Pattern Recognition Letters 37 (2014) 94–106

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

A new interactive semi-supervised clustering model for large image database indexing

Hien Phuong Lai [a,b,c,*], Muriel Visani [a], Alain Boucher [a,b,c], Jean-Marc Ogier [a]

[a] L3I, Université de La Rochelle, Avenue M. Crépeau, 17042 La Rochelle cedex 1, France
[b] IFI, Equipe MSI; IRD, UMI 209 UMMISCO, Institut de la Francophonie pour l'Informatique, 42 Ta Quang Buu, Hanoi, Vietnam
[c] Vietnam National University, Hanoi, Vietnam

ARTICLE INFO

Article history:
Available online 27 June 2013

Keywords:
Semi-supervised clustering
Interactive learning
Image indexing

ABSTRACT

Indexing methods play a very important role in finding information in large image databases. They organize indexed images in order to facilitate, accelerate and improve the results for later retrieval. Alternatively, clustering may be used for structuring the feature space so as to organize the dataset into groups of similar objects without prior knowledge (unsupervised clustering) or with a limited amount of prior knowledge (semi-supervised clustering).

In this paper, we introduce a new interactive semi-supervised clustering model where prior information is integrated via pairwise constraints between images. The proposed method allows users to provide feedback in order to improve the clustering results according to their wishes. Different strategies for deducing pairwise constraints from user feedback were investigated. Our experiments on different image databases (Wang, PascalVoc2006, Caltech101) show that the proposed method outperforms semi-supervised HMRF-kmeans (Basu et al., 2004).

1. Introduction

Content-Based Image Retrieval (CBIR) refers to the process which uses visual information (usually encoded using color, shape, texture feature vectors, etc.) to search for images in the database that correspond to the user's queries. Traditional CBIR systems generally rely on two phases. The first phase is to extract the feature vectors from all the images in the database and to organize them into an efficient index data structure. The second phase is to efficiently search in the indexed feature space to find the most similar images to the query image.

With the development of many large image databases, an exhaustive search is generally intractable. Feature space structuring methods (normally called indexing methods) are therefore necessary for facilitating and accelerating further retrieval. They can be classified into space partitioning methods and data partitioning methods.

Space partitioning methods (KD-tree (Bentley and Sep., 1975), KDB-tree (Robinson, 1981), LSD-tree (Henrich et al., 1989), Grid-File (Nievergelt et al., 1988) etc.) generally divide the feature space into cells (sometimes referred to as "buckets") of fairly similar cardinality (in terms of number of images per cell), without taking into account the distribution of the images in the feature space. Therefore, dissimilar points may be included in a same cell while similar points may end up in different cells. The resulting index is therefore not optimal for retrieval, as the user generally wants to retrieve similar images to the query image. Moreover, these methods are not designed to handle high dimensional data, while image feature vectors commonly count hundreds of elements.

Data partitioning methods (B-tree (Bayer and McCreight, 1972), R-trees (Guttman, 1984; Sellis et al., 1987; Beckmann et al., 1990), SS-tree (White and Jain, 1996), SR-tree (Katayama and Satoh, 1997), X-tree (Berchtold et al., 1996) etc.) also integrate information about image distribution in the feature space. However, the limitations on the cardinality of the space cells remain, causing the resulting index to be non-optimal for retrieval, especially in the case where groups of similar objects are unbalanced, i.e. composed of different numbers of images.

Our claim is that using clustering instead of traditional indexing to organize feature vectors, results in indexes better adapted to high dimensional and unbalanced data. Indeed, clustering aims to split a collection of data into groups (clusters) so that similar objects belong to the same group and dissimilar objects are in different groups, with no constraints on the cluster size. This makes the resulting index better optimized for retrieval. In fact, while in traditional indexing methods it might be difficult to fix the number of objects in each bucket (especially in the case of unbalanced data),

* Corresponding author at: L3I, Université de La Rochelle, Avenue M. Crépeau, 17042 La Rochelle cedex 1, France. Tel.: +33 6 46 51 12 32; fax: +33 5 46 45 82 42.
E-mail addresses: hien_phuong.lai@univ-lr.fr (H.P. Lai), muriel.visani@univ-lr.fr (M. Visani), alainboucher12@gmail.com (A. Boucher), jean-marc.ogier@univ-lr.fr (J.-M. Ogier).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.06.014

clustering methods have no limitation on the cardinality of the clusters, objects can be grouped into clusters of very different sizes. Moreover, using clustering might simplify the relevance feedback task, as the user might interact with a small number of cluster prototypes rather than numerous single images.

Because feature vectors only capture low level information such as color, shape or texture, there is a semantic gap between high-level semantic concepts expressed by the user and these low-level features. The clustering results are therefore generally different from the intent of the user. Our work aims to involve users in the clustering phase so that they can interact with the system in order to improve the clustering results. The clustering methods should therefore produce a hierarchical cluster structure where the initial clusters may be easily merged or split. We are also interested in clustering methods which can be incrementally built in order to facilitate the insertion or deletion of new images by the user. It can be noted that incrementality is also very important in the context of huge image databases, when the whole dataset cannot be stored in the main memory. Another very important point is the computational complexity of the clustering algorithm, especially in an interactive online context where the user is involved.

