Engineering Applications of Artificial Intelligence 46 (2015) 196–208

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Active learning for improving a soft subspace clustering algorithm

Amel Boulemnadjel [a,*], Fella Hachouf [a], Amel Hebboul [b], Khalifa Djemal [c]

[a] Laboratoire d'Automatique et de Robotique, Département d'Électronique, Faculté des sciences de l'ingénieur, Université des Freres Mentouri Constantine, Route d'Ain el bey, 25000 Constantine, Algeria
[b] Département des Sciences Exactes et Informatique, Ecole Normale Supérieure de Constantine, Ali Mendjli, Constantine 3, Algeria
[c] IBISC Laboratory, University Evry val D'Essonnes, 40 Pelvoux Street, 91080 EVRY Courcouronnes Cedex, France

ARTICLE INFO

Article history:
Received 26 October 2014
Received in revised form
4 July 2015
Accepted 6 August 2015

Keywords:
Subspace clustering
Density
Active learning
SVM

ABSTRACT

In this paper a new soft subspace clustering algorithm is proposed. It is an iterative algorithm based on the minimization of a new objective function. The classification approach is developed by acting at three essential points. The first one is related to an initialization step; we suggest to use a multi-class support vector machine (SVM) for improving the initial classification parameters. The second point is based on the new objective function. It is formed by a separation term and compactness ones. The density of clusters is introduced in the last term to yield different cluster shapes. The third and the most important point consists in an active learning with SVM incorporated in the classification process. It allows a good estimation of the centers and the membership degrees and a speed convergence of the proposed algorithm. The developed approach has been tested to classify different synthetic datasets and real images databases. Several indices of performance have been used to demonstrate the superiority of the proposed method. Experimental results have corroborated the effectiveness of the proposed method in terms of good quality and optimized runtime.

1. Introduction

Clustering problem seeks to partition a given dataset into clusters. Objects in a same cluster have high similarity but they are very dissimilar to objects in other clusters according to a similarity measure. In high dimensional datasets, objects representation varies from one dimension to another. Therefore, it is difficult to find a single notion of similarity to group all objects. Related classical algorithms are inefficient for classification data in different subspaces. They suffer from the curse of dimensionality and the similarity functions that use all input features with equal relevance and may not be effective. Dimensionality reduction methods such as feature transformation and selection (Hu et al., 2007) have been proposed to solve this type of problem. The bi-clustering algorithms (Madeira and Oliveira, 2004; Busygin et al., 2008) use a dissimilarity measure between the rows and the columns. These methods are mainly applied to cluster binary data. In data mining, dataset have hundreds or thousands of categorical attributes. In this case, the bi-clustering algorithms need a large number of binary attributes. This will inevitably increase the computation memory space. This problem led researchers to find other methods that detect different clusters in different subspaces. Subspace clustering is an extension of the traditional clustering (Parsons et al., 2004), with the goal of finding clusters that form on different subspaces. Aggrawal et al. (1999a) were the first to successfully introduce a methodology for locating different subspaces in different clusters. Subspaces clustering methods are divided into two broad categories: hard subspace clustering (Parsons et al., 2004) and soft subspace clustering (Deng et al., 2010; Xia et al., 2013; Wang et al., 2013). In the hard methods, each point of the dataset can belong to only one category. The first example of bottom-up methods was Clique algorithm while Proclus (Aggarwal et al., 1999b) was the first top-down one. In soft subspace clustering methods, each point of the dataset does not belong fully to one class but it has different degrees of membership in several classes. Weights are assigned to each feature to measure its contribution to build a particular cluster. In recent years, soft subspace clustering has emerged as an important research topic and many algorithms have been developed. Michele (2008) proposed a semi-supervised method. A metric learning approach is used to improve the classical fuzzy *C*-means (Bezdek, 1981). This method is based on two steps. In the first step, a series of different metrics is learnt based on the data. In the second one, the fuzzy *C*-means with the previously computed metric is executed. Wu et al. (2005) employed two important informations; between and separation clusters to develop a fuzzy compactness and separation algorithm (FCS). Liang et al. (2011) proposed a clustering method for high

* Corresponding author.
E-mail addresses: amel.boulemnadjel@yahoo.fr (A. Boulemnadjel), hachouf.fella@gmail.com (F. Hachouf), ahebboul@yahoo.fr (A. Hebboul), djemal@iup.univ-evry.fr (K. Djemal).

