Pattern Recognition Letters 37 (2014) 85–93

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Context-sensitive intra-class clustering

Yingwei Yu [a,*], Ricardo Gutierrez-Osuna [b], Yoonsuck Choe [b]

[a] IHS Global, Inc., 8584 Katy Freeway, Suite 400, Houston, TX 77024, USA
[b] Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843-3112, USA

ARTICLE INFO

Article history:
Available online 10 May 2013

Keywords:
Clustering
Intra-class clustering
ICC
Linear discriminant analysis
LDA
Semi-supervised learning

ABSTRACT

This paper describes a new semi-supervised learning algorithm for intra-class clustering (ICC). ICC partitions each class into sub-classes in order to minimize overlap across clusters from different classes. This is achieved by allowing partitioning of a certain class to be assisted by data points from other classes in a context-dependent fashion. The result is that overlap across sub-classes (both within- and across class) is greatly reduced. ICC is particularly useful when combined with algorithms that assume that each class has a unimodal Gaussian distribution (e.g., Linear Discriminant Analysis (LDA), quadratic classifiers), an assumption that is not always true in many real-world situations. ICC can help partition non-Gaussian, multimodal distributions to overcome such a problem. In this sense, ICC works as a preprocessor. Experiments with our ICC algorithm on synthetic data sets and real-world data sets indicated that it can significantly improve the performance of LDA and quadratic classifiers. We expect our approach to be applicable to a broader class of pattern recognition problems where class-conditional densities are significantly non-Gaussian or multi-modal.

1. Introduction

Many existing clustering methods based on unsupervised learning can partition data into clusters using the distribution of data points and the relative distance between the data points. A small amount of class information can help refine such clustering results by providing contextual information in a semi-supervised manner. In such a semi-supervised clustering situation, not only the distance among data points within the object class (e.g., targets to be recognized), but also the position of the context class (e.g., backgrounds or confounders) relative to the object class becomes important. As an example, the context class can serve as a boundary within the object class, and therefore influence how to best subdivide the object class.

An example of how a context class can affect the partitioning is shown in Fig. 1. Data points for the object class marked "+" were generated from a single Gaussian distribution (see Fig. 1 caption for details). Without the context data the distribution is best grouped into a single cluster as shown by the ellipse in Fig. 1(a). However, in the presence of intervening context data (marked "○"), a different strategy is needed (Fig. 1(b)). Namely, when the context class is present, the object class can be split into two clusters: one to the lower left and the other to the upper right. The intra-class clustering algorithm in this paper is designed specifically to account for the presence of such intervening context classes.

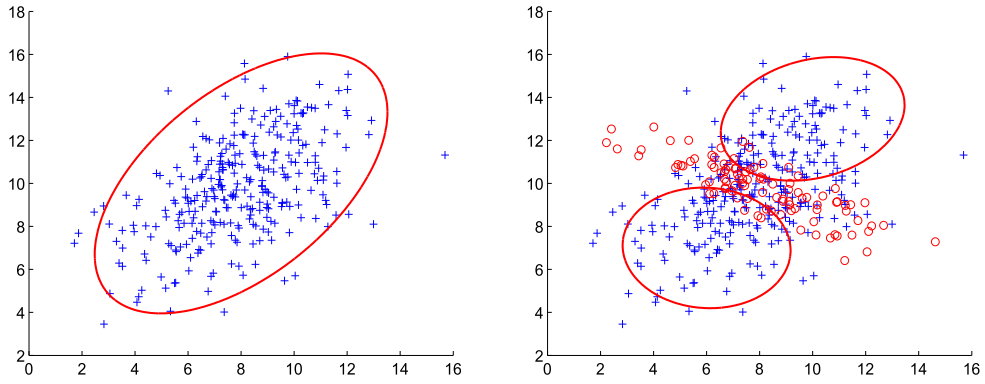
Unlike other clustering algorithms that partition data into clusters exclusively based on intrinsic information (Jain, 2010), our context-sensitive clustering method operates in a *semi-supervised* mode by utilizing external information from context classes in addition to intrinsic information from the object class. As defined by Chapelle et al. (2006), "semi-supervised" clustering algorithms make use of external information, or "side-information". This semi-supervised learning method of context-sensitive clustering can lead to improvements in classification performance, as we will see shortly.

As an example, Fisher's linear discriminant analysis (LDA) (Duda et al., 2001) can find an optimal projection to discriminate data from different classes under the assumption that each class is a unimodal Gaussian with the same covariance matrix. However, LDA can significantly underperform in two situations (Zhao, 2000): (1) when the class discriminant information is in the variance of the data set, not just in the mean (Wang et al., 2004) or (2) when the class data is markedly non-Gaussian. For the first problem, one approach is to use the combined distribution of all other classes to split the object class in a context-sensitive manner. For the second problem we can require that the intra-class clustering method generates unimodal sub-clusters that are Gaussian.

