Expert Systems With Applications 55 (2016) 297–312

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Exploiting clustering algorithms in a multiple-level fashion: A comparative study in the medical care scenario

Tania Cerquitelli*, Silvia Chiusano, Xin Xiao

Control and Computer Engineering Department, Politecnico di Torino, Corso Duca degli Abruzzi, 24–10129 Torino, Italy

ARTICLE INFO

Keywords:
Cluster analysis
Data with a variable distribution
Diabetic patient treatments
Sequential patterns
Mobile applications

ABSTRACT

Clustering real-world data is a challenging task, since many real-data collections are characterized by an inherent sparseness and variable distribution. An appealing domain that generates such data collections is the medical care scenario where collected data include a large cardinality of patient records and a variety of medical treatments usually adopted for a given disease pathology.

This paper proposes a two-phase data mining methodology (MLC) to iteratively analyze different dataset portions and locally identify groups of objects with common properties. Discovered cohesive clusters are then analyzed using sequential patterns to characterize temporal relationships among data features. To support an automatic classification of new data objects within one of the discovered groups, a classification model is created starting from the computed cluster set. A mobile application has been also designed and developed to visualize and update data under analysis as well as categorizing new unlabeled data objects.

The experimental evaluation conducted on real datasets in the medical care scenario showed the effectiveness of MLC to discover interesting knowledge items and to easily exploit them through a mobile application. Results have been also discussed from a medical perspective.

1. Introduction

Cluster analysis is an exploratory technique which aims at grouping a data object collection into subsets (clusters) based on object properties, without the support of additional a priori knowledge (Pang-Ning, Steinbach, & Kumar, 2006). Nevertheless clustering is a widely studied data mining problem, clustering real-world data collections may impose new challenges. Real datasets are usually characterized by an *inherent sparseness* and *variable distribution*, since they are generated by a large variety of events, and *high data dimensionality* because features used to model real objects and human actions may have very large domains. The variability in data distribution grows with data volume, thus increasing the complexity of mining such data. For example, health care data collections can have large volume due to the large cardinality of patient records. Because of the variety of medical treatments usually adopted for the different degrees of severity of a given pathology, patient data collections are also usually characterized by high dimensionality, variable data distribution and inherent sparseness. However, at present, most clustering algorithms perform better with uniform data distribution, while their performance as well as the quality of the extracted knowledge tend to decrease in non-uniform collections.

Aimed at addressing the above issues, this paper presents a *M*ultiple-*L*evel *C*lustering (MLC) framework which comprises two data mining phases. First MLC exploits clustering algorithms in a multiple-level fashion to iteratively focus on different dataset portions and *locally* identify groups of correlated objects. Cohesive and well-separated clusters with diverse data distributions are discovered. Then, the cluster content is concisely described in terms of data features most frequently appearing in the cluster and sequential patterns capturing temporal correlations among data features. Moreover, for supporting the automatic categorization of new data objects into one of the discovered cluster, a classification model is created starting from the cluster set. To allow ubiquitous real-time classification of new data, a two-tier architecture based on a mobile (Android) application has been designed and developed.

Before to apply the clustering analysis, in the MLC framework data are represented in the Vector Space Model (VSM; Salton, 1971) using the TF-IDF method (Pang-Ning et al., 2006) with the aim of highlighting the relevance of specific data characteristics. In this study, five different multiple-level clustering algorithms have been integrated into MLC, based on K-means (i.e., bisecting and

* Corresponding author. Tel.: +393487659476; fax: +390110907099.
E-mail addresses: tania.cerquitelli@polito.it, tania.cerquitelli@gmail.com (T. Cerquitelli), silvia.chiusano@polito.it (S. Chiusano), xin.xiao@polito.it (X. Xiao).

http://dx.doi.org/10.1016/j.eswa.2016.02.005
0957-4174/

refined K-means; Steinbach, Karypis, & Kumar, 2000), K-medoids (i.e., bisecting and refined K-medoids; Kashef & Kamel, 2008), and DBSCAN methods (i.e., multiple-level DBSCAN; Antonelli et al., 2013). Clustering results have been then analyzed and compared using some well-established quality indices, as SSE, Silhouette and overall similarity, and Rand Index (Pang-Ning et al., 2006). Maximal sequential patterns (Zaki, 2001) have been selected to concisely describe temporal correlations among data features appearing in each cluster. Decision trees (Pang-Ning et al., 2006) have been used to build the classification model, since they have been shown to provide accurate models in various application domains.

The MLC framework has been validated on three real datasets in the medical care scenario, i.e., underwent examinations by patients, drug prescriptions to patients, Twitter messages on healthcare job information. We considered as a reference case study the former dataset including the examination log data of (anonymized) patients with overt diabetes. Diabetic patients may suffer by various disease complications as eye problems, neuropathy, kidney and cardiovascular diseases. Patients affected by disease complications (or at risk of them) should be tested with more specific examinations in addition to routine tests to monitor its status (or reveal the pathology). The considered data collection is characterized by an inherently sparse distribution due to the variety of possible examinations, covering both routine tests and more specific examinations for different degrees of severity in diabetes.

