Expert Systems with Applications 41 (2014) 1218–1235

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Object clustering and recognition using multi-finite mixtures for semantic classes and hierarchy modeling

Taoufik Bdiri [a], Nizar Bouguila [b,*], Djemel Ziou [c]

[a] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G 1T7, Canada
[b] The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC H3G 1T7, Canada
[c] DI, Faculté des Sciences, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada

ARTICLE INFO

Keywords:
Data clustering
Mixture models
Hierarchical models
Semantic clustering
Inverted Dirichlet distribution
Maximum likelihood
Visual objects

ABSTRACT

Model-based approaches have become important tools to model data and infer knowledge. Such approaches are often used for clustering and object recognition which are crucial steps in many applications, including but not limited to, recommendation systems, search engines, cyber security, surveillance and object tracking. Many of these applications have the urgent need to reduce the semantic gap of data representation between the system level and the human being understandable level. Indeed, the low level features extracted to represent a given object can be confusing to machines which cannot differentiate between very similar objects trivially distinguishable by human beings (e.g. apple vs. tomato). In this paper, we propose a novel hierarchical methodology for data representation using a hierarchical mixture model. The proposed approach allows to model a given object class by a set of modes deduced by the system and grouped according to a labeled training data representing the human level semantic. We have used the inverted Dirichlet distribution to build our statistical framework. The proposed approach has been validated using both synthetic data and a challenging application namely visual object clustering and recognition. The presented model is shown to have a flexible hierarchy that can be changed on the fly within costless computational time.

1. Introduction

Available digital data has increased significantly in the recent years with the intensive use of technological devices and Internet. Due to the huge amount of such heterogeneous data, an urgent need has been triggered to automate its analysis and modeling for different purposes and applications. One challenging crucial aspect in data analysis is clustering, a form of unsupervised learning, which is defined as the process of assigning observations sharing similar characteristics to subgroups, such that their heterogeneity is minimized within a given subgroup and maximized between the subgroups. Such an assignment is not trivial especially when we deal with high dimensional data. Indeed, it has been shown that clustering is considered as one of the most important aspects of artificial intelligence and data mining (Jain, Murty, & Flynn, 1999; Fisher, 1996). Given a data set that we need to extract knowledge from it, the ultimate goal is to construct consistent high quality clusters using a computationally inexpensive way. Statistical-based approaches for data clustering have recently become an interesting and attractive research domain with the advancement of computational power that enables researchers to implement complex algorithms and deploy them in real time applications. One major approach based on statistics is model-based clustering using finite mixture models. A finite mixture model can be defined as a weighted sum of probability distributions where each distribution represents the population of a given subgroup. The authors in Fraley and Raftery (2002) traced the use of finite mixture models back to the 1960s and 1970s, citing amongst others, works in Edwards and Cavalli-Sforza (1965), Day (1969) and Binder (1978). Although their use backs at least as far as 1963, it is only in the recent decades that mixture models applications started to cover many fields including, but not limited to, digital image processing and computer vision (Sefidpour & Bouguila, 2012; Stauffer & Grimson, 2000; Allili, Ziou, Bouguila, & Boutemedjet, 2010), social networks (Couronne, Stoica, & Beuscart, 2010; Handcock, Raftery, & Tantrum, 2007; Morinaga & Yamanishi, 2004), medicine (Koestler et al., 2010; Tao, Cheng, & Basu, 2010; Neelon, Swamy, Burgette, & Miranda, 2011; Schlattmann, 2009; Rattanasiri, Böhning, Rojanavipart, & Athipanyakom, 2004), and bioinformatics (Kim, Cho, & Kim, 2010; Meinicke, Ahauer, & Lingner, 2011; Ji, Wu, Liu, Wang, & Coombes, 2005).

