Expert Systems With Applications 85 (2017) 146–157

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Semi-supervised model-based clustering with controlled clusters leakage

Marek Śmieja*, Łukasz Struski, Jacek Tabor

Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, 30-348 Krakow, Poland

ARTICLE INFO

Article history:
Received 12 December 2016
Revised 28 April 2017
Accepted 13 May 2017
Available online 17 May 2017

Keywords:
Semi-supervised clustering
Projection pursuit
Model-based clustering
Cross-entropy
Constrained Gaussian mixture model

ABSTRACT

In this paper, we focus on finding clusters in partially categorized data sets. We propose a semi-supervised version of Gaussian mixture model, called C3L, which retrieves natural subgroups of given categories. In contrast to other semi-supervised models, C3L is parametrized by user-defined leakage level, which controls maximal inconsistency between initial categorization and resulting clustering. Our method can be implemented as a module in practical expert systems to detect clusters, which combine expert knowledge with true distribution of data. Moreover, it can be used for improving the results of less flexible clustering techniques, such as projection pursuit clustering. The paper presents extensive theoretical analysis of the model and fast algorithm for its efficient optimization. Experimental results show that C3L finds high quality clustering model, which can be applied in discovering meaningful groups in partially classified data.

1. Introduction

Model-based clustering aims at finding a mixture of probability models, which optimally estimates true probability distribution on data space. Contrary to other clustering techniques, it does not only recover meaningful groups, but also gives a rule (probability model) for generating elements from clusters. Therefore, it is commonly used in various areas of machine learning and data analysis (Salah, Rogovschi, & Nadif, 2016; Spurek, 2017; Wehrens, Buydens, Fraley, & Raftery, 2004).

Although clustering is an unsupervised technique, one can introduce additional information to guide the algorithm what is the expected structure of clusters. Semi-supervised learning methods usually use partial labeling (Liu & Fu, 2015) or pairwise constraints (Lu & Leen, 2007) to transfer expert knowledge into clustering process, while consensus and alternative clustering gather information from several partitions of data into one general view (Gondek & Hofmann, 2007; Nguyen, 2007). In this paper, we assume that we have the knowledge about division of data set into two categories and focus on the following problem: *How to find the best model of clusters that preserves a fixed amount of information about existing categories?* In other words, we focus on finding interesting clusters, which are very likely to belong to one category.

To explain a basic motivation behind our model, let us consider an expert system used for automatic text translation. It is a common practice to construct several translation models, each designed for one cluster retrieved from a data set (Aggarwal & Zhai, 2012). Alternatively, since texts are often categorized into specific domains, e.g. sport, politics, etc., then each translator can be fitted to one of these categories. To consider together both options, we could implement a separate module responsible for finding clusters, which (a) are described by compact models (e.g. Gaussians) and (b) are related with predefined topics. Observe that optimization of these two conflicting goals simultaneously is non-trivial. We cannot cluster elements from each category individually, because this strategy does not lead to optimal solution for the entire data set (in terms of likelihood). Moreover, existing categorization might be inaccurate as well as the interesting groups can cross the boundary between predefined domains. Therefore, a better approach is to incorporate the constraint to the clustering process and always work with the entire data set.

Our method can also be applied to strictly unsupervised situations, where no initial categorization is given. Let us recall that one way to analyze clusters in complex data spaces relies on finding projections onto one dimensional subspaces, where groups can be easily identified. Projection pursuit focuses on choosing such a direction, which optimizes selected statistical index such as kurtosis (Peña & Prieto, 2001) or skewness (Loperfido, 2013). Since one dimensional views generate linear decision boundaries in original data space, it is not possible to find flexible cluster structures.

* Corresponding author.
E-mail addresses: marek.smieja@ii.uj.edu.pl (M. Śmieja), lukasz.struski@uj.edu.pl (Ł. Struski), jacek.tabor@ii.uj.edu.pl (J. Tabor).

http://dx.doi.org/10.1016/j.eswa.2017.05.032
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

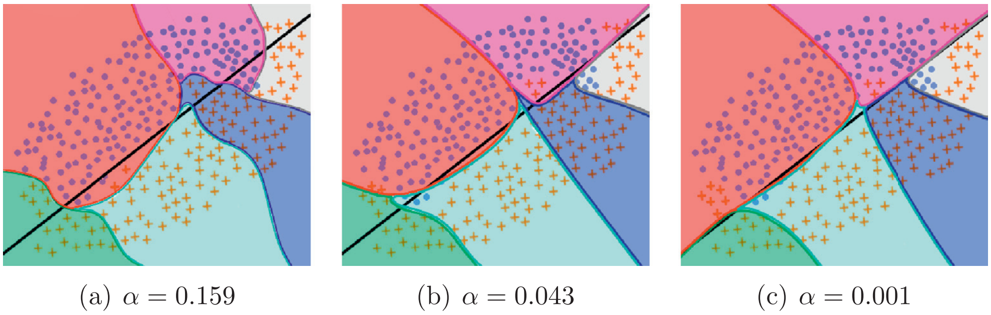
(a) $\alpha = 0.159$ (b) $\alpha = 0.043$ (c) $\alpha = 0.001$

Fig. 1. The effects of C3L for different values of the leakage level α.

