Knowledge-Based Systems 82 (2015) 70–79

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Mining patterns for clustering on numerical datasets using unsupervised decision trees

A.E. Gutierrez-Rodríguez [a,b,*], J. Fco Martínez-Trinidad [a], M. García-Borroto [c], J.A. Carrasco-Ochoa [a]

[a] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, Mexico
[b] Centro de Bioplantas, Universidad de Ciego de Ávila (UNICA), Ciego de Avila, Cuba
[c] Instituto Superior Politécnico José Antonio Echeverría (CUJAE), Havana, Cuba

ARTICLE INFO

Article history:
Received 10 October 2014
Received in revised form 26 January 2015
Accepted 22 February 2015
Available online 28 February 2015

Keywords:
Pattern-based clustering
Frequent patterns
Unsupervised decision trees
Numerical datasets
Cluster validity indices

ABSTRACT

Pattern-based clustering algorithms return a set of patterns that describe the objects of each cluster. The most recent algorithms proposed in this approach extract patterns on numerical datasets by applying an a priori discretization process, which may cause information loss. In this paper, we introduce a new pattern-based clustering algorithm for numerical datasets, which does not need an a priori discretization on numerical features. The new algorithm extracts, from a collection of trees generated through a new induction procedure, a small subset of patterns useful for clustering. Experimental results show that the patterns extracted by the proposed algorithm allows to build a pattern-based clustering algorithm, which obtains better clustering results than recent pattern-based clustering algorithms. In addition, the proposed algorithm obtains similar clustering results, in quality, than traditional clustering algorithms.

1. Introduction

Non-overlapping clustering is about partitioning a set of unlabeled objects into disjoints clusters, according to a certain criterion [1]. From now, when we refer to clustering, we will refer to non-overlapping clustering. A widely used clustering criterion is that objects of the same cluster should be more similar than objects from different clusters [2], in terms of a comparison function [3]. However, lack of comprehensibility of the results is a common issue of clustering algorithms based on comparison functions.

In some applications of data mining and knowledge discovery, users need an explanation about the clustering results, more than just a list of objects for each cluster [4]. Pattern-based clustering constitutes a different approach for clustering [5]. Algorithms under this approach, in addition to the list of objects belonging to each cluster, return a set of patterns that describe each cluster.

A pattern is an expression, in some language (relating features and their values), that describes, or covers, an object subset [6]. Several works have been reported into the pattern-based clustering approach [7–11,5]. One of the most recent algorithms is proposed in [5], but for applying this algorithm over numerical datasets, all numerical features must be a priori discretized, which may cause information loss.

In this paper, we introduce a new pattern-based clustering algorithm for numerical datasets, which extracts a small subset of patterns useful for clustering without applying an a priori discretization on numerical features. The proposed algorithm allows creating understandable numerical patterns. This algorithm extracts patterns from a collection of unsupervised decision trees created through a novel induction procedure. Our experiments show that the proposed algorithm obtains better results than recent pattern-based clustering algorithms. Additionally, our clustering results are competitive, in quality, with traditional clustering algorithms like K-Means [12] and EM [13].

One of the main contributions of our work is a novel procedure to induce a collection of binary unsupervised decision trees. This procedure generates a fix number of trees by selecting just some splits in some of the top levels of each tree. Another important contribution of the paper is a new split evaluation criterion for numerical features. This criterion allows selecting those splits that maximize the difference between the centroids of the children nodes, which is a widely used clustering criterion. Moreover, we propose an algorithm for extracting patterns from the induced

* Corresponding author at: Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, Mexico. Tel.: +52 222 266 3100, +53 33 22 4016.
E-mail address: aegr82@gmail.com (A.E. Gutierrez-Rodríguez).

http://dx.doi.org/10.1016/j.knosys.2015.02.019
0950-7051/

trees. Finally, we present a strategy for clustering, which first groups the patterns, and then groups the objects.

This paper is organized as follows: Section 2 presents a brief review about pattern-based clustering. Section 3 introduces the proposed algorithm for mining a subset of patterns useful for clustering, without applying an a priori discretization on numerical features; additionally, a pattern-based clustering algorithm that uses the mined patterns is also presented. Section 4 shows the experimental results. Conclusions and future work appear in Section 5.

2. Related work

In the literature, several works on pattern-based clustering have been reported. Michalski proposed CLUSTER/2 [7] in 1983. This algorithm groups a dataset in k predefined clusters trying that the descriptions (patterns) of the clusters be simple and fit well the dataset. In a first step, CLUSTER/2 selects k objects as seeds. Then, for each seed, a pattern is built using those feature-value items different from those in other seeds. Based on these patterns, clusters are built and the patterns are modified in order to obtain disjoint clusters. These steps are repeated until a quality criterion is fulfilled.

