Expert Systems With Applications 56 (2016) 349–359

Contents lists available at ScienceDirect

Expert Systems With Applications

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

DENDIS: A new density-based sampling for clustering algorithm

Frédéric Ros[a,*], Serge Guillaume[b]

[a] Laboratory PRISME, Orléans university, 8 Rue Léonard de Vinci - 45072 Orléans, France
[b] Irstea, UMR ITAP, 361, rue J.F. Breton B.P. 5095 34196, Montpellier Cedex 5, France

ARTICLE INFO

Article history:
Received 6 November 2015
Revised 6 February 2016
Accepted 3 March 2016
Available online 17 March 2016

Keywords:
Density
Distance
Space coverage
Clustering
Rand index

ABSTRACT

To deal with large datasets, sampling can be used as a preprocessing step for clustering. In this paper, an hybrid sampling algorithm is proposed. It is density-based while managing distance concepts to ensure space coverage and fit cluster shapes. At each step a new item is added to the sample: it is chosen as the furthest from the representative in the most important group. A constraint on the hyper volume induced by the samples avoids over sampling in high density areas. The inner structure allows for internal optimization: only a few distances have to be computed. The algorithm behavior is investigated using synthetic and real-world data sets and compared to alternative approaches, at conceptual and empirical levels. The numerical experiments proved it is more parsimonious, faster and more accurate, according to the Rand Index, with both k-means and hierarchical clustering algorithms.

1. Introduction

Summarizing information is a key task in information processing, either in data mining, knowledge induction or pattern recognition. Clustering (Ling, 1981) is one of the most popular techniques. It aims at grouping items in such a way that similar ones belong to the same cluster and are different from the ones which belong to other clusters. Many methods (Andreopoulos, An, Wang, & Schroeder, 2009) have been proposed to identify clusters according to various criteria. Some of them (Nagpal, Jatain, & Gaur, 2013) are based on an input space partition (k-means, spectral clustering, Clarans) or grid techniques (like Sting or Clique), others are density-based (Dbscan, Denclue, Clique). Some of these techniques benefit a tree implementation: Birch, Cure, Diana, Chameleon, Kd-tree.

Algorithms are becoming more and more sophisticated in order to be able to manage data with clusters of various shapes and densities. This leads to an increased computational cost which limits their practical use, especially when applications concern very large database like records of scientific and commercial applications, telephone calls, etc. Clearly, most of mature clustering techniques address small or medium databases (several hundreds of patterns) but fail to scale up well with large data sets due to an excessive computational time. Therefore, in addition to the usual performance requirements, response time is of major concern to most data clustering algorithms nowadays. Obviously, algorithms with quadratic or exponential complexity, such as hierarchical approaches, are strongly limited, but even algorithms like *k-means* are still slow in practice for large datasets.

While some approaches aim to optimize and speed up existing techniques (Chiang, Tsai, & Yang, 2011; Viswanath, Sarma, & Reddy, 2013), sampling appears as an interesting alternative to manage large data sets. In our case, sampling is a preprocessing step for clustering and clustering is assessed according to cluster homogeneity and group separability. This calls for two basic notions: density and distance. Clusters can be defined as dense input areas separated by low density transition zones. Sampling algorithms are based upon these two notions, one driving the process while the other is more or less induced.

Various techniques have been proposed in the abundant literature. Some algorithms estimate local density, using neighborhood or kernel functions (Kollios, Gunopulos, Koudas, & Berchtold, 2003), in order to bias the random sampling to make sure small clusters are represented in the sample (Ilango & Mohan, 2010; Palmer & Faloutsos, 2000). Others work at a global scale, like the popular *k-means* or evolutionary approaches (Naldi & Campello, 2015). In the former, the number of representatives is a priori set, each center induces an attraction basin. The third category includes incremental algorithms. They can be driven either by the attraction basin size (Yang & Wu, 2005) favoring the density search or by distance concepts (Rosenkrantz, Stearns, & Lewis, 1977; Sarma, Viswanath, & Reddy, 2013) promoting the coverage aspect. Incremental algorithms differ in the heuristics introduced to balance the

* Corresponding author. Tel.: +33 238642588.
E-mail addresses: frederic.ros@univ-orleans.fr, frederic.ros@free.fr (F. Ros), serge.guillaume@irstea.fr (S. Guillaume).

http://dx.doi.org/10.1016/j.eswa.2016.03.008
0957-4174/

density and distance concepts, and also in the parametrization. A comparison of algorithms for initializing *k-means* can be found in Celebi, Kingravi, and Vela (2013).

Fulfilling the two conflicting objectives of the sampling, ensuring small clusters coverage while favoring high local density areas, especially around the modes of the spatial distribution, with a small set of meaningful parameters, is still an open challenge.

The goal of this paper is to introduce a new incremental algorithm (http://frederic.rosresearch.free.fr) to meet these needs. DENDIS combines density and distance concepts in a really innovative way. Density-based, it is able to manage distance concepts to ensure space coverage and fit cluster shapes. At each step a new item is added to the sample: it is chosen as the furthest from the representative in the most important group. A constraint on the hyper volume induced by the samples avoids over sampling in high density areas. The attraction basins are not defined using a parameter but are induced by the sampling process. The inner structure allows for internal optimization. This makes the algorithm fast enough to deal with large data sets.

