Pattern Recognition Letters 80 (2016) 52–57

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A proposal for supervised clustering with Dirichlet Process using labels☆

Billy Peralta[a,*], Alberto Caro[a], Alvaro Soto[b]

[a] Catholic University of Temuco, Manuel Montt 056, Temuco, Chile
[b] Pontifical Catholic University of Chile, Av. Vicuna Mackenna 4860, Santiago, Chile

ARTICLE INFO

Article history:
Received 2 August 2015
Available online 24 May 2016

Keywords:
Dirichlet Process
Supervised clustering
Clustering

ABSTRACT

Supervised clustering is an emerging area of machine learning, where the goal is to find class-uniform clusters. However, typical state-of-the-art algorithms use a fixed number of clusters. In this work, we propose a variation of a non-parametric Bayesian modeling for supervised clustering. Our approach consists of modeling the clusters as a mixture of Gaussians with the constraint of encouraging clusters of points with the same label. In order to estimate the number of clusters, we assume a-priori a countably infinite number of clusters using a variation of Dirichlet Process model over the prior distribution. In our experiments, we show that our technique typically outperforms the results of other clustering techniques.

1. Introduction

The problem of finding K groups from a set of N points is known as the clustering problem. Clustering is usually used in an unsupervised learning framework using typical error functions, e.g. a minimization of the intra-cluster distance. Many types of clustering algorithms have been proposed as partitional, agglomerative [10], spectral [12], model-based [9], and subspace clustering [17]. Model-based clustering techniques have the advantage of being based on a solid framework that facilitates clustering analysis. A typical clustering model clustering is the mixture of Gaussians due to its flexibility and mathematical soundness [3]. Nonetheless, this technique has the drawback that it requires a fixed number of clusters [19].

A typical problem for model-based clustering algorithms is to choose the number of clusters, which is typically an user-defined parameter [10]. Nonetheless, the number of clusters is usually an unknown variable because the problem context is not sufficiently understood or there are multiple valid options. A typical approach for performing clustering without knowing the number of clusters is using non-parametric Bayesian models which typically consider a Dirichlet Process as a prior distribution [15]. The extension of Gaussian mixtures models to non-parametric Bayesian setting is the Infinite Gaussian Mixture models [19].

In contrast to classical clustering, supervised clustering assumes that the dataset is labeled and has the goal of finding clusters with a high purity, where the purity of a cluster is defined as the percentage of data in a cluster that belongs to its most frequent class objective. Moreover, supervised clustering has the constraint of keeping the number of clusters as small as possible; in this way, it avoids having many clusters as points. The benefits of supervised clustering are the following: improving the quality of clusters, dataset compression, and enhancement of classification algorithms [8].

An example of application is the clustering of profile customers according to continuous measures (for instance: age, height, coordinate residence, etc.) into clusters that are discriminative in regard to the buying behavior of the customers across product categories, which would be the label. Other potential uses would include identifying patterns in genetics and finances [20]. It is still applied to computer vision tasks such as building of visual dictionaries by replacing the typical K-Means algorithm [18].

In this work, we present a heuristic variant of the Infinite Gaussian Mixture model for supervised clustering. Our idea is based on the hypothesis that in a wide variety of applications a combination of supervised and unsupervised information can lead us to more informative clusters because they are complementary. The proposed variant is based on modifying the Dirichlet Process model for prior distribution of partitions, considering the label information. We use the Pólya urn interpretation of the Dirichlet Process model as our model, with the intuition is that if we have two clusters with similar distance, a new recsord with label l should prefer the cluster that has the majority of points with label l. The main

☆ This paper has been recommended for acceptance by Maria De Marsico.

* Corresponding author. Tel.: +56 45 2205205; fax: +56 45 2211034.
E-mail address: bperalta@uct.cl, billymark@gmail.com, bmperalt@uc.cl (B. Peralta).

http://dx.doi.org/10.1016/j.patrec.2016.05.019
0167-8655/

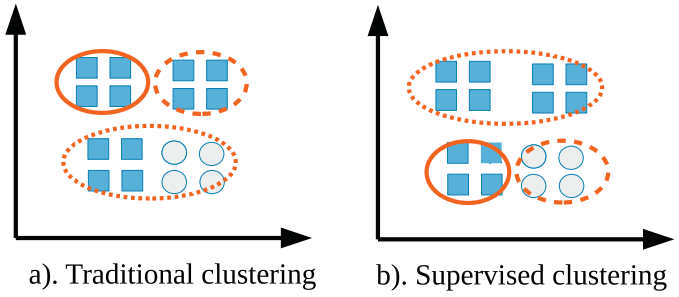

Fig. 1. Toy example comparing traditional and supervised clustering results. There are two class labels denoted by dark squares and gray circles. The results are different, while classical clustering considers distance between points, supervised clustering prioritizes clusters that have the same class labels.

contributions of this paper are: i) Presenting the Labeled Dirichlet Process model (LDPM), a variant of the classical non-parametric Dirichler Process model (DPM) that extends those by incorporating label information, and ii) Empirical evidence showing that LDPM outperforms DPM in terms of quality clustering based on label information and competitiveness with specialized supervised clustering algorithms (Fig. 1).