The consideration of mixture models is practical for many applications. In many cases, however, the complexity of the observed data may render the use of one single distribution to

* Corresponding author. Tel.: +1 5148482424; fax: +1 5148483171.
E-mail addresses: t_bdiri@encs.concordia.ca (T. Bdiri), nizar.bouguila@concordia.ca (N. Bouguila), djemel.ziou@usherbrooke.ca (D. Ziou).

0957-4174/$ - see front matter
http://dx.doi.org/10.1016/j.eswa.2013.08.005

represent a given class insufficient for inference. Many techniques have been proposed to select the best number of mixture components that best represents the data. Examples include Bayesian inference criterion (BIC) (Schwarz, 1978), minimum description length (MDL) (Rissanen, 1999) and minimum message length (MML) (Wallace, 2005) criteria. These criteria are mainly used in unsupervised algorithms where the system handles data modeling without any intervention during the learning process. This approach has a serious drawback in many applications as the semantic meaning of the mixture modes in the selected model does not necessarily fit with a human comprehensible semantic. Consider for instance an object recognition application where the system has to recognize different objects according to the user need. An MML, BIC or MDL criterion would in most cases consider classes with important visual similarities (e.g., apple and tomato), as being the classes that should be represented by a single mode in the mixture. This is not always the case in real applications, where a human being or an application would need to differentiate between different classes even when they have close visual properties and similarities, thus we talk here about a semantic meaning of the mixture modes. Therefore, the gap between the system representation and the human representation of data is still high when using these methods. An eventual solution is to form a hierarchical model based on some ontology as in Ziou, Hamri, and Boutemedjet (2009), Bakhtiari and Bouguila (2010), Bakhtiari and Bouguila (2011) and Bakhtiari and Bouguila (2012) where the data is grouped into clusters and sub clusters (i.e., tree-structured clustering).Yet, this approach still form the model using the visual similarities and groups data according to the system choice. Furthermore, since the model is based on estimating the model's parameters as the algorithm goes deeper in the hierarchy (the distributions' parameters of the children clusters depend on the parameters of the parents clusters), when changing the hierarchical model, a whole new estimation should take place, which increases the computational cost. The user intervention to build a hierarchical mixture has been introduced in Bishop and Tipping (1998) which developed the concept of hierarchical visualization where the construction of the hierarchical tree proceeds top-down, and for each level, the user decides on the suitable number of models to fit at the next level down. Indeed this interaction, may serve to have an optimal number of clusters for each level according to the user, but it does not permit the user to define any semantic meaning to the clusters or group the clusters as he/she needs. Moreover, the user cannot define any ontological model to the data, and there is a new estimation of the parameters to be calculated at each level when the model goes deeper in the tree.

In this work, we present a novel way to model data and assign a semantic meaning to clusters according to the user needs which can reduce significantly the gap between the system representation and the user level representation. We tackle the challenging problem of object clustering, and recognition of new unseen data in terms of affectation to the appropriate clusters forming the object class. Naturally, the choice of the distribution forming the mixture model is crucial in terms of clustering efficiency and accuracy of the classification of unseen data. Indeed, many works have focused on Gaussian mixture models (GMM) to build their applications such as in Permuter, Francos, and Jermyn (2003), Zivkovic (2004), Yang and Ahuja (1999) and Weiling, Lei, and Yang (2010), but recent researches have shown that it is not appropriate to always assume that data follows a normal distribution. For instance, the works in Boutemedjet, Bouguila, and Ziou (2009), Bouguila, Ziou, and Vaillancourt (2004), Bouguila and Ziou (2006), Bouguila, Ziou, and Hammoud (2009) have considered the Dirichlet and generalized Dirichlet mixture models, to model proportional data, which have been shown to outperform the GMM. We have developed in our previous work the inverted Dirichlet mixture model (IDMM) which has better capabilities than the GMM when modeling positive data that occurs naturally in many real applications (Bdiri & Bouguila, 2012; Bdiri & Bouguila, 2011). Hence, we propose our new methodology using IDMM, although it is noteworthy to bear in mind than any other distribution can be used as the presented framework is general.