The paper is organized as follows. Section 2 reports the main sampling techniques. Then DENDIS is introduced in Section 3 and compared at a conceptual level to alternative approaches in Section 4. The optimization procedure is detailed in Section 5. Section 6 is dedicated to numerical experiments, using synthetic and real world data, to explore the algorithm behavior and to compare the proposal with concurrent approaches. Finally Section 7 summarizes the main conclusions and open perspectives.

2. Literature review

The simplest and most popular method to appear was uniform random sampling, well known to statisticians. The only parameter is the proportion of the data to be kept. Even if some work has been done to find the optimal size by determining appropriate bounds (Guha, Rastogi, & Shim, 1998), random sampling does not account for cluster shape or density. The results are interesting from a theoretical point of view (Chernoff, 1952), but they tend to overestimate the sample size in non worst-case situations.

Density methods (Menardi & Azzalini, 2014) assume clusters are more likely present around the modes of the spatial distribution. They can be grouped in two main families for density estimation: space partition (Ilango & Mohan, 2010; Palmer & Faloutsos, 2000) and local estimation, using neighborhood or kernel functions (Kollios et al., 2003).

The main idea of these methods is to add a bias according to space density, giving a higher probability for patterns located in less dense regions to be selected in order to ensure small cluster representation. The results are highly dependent upon the bias level and the density estimation method. The local estimation approaches (*kernel* or *k-nearest-neighbors*) require a high computational cost. Without additional optimization based on preprocessing, like the bucketing algorithm (Devroye, 1981), they are not scalable. However this new step also increases their complexity.

Distance concepts are widely used in clustering and sampling algorithms to measure similarity and proximity between patterns. The most popular representative of this family remains the *k-means* algorithm, and its robust version called *k-medoids*. It has been successfully used as a preprocessing step for sophisticated and expensive techniques such as hierarchical approaches or Support Vector Machine algorithms (SVM) (Xiao, Liu, Hao, & Cao, 2014). It is still the subject of many studies to improve its own efficiency and tractability (Khan & Ahmad, 2013; Lv et al., 2015; Zhong, Malinen, Miao, & Fränti, 2015). The proposals are based on preprocessing algorithms which are themselves related to sampling or condensation techniques (Arthur & Vassilvitskii, 2007; Zahra et al., 2015) including evolutionary algorithms (Hatamlou, Abdullah, & Nezamabadi-pour, 2012; Naldi & Campello, 2015). These algorithms are still computationally expensive (Tzortzis & Likas, 2014).

While the *k-means* is an iterative algorithm, whose convergence is guaranteed, some single data-scan distance based algorithms have also been proposed, such as *leader family* (Sarma et al., 2013; Viswanath et al., 2013) clustering or the furthest-first-traversal (fft) algorithm (Rosenkrantz et al., 1977). The pioneering versions of distance based methods are simple and fast, but they also are limited in the variety of shapes and densities they are able to manage. When improved, for instance by taking density into account, they become more relevant but their overall performance depends on the way both concepts are associated and, also, on the increase of the computational cost. The *mountain method* proposed by Yager and its modified versions (Yang & Wu, 2005) are good representatives of hybrid methodologies as well as the recent work proposed by Feldman, Faulkner, and Krause (2011). Density is managed by removing from the original set items already represented in the sample.

Strategies usually based on stratification processes have also been developed to improve and speed up the sampling process (Gutmann & Kersting, 2007). Reservoir algorithms (Al-Kateb & Lee, 2014) can be seen as a special case of stratification approaches. They have been proposed to deal with dynamic data sets, like the ones to be found in web processing applications. These method need an accurate setting to become really relevant.

Even if the context is rather different, Vector Quantization techniques (Chang & Hsieh, 2012), coming from the signal area especially for data compression, involve similar mechanisms. The objective is to provide a codebook representative of the original cover without distortion. The *LBG* algorithm and its variations (Bardekar & Tijare, 2011) appear to be the most popular. The methods are incremental, similar to the global k-means family approaches (Bagirov, Ugon, & Webb, 2011; Likas, Vlassis, & Verbeek, 2003), as at each step a new representative is added according to an appropriate criterion. Recent literature (Ma, Pan, Li, & Fang, 2015; Tzortzis & Likas, 2014) reports the difficulty to find the balance between length of codebook entries, its quality and time required for its formulation.

This short review shows that sampling for clustering techniques have been well investigated. Both concepts, density and distance, as well as the methods have reached a good level of maturity. Approaches that benefit from a kd-tree implementation (Nanopoulos, Manolopoulos, & Theodoridis, 2002; Wang, Wang, & Wilkes, 2009) seem to represent the best alternative, among the known methods, in terms of accuracy and tractability. However, they are highly sensitive to the parameter setting. The design of a method that would be accurate and scalable allowing to process various kinds of large data sets with a standard setting, remains an open challenge.

3. DENDIS: the proposed sampling algorithm

The objective of the algorithm is to select items from the whole set, T, to build the sample set, S. Each item in S is called a representative, each pattern in T is attached to its closest representative in S. The S set is expected to behave like the T one and to be as small as possible. *DENDIS* stands for DENsity and DIStance, meaning the proposal combines both aspects.

Overview of the algorithm It is an iterative algorithm that add a new representative at each step in order to reach two objectives. Firstly, ensure high density areas are represented in S, and, keeping in mind the small size goal, avoiding over representation. The second objective aims at homogeneous space covering to fit cluster shapes. To deal with the density requirement the new representative is chosen in the most populated set of attached patterns. For space covering purposes, the new representative is the