10,14] for elaboration of data granules in subspaces of the original task space.

Many subspace algorithms can be classified as top-down or bottom-up techniques. The top-down algorithms start with all dimensions and try to throw away dimensions of lower importance (e.g. PROCLUS [1], ORCLUS [2], δ-Clusters [22]. The bottom-up approach splits the data with a grid, tests the density of regions, and extract relevant dimensions (e.g. CLIQUE [3], ENCLUS [4], MAFIA [11]). In algorithms mentioned above the weights of dimensions in clusters is either 0 or 1 (hard weights).

The crisp (hard) weights of attributes may not be satisfactory in some applications. Algorithms that elaborate fuzzy (soft, nonbinary) weights for attributes are an interesting direction of research. Some algorithm have been recently

E-mail address: Krzysztof.Siminski@polsl.pl.

http://dx.doi.org/10.1016/j.fss.2017.01.001

Table 1
Symbols used in the paper.

Symbol	Meaning
$\mathbb{C}$	set of clusters
C	number of clusters, $C = \|\mathbb{C}\|$
$\mathbb{X}$	set of data items, $\mathbb{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_X\}$
X	number of data items, $X = \|\mathbb{X}\|$
$\mathbf{x}$	data item, $\mathbf{x} = [x_1, \ldots, x_D]$
D	number of attributes
$\mathbf{U}$	membership matrix $[C \times X]$
u_{cx}	membership of the xth item to the cth cluster
$\mathbf{V}$	matrix of cluster centres, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_C]^{\mathrm{T}}$
$\mathbf{v}_c$	centre of cth cluster, prototype, $\mathbf{v}_c = [v_{c1}, \ldots, v_{cD}]$
$\mathbf{Z}$	weight matrix $[C \times D]$
z_{cd}	weight of dth attribute in cth cluster
$\hat{z}_{cd}$	augmented weight of dth attribute in cth cluster, cf. Eq. (27)
β_{ck}	typicality of the kth data item with respect to cth cluster
f_k	global typicality of the k data item, cf. Eq. (29)
m	weighting exponent for memberships
ϕ	weighting exponent for weights
h	loss function, cf. Eq. (31)–(37)
e_{cdk}	residual of dth of kth datum from the centre of cth cluster
$\diamond$	s-norm, cf. Eq. (29)
$\varkappa$	data item's ordinal number, cf. Eq. (22)

proposed [10,8,17,19,9,20]. Subspace FCM (fuzzy weighted C-means) is an algorithm that assigns weights (from interval [0, 1]) to attributes [18]. The algorithm is based on minimisation of a criterion function and calculates both fuzzy membership of data item to clusters and fuzzy weights for attributes in clusters. Automatic feature grouping fuzzy k-means (AFGFKM) algorithm groups attributes and assigns weights to the groups [9]. The membership of an data item to the group is crisp. The attribute weights in groups are soft.

One of the problems of the soft weight clustering algorithms is their vulnerability to noise and outliers [9]. The papers [12,13] present an ordering technique in clustering. The fuzzy C-ordered means algorithm (FCOM) [13] does not assign weights to attributes, but calculates typicality of each data item. The data items are ordered and their typicality is updated in each iteration of the clustering procedure. The distant items from all prototype centres have lower weights. Outliers and noise data items are assigned low typicality and do not distort the clustering process. Both algorithms are robust to outliers and even their high ratio does not distort the clustering results severely [12,13].

In the paper we propose a new fuzzy weighted C-ordered-means clustering algorithm. The algorithm finds clusters in fuzzy subspaces of the original task space. The algorithm assigns weights to dimensions (attributes) in each cluster. The weights are numbers from unit interval [0, 1]. To make this algorithm more robust to outliers and noise it incorporates the ordering technique.

In the paper we follow the general rule for symbols: the blackboard bold uppercase characters ($\mathbb{A}$) are used to denote the sets, uppercase italics (A) – the cardinality of sets, uppercase bolds ($\mathbf{A}$) matrices, lowercase bolds ($\mathbf{a}$) vectors, lowercase italics (a) scalars and set elements. Table 1 lists symbols used in the paper.

The paper is organised as follows: The novel fuzzy weighed c-ordered clustering algorithm is described in Sec. 2. The numerical experiments are presented in Sec. 3. Finally Sec. 4 summaries the paper.

2. Fuzzy weighted C-ordered-means clustering algorithm

The clustering algorithm minimises a criterion function J defined as:

$$J(\mathbf{U}, \mathbf{V}, \mathbf{Z}) = \sum_{c=1}^{C} \sum_{i=1}^{X} \beta_{ci} \, (u_{ci})^m \sum_{d=1}^{D} (z_{cd})^{\phi} \, (x_{id} - v_{cd})^2 , \tag{1}$$

where $\mathbf{U}$ is a $C \times X$ membership matrix (for C clusters and X data items) whose each element $u_{cx} \in [0, 1]$ denotes the membership grade of the xth item to the cth cluster (group); $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_C]^{\mathrm{T}} \in \mathbb{R}^{C \times D}$ is a matrix of prototypes

Please cite this article in press as: K. Siminski, Fuzzy weighted C-ordered means clustering algorithm, Fuzzy Sets Syst. (2017), http://dx.doi.org/10.1016/j.fss.2017.01.001