The rest of this paper is organized as follows. In Section 2, we present our statistical framework by considering a two-levels hierarchy for ease of representation and understanding of the general methodology. In Section 3, we propose a detailed approach to learn the proposed statistical model. In Section 4, we propose a generalization of our modeling framework to cover many hierarchical levels. Section 5 is devoted to present the experimental results using both synthetic data and a real-life application concerning object recognition. Finally, Section 6 gives a conclusion and future perspectives for research.

2. Statistical framework: the model

We propose to develop a statistical framework that can model data in a hierarchical fashion. The attribution of a semantic meaning to the model is discussed in sub Section 5.2.1. In this section, we consider a two-levels hierarchy where we have a set of *super classes*, composed each, of a set of classes. The generalization of the model is discussed in Section 4. Let us consider a set $\mathcal{X}$ of N D-dimensional vectors, such that $\mathcal{X} = (\vec{X}_1, \vec{X}_2, \dots, \vec{X}_N)$. Let M denotes the number of different *super classes* and K_j the number of classes forming the *super class* j. We assume that $\mathcal{X}$ is controlled by a mixture of mixtures, such that each *super class* j is represented by a mixture of K_j components and the parent mixture is composed of M mixtures representing the *super classes*. Thus, we consider two views or levels for the statistical model. The first view focuses on the *super classes* and the second one zooms on the classes (see Fig. 1). We suppose that the vectors follow a common but unknown probability density function $p(\vec{X}_n|\Xi)$, where Ξ is the set of its parameters. Let $Z = \{\vec{Z}_1, \vec{Z}_2, \dots, \vec{Z}_N\}$ denotes the missing group indicator, where $\vec{Z}_n = (z_{n1}, z_{n2}, \dots, z_{nM})$ is the label of $\vec{X}_n$, such that $z_{nj} \in \{0,1\}$, $\sum_{j=1}^{M} z_{nj} = 1$ and z_{nj} is equal to one if $\vec{X}_n$ belongs to *super class* j and zero, otherwise. Then, the distribution of $\vec{X}_n$ given the *super class* label $\vec{Z}_n$ is:

$$p(\vec{X}_n|\vec{Z}_n, \Theta) = \prod_{j=1}^{M} p(\vec{X}_n|\theta_j)^{z_{nj}} \quad (1)$$

where $\Theta = \{\theta_1, \theta_2, \dots, \theta_M\}$ and θ_j is the set of parameters of the *super class* j. In practice, $p(\vec{X}_n|\Theta)$ can be obtained by marginalizing the complete likelihood $p(\vec{X}_n, \vec{Z}_n|\Theta)$ over the hidden variables. We define the prior distribution of $\vec{Z}_n$ as follows:

$$p(\vec{Z}_n|\vec{\pi}) = \prod_{j=1}^{M} \pi_j^{z_{nj}} \quad (2)$$

where $\vec{\pi} = (\pi_1, \dots, \pi_M)$, $\pi_j > 0$ and $\sum_{j=1}^{M} \pi_j = 1$, then we have:

$$p(\vec{X}_n, \vec{Z}_n|\Theta, \vec{\pi}) = p(\vec{X}_n|\vec{Z}_n, \Theta)P(\vec{Z}_n|\vec{\pi}) = \prod_{j=1}^{M} \left(p(\vec{X}_n|\theta_j)\pi_j\right)^{z_{nj}} \quad (3)$$

We proceed by the marginalization of Eq. (3) over the hidden variable (see Appendix A), so the first level of our mixture for a given vector $\vec{X}_n$ can be written as follows:

$$p(\vec{X}_n|\Theta, \vec{\pi}) = \sum_{j=1}^{M} p(\vec{X}_n|\theta_j)\pi_j \quad (4)$$

Thus, according to the previous equation, the set of parameters Ξ corresponding to the first level is $\Xi = (\Theta, \vec{\pi})$. When we examine the second level which considers the classes, given that $\vec{X}_n$ is generated from the mixture j, we suppose that it is also generated