The experimental evaluation showed that the multiple-level clustering strategy can effectively partition the initial data collection into cohesive groups, that can be then locally analyzed. Specifically, in the considered use case, interesting clusters containing patients with a similar examination history (with standard or more specific examinations) can be discovered. It also pointed out that, nevertheless both the multiple-level DBSCAN and the refined k-means algorithms generate cluster sets with good quality and agreement, from a medical perspective the multiple-level DBSCAN algorithm appears as the more suitable approach for patient analysis in the considered case study. Maximal sequential patterns characterizing cluster content highlight how examinations are interleaved and distributed over time. The classification performance showed the goodness of the constructed model and its efficiency in classifying new unlabeled data through a mobile application.

This paper is organized as follows. Section 2 describes previous work using clustering techniques in the medical care scenario. Section 3 presents the MLC framework and how the selected algorithms have been tailored to MLC. Section 4 reports the experimental study on real datasets, while Section 5 compares algorithm performance and analyses the results from a medical perspective. Section 6 draws the future developments of the proposed approach.

2. Related work

Clustering algorithms find application in a wide range of different domains, including sensor network data (Abbasi & Younis, 2007), biological data (Au, Chan, Wong, & Wang, 2005), and network traffic data (Eriksson, Barford, & Nowak, 2008). Clustering algorithms have been also widely used to analyze medical data (Esfandiari, Babavalian, Moghadam, & Tabar, 2014). Many studies addressed the identification of correlated groups of patients affected by different diseases. For example, Sengur and Turkoglu (2008) reviewed the cluster methods used to diagnose heart valve diseases. In Zheng, Yoon, and Lam (2014), clustering techniques were used to diagnose breast cancer based on tutor features, by recognizing hidden patterns of benign and malignant tumors. Khaing (2011) exploited the K-means algorithm to cluster a collection of patient records aimed at identifying relevant features of patients subjected to heart attack.

Some research efforts have been devoted to exploiting clustering techniques on data related to diabetic patients (Esfandiari et al., 2014). Different issues have been addressed as food analysis (Phanich, Pholkul, & Phimoltares, 2010), gait patterns (Sawacha, Guarneri, Avogaro, & Cobelli, 2010), discovering relationships among diabetes and risk factors (Chaturvedi, 2003), analyses of various imputation techniques (Purwar & Singh, 2015), and discovering similar medical treatments (Antonelli et al., 2013). Purwar and Singh (2015) focus on diabetes datasets using the K-means algorithm aimed at analysing various imputation techniques. Different from Purwar and Singh (2015), in this work we aim at identifying groups of patients with similar examination histories to provide a preliminary patient categorization into a set of predefined classes. Thus, we detailed each cluster with sequential patterns to discover how examinations are interleaved and distributed over time.

The idea of exploiting a clustering algorithm in a multiple-level fashion was first introduced in Antonelli et al. (2013) and used in Baralis, Cerquitelli, Chiusano, Grimaudo, and Xiao (2013) to analyze twitter messages. A first study towards a combined distance measure for clustering medical records has been presented in Bruno, Cerquitelli, Chiusano, and Xiao (2014). A parallel effort devoted to clustering documents proved that bisecting K-means was preferable to other clustering methods as standard K-means and hierarchical approaches (Steinbach et al., 2000).

The MLC data analysis framework presented in this study enhances the methodology proposed in Antonelli et al. (2013) by providing a more general approach which (i) integrates different clustering algorithms, (ii) uses more indices to evaluate cluster quality, (iii) characterizes temporal aspects of interlived examinations through sequential patterns, (iv) exploits cluster set enriched with domain semantics to train a classification model, and (v) allows ubiquitous classification on new unlabeled examination histories through a mobile application. MLC does not exploit the distance measure proposed in Bruno et al. (2014) because information on patient profiles (i.e., patient age and gender) are not available on the real data collection discussed as a reference case study. Among the different categories of clustering algorithms, i.e., prototype (e.g., K-means; Juang and Rabiner, 1990, K-medoids; Kaufman & Rousseeuw, 1990), density (e.g., DBSCAN; Ester, Kriegel, Sander, & Xu, 1996), model (e.g., EM; McLachlan & Krishnan 1997), and hierarchical based methods (Pang-Ning et al., 2006), in this study we focused on the two popular categories of prototype and density based methods for the development of the MLC framework. Furthermore, we integrated in MLC the maximal sequential pattern miner (Fournier-Viger, Wu, Gomariz, & Tseng, 2014b) to characterize cluster content and identify how patient examinations are interlived and distributed over time. To ease the exploitation of cluster results, the decision tree (as in Bruno et al., 2014), has been integrated in MLC to train a classification model. The latter is then exploited in an Android application to allow ubiquitous patient classification to new unlabeled examination histories.

The wide diffusion of mobile technologies and the increasing capabilities of mobile computing devices caused an increased interest in designing, implementing and testing innovative applications running on mobile devices to provide a wide range of useful services. In the medical care scenario, some efforts (Karan, Bayraktar, Gümüşkaya, & Karlık, 2012; Menshawy, Benharref, & Serhani, 2015; Polat, Güneş, & Arslan, 2008) have been devoted on this appealing research. In Karan et al. (2012), a distributed end-to-end pervasive healthcare system utilizing neural network computations for diagnosing diabetes was developed in small mobile devices. Menshawy et al. (2015) developed a new mobile-based approach to automatically detect seizures, using k-means as unsupervised classification technique. Polat et al. (2008) have presented Generalized Discriminant Analysis and Least Square Support Vector Machine models