However, we can input such a linear boundary to our model in order to improve existing clusters. Our method directly uses the information from initial splitting, but can extend linear decision surfaces to nonlinear ones generated by probabilistic mixture models.

Following the above motivation, we propose a semi-supervised clustering with controlled clusters leakage model (C3L), which integrates a distribution of data with a fixed division of the space into two categories. C3L focuses on finding a type of Gaussian mixture model (GMM) (McLachlan & Peel, 2004), which maximizes the likelihood function and preserves the information contained in the initial splitting with a predefined probability (leakage level). Intuitively, we allow for the flow of clusters densities over decision surface, but with a full control of total probability assigned to the opposite category, which is defined as the leakage level $\alpha \in (0, 1)$ (see Fig. 1). This general idea is formulated as a constrained optimization problem (Section 3).

The advantages of C3L can be summarized as follows:

1. It has a closed form solution in a special case of cross-entropy clustering (a type of GMM) (Tabor & Spurek, 2014).
2. It can be efficiently implemented and optimized by a modified on-line Hartigan algorithm (Section 4).
3. The user can directly parametrize C3L by a maximal inconsistency level between initial categorization and final clustering model (leakage level).
4. The selection of the leakage level α allows to move from a strictly unsupervised GMM for $\alpha = 0.5$, where decision boundary has no effect on clustering, to the limiting case of $\alpha \to 0$, where every group is fully condensed in one category (Section 5).

Experimental studies confirm that the proposed approach builds a high quality model under a given constraint in terms of inner clustering measures, such as Bayesian Information Criterion (Section 6.1). It can be successfully used to discover meaningful groups in partially classified data (Section 6.2) as well as to improve existing clusters obtained by applying projection techniques (Section 6.3). We present a real-life case study, in which the use of C3L allows to detect subgroups of chemical space given their division into active and inactive classes (Section 6.4).

2. Related work

Semi-supervised clustering incorporates the knowledge about class labels to partitioning process (Basu, Davidson, & Wagstaff, 2008). This information can be presented as partial labeling, which gives a division of a small portion of data into categories, or as pairwise constraints, which indicate whether two data points originate from the same (must-links) or distinct classes (cannot-links). Although pairwise constraints provide less amount of information than partial labeling, it is easier to assess whether two instances come from the same group than assign them to particular classes.

Clustering with pairwise constraints was introduced by Wagstaff, Cardie, Rogers, Schrödl et al. (2001), who created a variant of k-means, which focuses on preserving all constraints. Shental, Bar-hillel, Hertz, and Weinshall (2004) constructed a version of Gaussian mixture model, which gathers data points into equivalence classes (called chunklets) using must-link relation and then applied EM algorithm on such generalized data set of chunklets. This approach was later modified to multi-modal clustering models (Śmieja & Wiercioch, 2016). The aforementioned methods work well with noiseless side information, but deteriorate the results when some constraints are mislabeled. To overcome this problem, the authors of (Basu, Bilenko, & Mooney, 2004; Lu & Leen, 2005) applied hidden Markov random fields (HMRF) to construct more sophisticated dependencies between linked points. However, the use of HMRF leads to complex solutions, which are difficult to optimize. In recent years, Asafi and Cohen-Or (2013) suggested reducing distances between data points with a must-link constraint and adding a dimension for each cannot-link constraint. After updating all other distances to, e.g., satisfy the triangle inequality, the thus obtained pairwise distance matrix can be used for unsupervised learning. Wang and Davidson (2010) proposed a version of spectral clustering, which relies on solving a generalized eigenvalue problem.

Partial labeling is used in clustering to define sample data points from particular classes. Liu and Fu (2015) added additional attributes to feature vectors and proposed modified k-means algorithm. There is also a semi-supervised version of fuzzy c-means (Pedrycz, Amato, Di Lecce, & Piuri, 2008; Pedrycz & Waletzky, 1997), where the authors supplied the cost function with a regularization term that penalizes fuzzy partitions that are inconsistent with the side information. GMMs can be adapted to make use of class labels by combining the classical unsupervised GMM with a supervised one (Ambroise, Denoeux, Govaert, & Smets, 2001; Zhu & Goldberg, 2009).

Since assigning data points to classes or labeling pairwise constraints requires extensive domain knowledge, then many clustering methods were adapted to use additional information about data, which does not require human intervention. One example is consensus clustering, which considers gathering information coming from different domains (Nguyen, 2007). On the other hand, complementary (alternative) clustering aims at finding groups which provide a perspective on the data that expands on what can be inferred from previous partitions (Gondek & Hofmann, 2007).

C3L is a version of Gaussian mixture model, which uses side information given by class labels or more generally by a decision boundary between classes. In contrast to classical methods applying partial labeling, it focuses on finding subgroups of original classes. This goal is similar to information bottleneck method (Chechik, Globerson, Tishby, & Weiss, 2005; Tishby, Pereira, & Bialek, 1999). Roughly speaking, this approach tries to construct compact clusters (compressed representation), which contain high