http://dx.doi.org/10.1016/j.engappai.2015.08.005
0952-1976/

dimensionality data based on the *k*-modes algorithm. Two types of weights are introduced to identify important attributes and delete non-significant ones. Feature groups *k*-means (FG-*k*-means) is an algorithm proposed in Xiaojun et al. (2012). Two terms have been added to compute two types of weight attributes. Then, they are introduced to simultaneously identify the importance of groups and individual features in categorizing each cluster. Shortcomings of these methods are in the initialization step which makes unstable the final results. Soft subspace clustering has been enhanced by introducing the concept of entropy weights tuning (Anil Kumar et al., 2010; Domeniconi et al., 2007; Friedman and Meulman, 2004). LAC (Domeniconi et al., 2007) and COSA (Friedman and Meulman, 2004) associate to each cluster a weight vector. Both of them have been developed on an exponential weighting scheme but they are fundamentally different in their search strategies and outputs. Only inter-cluster dissimilarity is used in the objective function. In Deng et al. (2010) the objective function of an enhanced soft subspace clustering (ESSC) algorithm is based on two important informations; within-cluster compactness and between-cluster separation. It suffers from the initialization step inducing instability in the results. The effectiveness of this type of algorithm drops if the clusters have different densities and shapes. Better results are obtained with methods based on the cluster density (Sunita and Parag, 2009; Sembiring and Zain, 2010). Among these methods, DBSCAN (Ester et al., 1996) is the oldest. It is based on the density of reachability and connectability. Two parameters are selected to initialize the clustering process; neighborhood size and minimum density of clusters. This method presents the problem of parameters selection and computational complexity. An improved version of this method is given in Damodar and Prasanta (2012). The *K*-means algorithm is used to perform an automatic prototype selection.

The paper is organized as follows: in Section 3, the proposed approach built around an active learning for improving the soft subspace clustering algorithm (ALISSC) is presented. Section 4 is dedicated to the experimental results and their discussion; cluster analysis and technical validation of the proposed method are developed. Finally, in Section 5, a conclusion and perspectives to improve the proposed work are given.

2. Motivations

Classification methods based on optimization models (Benaichouche et al., 2013) are iterative methods that need an initialization step. An arbitrary initialization induces more computing time and influences the stability of the results. The cluster centers are poorly localized. Indeed, errors in estimating distances between centers are amplified through the iterations, inducing a classification low rate and a high processing time. The complexity of the objective function makes difficult to find the local minima. To overcome these shortcomings, a new soft subspace clustering approach is proposed. It is based on a new objective function. It contains two terms: weighting within cluster compactness and weighting between cluster separation. The weighting within cluster compactness term is based on the local variance and the clusters density. In this term, the weights are assigned to features according to the local variance of data along each dimension. They are computed using a new and simple formulation based on the local entropy for each feature. The learned weights perform a directional local reshaping of distances. Hence, they allow a better clusters separation and therefore the discovery of new patterns in different subspaces. An initial classification step is added by using a support vector machine (SVM) algorithm (Vapnik, 1999; Sangeetha et al., 2011; Langone et al., 2015), to generate the initial set of cluster centers and membership degrees. Clusters density is introduced to improve the efficiency separation and to yield different clusters shapes. To take the advantages of the proposed objective function and the multi-class SVM, a combination of these two concepts is achieved. A novel multi-label strategy based on active learning (Jain and Kapoor, 2009; Hua Ho et al., 2011) is proposed to accelerate the convergence of the SVM and to estimate the cluster centers (Fig. 1). In the processing steps all used parameters are automatically selected.

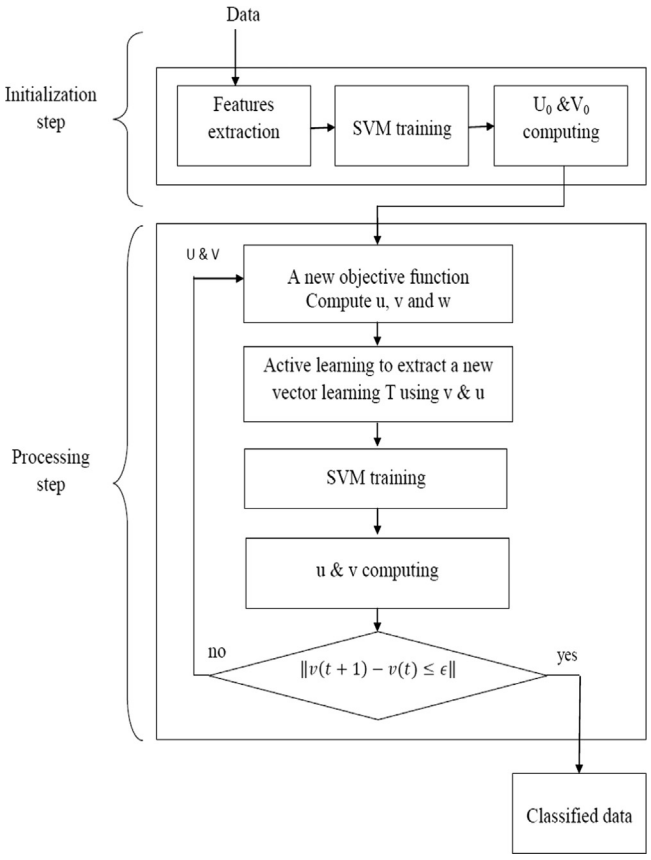

Fig. 1. Flowchart of the proposed method.

3. Proposed method

3.1. Nomenclature

To improve the readability of the equations, the following notations are used:

ALISSC	Active learning to improve a soft subspace clustering
ESSC	Enhanced soft subspace clustering
FSC	Fuzzy compactness and separation
J_{ALISSC}	Proposed objective function
v	Cluster center
u	Membership degree
m	Fuzzy parameter
c	Clusters number
N	Data size
D	Features number (subspaces)
w	Weight matrix
x	Data
n_{ik}	Cluster density in the ith cluster and kth subspace
η	Parameter controlling the influence of the weights in clusters separation