In 1987, Fisher developed the COBWEB [8] clustering algorithm. This algorithm incrementally builds a clustering by means of a classification tree where each node is a probabilistic pattern that represents an object class. The classification of a new object is performed by descending the tree along the appropriate path. COBWEB is an incremental algorithm for hierarchical clustering and it does not partition the dataset in a k predefined number of clusters.

Ralambondrainy proposed a pattern-based K-Means algorithm in 1995, called CKM [9]. This algorithm groups objects with the traditional K-Means clustering algorithm. Then, a characterization phase builds the patterns. This algorithm has the limitation that the patterns extracted in a post clustering step are not taken into account for building the clusters, which is out of the main idea of the pattern-based clustering approach.

In 2004, Mishra [10] introduced a graph formulation for pattern-based clustering with the objective of identifying a collection of patterns that describe the objects. The author connects the pattern-based clustering problem with the maximum edge biclique problem [14]. The clusters discovered by this approach may overlap and also they may not cover all the objects.

In 2008, Wong proposed a pattern-based clustering algorithm to simultaneously cluster patterns and data [11]. Unlike traditional pattern mining algorithms, which extract patterns based on feature-value frequencies, this algorithm extracts patterns from categorical data using correlation relationships between features. This algorithm defines several distance measures between patterns. Then, applying a traditional agglomerative hierarchical algorithm, the patterns and their associated data are clustered using these distances. This algorithm cannot partition the dataset in a k predefined number of clusters.

Fore and Dong reported the CPC algorithm [5] in 2012. In the first step, CPC mines all patterns in a dataset using the FP-Growth algorithm [15]. Then, the extracted patterns are filtered by equivalence classes. This algorithm defines a relationship between patterns, which is used to cluster the set of patterns in k clusters. After building clusters of patterns, the objects in the dataset are classified into these clusters.

Finally, it is important to separate those works that, at first sight, could be considered as related to pattern-based clustering, but indeed they follow a different approach. Some works use the term pattern-based clustering for *subspace clustering* [16–20]; however, subspace clustering is about clustering objects using subsets of features. Into *text clustering*, there are several works that report the use of patterns for building clusters [21–25]; nevertheless, these works use patterns to create a bag-of-words like vector representation, and then a traditional clustering algorithm is applied.

In the pattern-based clustering approach, usually all numerical features are a priori discretized in order to extract patterns, which may cause information loss. Moreover, it is desirable that algorithms return just a few comprehensible patterns for describing the clusters; since they are designed for situations where an explanation of the results is required. For this reason, in this paper, we propose a pattern-based clustering algorithm that extracts just a subset of useful patterns for clustering, without applying an a priori discretization on numerical data.

3. Our proposal

In this section, we introduce a **P**attern-based **C**lustering algorithm for **N**umerical datasets (**PCN**) that extracts just a subset of patterns useful for clustering, without applying an a priori discretization on numerical features.

A pattern is a conjunction of relational statements $X_i \# R_i$, where R_i is a value in the domain of the feature X_i, and $\#$ is a relational operator [7]. Traditional pattern mining algorithms only use "$=$" as relational operator because they are defined for categorical data and, for this reason, these algorithms need to apply an a priori discretization on numerical features; for example, a pattern extracted from a numerical feature might be $[Age = [0, 20)]$. In our proposal, for avoiding an a priori discretization step, we propose using the relational operators "$\leqslant$" and "$>$" for numerical features, which allows to build patterns as $[Age \leqslant 15]$.

For mining patterns, we use a collection of binary *unsupervised decision trees*, generated through a new induction procedure, which includes a new split evaluation criterion to build internal nodes of the tree. In order to generate diversity among trees, our proposal, unlike traditional methods for building unsupervised decision trees [26], generates more candidate splits than those generated by conventional methods.

3.1. Inducing unsupervised decision trees

We propose to build binary unsupervised decision trees by creating splits that maximize the difference between the feature-value means (centroids) of the children nodes. This reflects the criterion that the objects in different clusters should be as dissimilar as possible, which is a widely used clustering criterion.

For each feature f_j, the induction algorithm creates, for each value v_i, two candidate splits (two new nodes) with the properties $f_j \leqslant v_i$ and $f_j > v_i$, respectively. We associate to each new node those objects in the parent node that fulfill the corresponding property. Then, for each new node, we compute the average of the values in the feature f_j. This average represents the mean of that node. In Eq. (1) we define the mean of a node:

$$M = \frac{\sum_{v_i \in V} v_i}{|V|}, \tag{1}$$

where V is the set of values that the feature f_j takes on objects in the node.

In Eq. (2), we define a split evaluation criterion Q, for a decision node N, as the difference between the means of the left and right children nodes:

$$Q(f_j, v_i) = \frac{|M_L - M_R|}{\max\{V\} - \min\{V\}}, \tag{2}$$