In the case of large image database indexing, we may be interested in traditional clustering (unsupervised) (Jain et al., 1999; Xu et al., 2005) or semi-supervised clustering (Basu et al., 2002, 2004; Dubey et al., 2010; Wagstaff et al., 2001). While no information about ground truth is provided in the case of unsupervised clustering, a limited amount of knowledge is available in the case of semi-supervised clustering. The provided knowledge may consist of class labels (for some objects) or pairwise constraints (must-link or cannot-link) between objects.

In Lai et al. (2012a), we proposed a survey of unsupervised clustering techniques and analyzed the advantages and disadvantages of different methods in a context of huge masses of data where incrementality and hierarchical structuring are needed. We also experimentally compared five methods (global k-means (Likas et al., 2003), AHC (Lance and Williams, 1967), R-tree (Guttman, 1984), SR-tree (Katayama and Satoh, 1997) and BIRCH (Zhang et al., 1996)) with different real image databases of increasing sizes (Wang, PascalVoc2006, Caltech101, Corel30k) (the number of images ranges from 1000 to 30,000) to study the scalability of different approaches relative to the size of the database. In Lai et al. (2012b), we presented an overview of semi-supervised clustering methods and proposed a preliminary experiment of an interactive semi-supervised clustering model using the HMRF-kmeans (Hidden Markov Random Fields kmeans) clustering (Basu et al., 2004) on the Wang image database in order to analyze the improvement in the clustering process when user feedback is provided.

There are three main parts to this paper. Firstly, we propose a new interactive semi-supervised clustering model using pairwise constraints. Secondly, we investigate different methods for deducing pairwise constraints from user feedback. Thirdly, we experimentally compare our proposed semi-supervised method with the widely known semi-supervised HMRF-kmeans method.

This paper is structured as follows. A short review of semi-supervised clustering methods is presented in Section 2. Our interactive semi-supervised clustering model is proposed in Section 3. Some experiments are presented in Section 4. Some conclusions and further works are provided in Section 5.

2. A short review of semi-supervised clustering methods

For unsupervised clustering only similarity information is used to organize objects; in the case of semi-supervised clustering a small amount of prior knowledge is available. Prior knowledge is either in the form of class labels (for some objects) or pairwise constraints between objects. Pairwise constraints specify whether two objects should be in the same cluster (must-link) or in different clusters (cannot-link). As the clusters produced by unsupervised clustering may not be the ones required by the user, this prior knowledge is needed to guide the clustering process for resulting clusters which are closer to the user's wishes. For instance, for clustering a database with thousands of animal images, an user may want to cluster by animal species or by background landscape types. An unsupervised clustering method may give, as a result, a cluster containing images of elephants with a grass background together with images of horses with a grass background and another cluster containing images of elephants with a sand background. These results are ideal when the user wants to cluster by background landscape types. But they are poor when the user wants to cluster by animal species. In this case, must-link constraints between images of elephants with a grass background and images of elephants with a sand background and cannot-link constraints between images of elephants with a grass background and images of horses with a grass background are needed to guide the clustering process. The objective of our work is to make the user interact with the system so as to define easily these constraints with only a few clicks. Note that the available knowledge is too poor to be used with supervised learning, as only a very limited ratio of the available images are considered by the user at each step. In general, semi-supervised clustering methods are used to maximize intra-cluster similarity, to minimize inter-cluster similarity and to keep a high consistency between partitioning and domain knowledge.

Semi-supervised clustering has been developed in the last decade and some methods have been published to date. They can be divided into semi-supervised clustering with labels, where partial information about object labels is given, and semi-supervised clustering with constraints, where a small amount of pairwise constraints between objects is given.

Some semi-supervised clustering methods using labeled objects have been put forward: seeded-kmeans (Basu et al., 2002), constrained-kmeans (Basu et al., 2002), etc. Seeded-kmeans and constrained-kmeans are based on the k-means algorithm. Prior knowledge for these two methods is a small subset of the input database, called seed set, containing user-specified labeled objects of k different clusters. Unlike k-means algorithm which randomly selects the initial cluster prototypes, these two methods use the labeled objects to initialize the cluster prototypes. Following this we repeat, until convergence, the re-assignment of each object in the dataset to the nearest prototype and the re-computation of the prototypes with the assigned objects. The seeded-kmeans assigns objects to the nearest prototype without considering the prior labels of the objects in the seed set. In contrast, the constrained-kmeans maintains the labeled examples in their initial clusters and assigns the other objects to the nearest prototype. An interactive cluster-level semi-supervised clustering was proposed in Dubey et al. (2010) for document analysis. In this model, knowledge is progressively provided as assignment feedback and cluster description feedback after each interactive iteration. Using assignment feedback, the user moves an object from one cluster to another cluster. Using cluster description feedback, the user modifies the feature vector of any current cluster (e.g. increase the weighting of some important words). The algorithm learns from all the feedback to re-cluster the dataset in order to minimize average distance between points and their cluster centers while minimizing the violation of constraints corresponding to feedback.

Among the semi-supervised clustering methods using pairwise constraints between objects, we can cite COP-kmeans (constrained-kmeans) (Wagstaff et al., 2001), HMRF-kmeans (Hidden Markov Random Fields Kmeans) (Basu et al., 2004), semi-supervised kernel-kmeans (Kulis et al., 2005), etc. The input data of these