Expectation Maximization (EM) (Dempster et al., 1977) based algorithms can divide non-Gaussian data into unimodal clusters, but EM is not context-sensitive, so the resulting clusters may have overlapping distributions. This problem is related to the first

* Corresponding author. Tel.: +1 281 844 4955.
E-mail addresses: yingweiy@gmail.com (Y. Yu), rgutier@cs.tamu.edu (R. Gutierrez-Osuna), choe@cs.tamu.edu (Y. Choe).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.04.031



(a) Single Gaussian distribution (b) Two clusters with context data

Fig. 1. Illustration of how object class and context class can interact. (a) The data points of the object class (marked "+") in the plot can be best modeled by a single Gaussian (mean at $(8, 10)$, with covariance $\begin{bmatrix} 10 & 30 \\ 30 & 30 \end{bmatrix}$) when no context data are present. (b) Data in the object class (marked "+") are those in (a), but whenthe context data are present (marked "$\circ$") with mean at $(8, 10)$, and covariance $\begin{bmatrix} 6 & -3 \\ -3 & 2 \end{bmatrix}$, the object class splits into two sub-clusters.

problem of LDA we discussed earlier. For example, the illustration in Fig. 2(a) shows data points from two Gaussian distributions (see figure caption for details). When EM is applied blindly, three of the resulting clusters significantly overlap. A better approach would be one that minimizes overlap among the clusters (see Fig. 2(b)).

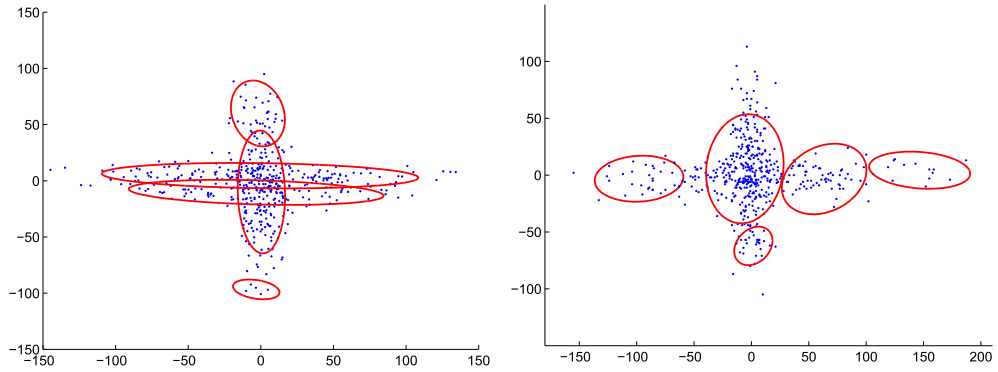
To address these issues, this paper proposes an intra-class clustering algorithm that can generate non-overlapping unimodal clusters (e.g., as those in Fig. 2(b)). To achieve this, we propose a new semi-supervised learning algorithm called context-sensitive intra-class clustering algorithm (ICC). ICC can be used in unsupervised mode to cluster data as shown in Fig. 2(b) when there is no context class. Or, in semi-supervised mode, it can identify unimodal sub-clusters that reduce overlap within and across classes. The separation of data into single unimodal clusters makes the resulting distribution suitable for LDA as well as for low-cost Gaussian classifiers (e.g., the quadratic classifier). Our context-sensitive clustering algorithm is novel compared to other clustering algorithms in that it can not only cluster the object data based on the distance between samples, but also take into account the intervening nature of the context class.

The rest of the paper is organized as follows. In Section 2, we describe in detail our intra-class clustering algorithm. Section 3 presents experimental results of the proposed method on synthetic data and real-world data. Finally, Section 4 presents a brief discussion of our algorithm, followed by the conclusion.

2. Proposed algorithm: intra-class clustering

In most pattern recognition problems, the input space is very high dimensional which leads to serious problems due to curse of dimensionality as well as high computational cost. These issues can in part be overcome by applying dimensionality reduction algorithms. However, low-dimensional projections of the data are not guaranteed to minimize the overlap among different classes. If such overlap in low dimensional mapping can be alleviated by breaking down the classes into non-overlapping sub-classes, a highly efficient and accurate algorithm can be derived.

Our method starts by projecting the data in 2D space (or in 3D). Next, from these projections, density maps are generated for both the object class and the context class. Finally, the difference of the densities is calculated. A number of dimensionality reduction



(a) EM clustering (b) Intra-class clustering

Fig. 2. EM vs. ICC in unsupervised mode. (a) The data are generated by two overlapping Gaussian distributions where both of their means are at $(0, 0)$. The first Gaussian has covariance $\begin{bmatrix} 10 & 0 \\ 0 & 40 \end{bmatrix}$, while the second one $\begin{bmatrix} 60 & 0 \\ 0 & 10 \end{bmatrix}$. Each Gaussian has 500 data points. Assume that we do not know the true number of underlying distributions and guessed that there are five. Using the EM algorithm, we can fit five Gaussian distributions to the data. There is obvious overlap between the resulting distributions near the center. (b) The intra-class clustering method clusters the data into five non-overlapping unimodal Gaussians. This method ensures that the resulting clusters are unimodal and reduces the unnecessary overlap between the clusters.