This paper is organized as follows: Section 2 describes previous works and background information about our technique. Section 3 presents the method proposed in this paper. Section 4 describes the experiments. Finally, Section 5 presents the main conclusions of this work.

2. Background and related work

2.1. Related work

Some researchs of semi-supervised clustering bear similarity to our approach. The semi-supervised clustering techniques usually consist of improving the clustering results by using some labeled examples. These algorithms generally aim to maximize the purity of clusters, where the purity is the percentage of the points labeled with the label mode. The semi-supervised clustering methods can be divided into two groups: similarity-based methods and search-based methods [2]. Similarity-based methods use a traditional clustering algorithm to group the data considering a modified distance function constrained by the data labels. On the other hand, search-based methods modify the clustering algorithm itself but do not change the distance function.

Tishby et al. [22] introduce the information bottleneck method which is based on an information theory approach. Their method consists of applying an agglomerative clustering algorithm [23] that minimizes information loss with respect to the conditional distribution $P(C|A)$, where C is the class variable and A is the training dataset.

Embrechts et al. [7] propose a genetic algorithm for a k-means, where the objective of the search process is to get clusters that minimize a combination of the cluster dispersion and cluster impurity. Cohn [5] varies the popular EM algorithm by incorporating similarity and dissimilarity constraints. Basu et al. [2] modify the k-means clustering algorithm to cope with class knowledge.

Sinkkonen et al. [20] propose discriminative clustering, which minimizes the distortion within clusters. Distortion is defined as the loss of mutual information between classes and the clusters caused by representing each cluster with one prototype. This technique seeks to produce clusters that are as internally homogeneous as possible in conditional distributions $p(C|X)$ of the auxiliary label variable, which implies that the clusters tend to belong to a single class.

Jordan et al. [24] and Shental et al. [1] transform the training examples into constraints where the points of different classes have a distance larger than a given bound. Then they derive a modified distance metric that minimizes the distance between points considering such constraints. Finally, they use the K-means clustering algorithm in conjunction with the modified distance function to compute clusters.

Eick et. al. [8] propose the supervised clustering where the idea is to maximize the purity of clusters with the lowest possible number of clusters. They propose supervised versions of some typical clustering algorithms such as CLARANS and PAM. We compare our method with the proposed SRIDHCR algorithm (Single Representative Insertion/Deletion Steepest Decent Hill Climbing with Randomized Restart) because it is shown to have good results in this paper.

Ye et al. [25] present a discriminative version of the K-Means algorithm. Their algorithm solves simultaneously linear discriminant analysis (LDA) and clustering problem using matrix algebra to finally obtain an iterative algorithm. Another advantage of their approach is that it makes a feature transformation by using LDA properties. All of the aforementioned methods have the drawback of requiring the number of clusters. This is precisely the problem on which we focus. Our approach to modeling labeled clusters is based on Dirichlet Process Mixture, therefore we describe this technique before describing our method. We found a similar work by [6] where they model the supervised clustering using a Bayesian approach based on Dirichlet Process, however, we propose an alternative to the classical Dirichlet Process prior.

2.2. Dirichlet Process Mixture

We assume $X = \{x_1, \ldots, x_n\}$ with n independent observations, with dimensionality p arising from a mixture of distributions $F(\theta_i)$, where i is the index of mixture component. The model parameters θ_i are assumed to be independent draws from a prior probability distribution, G, which follows a Dirichlet Process prior. This leads to the following hierarchical mixture model:

$$\begin{aligned} x_i|\theta_i &\sim F(x|\theta_i) \\ \theta_i|G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{1}$$

where α is the concentration parameter, and G_0 is the baseline distribution for the Dirichlet Process prior (DP), such that $E(G) = G_0$. We use the Pólya urn scheme representation of the Dirichlet Process to fit this model [4]. Considering the marginalization over G, θ_i can be written in terms of successive conditional distributions:

$$\theta_i|\theta_{-i} \sim \frac{1}{n-1+\alpha}\sum_{k\neq i}\delta(\theta_k) + \frac{\alpha}{n-1+\alpha}G_0 \tag{2}$$

where $\delta(\theta_k)$ is a point mass distribution at θ_k.

It has been shown that equivalent models can be obtained by taking the limit as $K \to \infty$ of finite mixture models with K clusters [19]. This can be expressed as:

$$\begin{aligned} x_i|c_i, \phi &\sim F(x|\phi_{c_i}) \\ c_i|p &\sim Discrete(p_1, \ldots, p_K) \\ \phi_c &\sim G_0 \\ q &\sim Dir(\alpha/K, \ldots, \alpha/K) \end{aligned} \tag{3}$$

where q are the mixing proportions, c_i is the latent variable that indicates the cluster allocation of sample i, and ϕ_{c_i} corresponds to identical θ_i's. The conditional prior of c_i can be obtained of the integration over q [16], then by taking $K \to \infty$, it is