Contents lists available at ScienceDirect

Pattern Recognition

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

A unifying criterion for unsupervised clustering and feature selection

Mihaela Breaban *, Henri Luchian

Faculty of Computer Science, Alexandru Ioan Cuza University, Iasi, Romania

ARTICLE INFO

Article history:
Received 20 December 2009
Received in revised form
21 July 2010
Accepted 6 October 2010

Keywords:
Unsupervised feature selection
Unsupervised clustering
Global optimization

ABSTRACT

Exploratory data analysis methods are essential for getting insight into data. Identifying the most important variables and detecting quasi-homogenous groups of data are problems of interest in this context. Solving such problems is a difficult task, mainly due to the unsupervised nature of the underlying learning process. Unsupervised feature selection and unsupervised clustering can be successfully approached as optimization problems by means of global optimization heuristics if an appropriate objective function is considered. This paper introduces an objective function capable of efficiently guiding the search for significant features and simultaneously for the respective optimal partitions. Experiments conducted on complex synthetic data suggest that the function we propose is unbiased with respect to both the number of clusters and the number of features.

1. Introduction

Clustering is the task of identifying *natural* groups in data. The problem can be stated more formally as follows:

Given a set S of n data items each of which is described by m numerical attributes: $S=\{d_1,d_2,\ldots,d_n\}$ where $d_i=\{f_{i1},f_{i2},\ldots,f_{im}\}\in\Im_1\times\Im_2\times\cdots\times\Im_m\subset\Re^m\forall i=\overline{1..n}$, find

$$C^*=\underset{C\in\Omega}{\mathrm{argmax}}F(C)$$

where

- Ω is the set of all possible hard partitions C of the data set S, where each C is a hard partition if $C=\{C_1, C_2,\ldots,C_k\}$, $\bigcup_{i=1}^{k} C_i=S$ and $C_i\bigcap C_j=\emptyset\ \forall i,j=\overline{1..k},i\neq j,k\in\{1,2,\ldots,card(C)\}$.
- F is a function which measures the quality of each partition $C\in\Omega$ with respect to the requirement implicitly described above by the word *natural*: similar data items should belong to the same cluster and dissimilar items should reside in distinct clusters.

The notion of *similarity* is seldom given in the problem statement.

If the number of clusters k is known in advance the problem is called *supervised clustering*; otherwise, it is called *unsupervised clustering*.

This definition leaves space to a wide choice of objective functions and similarity functions, depending strongly on the domain under investigation. The choice is rarely straightforward. The literature records a lot of comparative studies regarding the impact of various objective functions on the solution especially in the case of unsupervised clustering. As for the similarity function, if extra-information is available in the form of pairwise constraints of data items that must reside in the same cluster (the case of semi-supervised clustering and supervised classification) then an optimal distance metric can be learned. For unsupervised clustering, metric learning is usually performed in a pre-processing step, using methods that reduce data dimensionality through statistical analysis.

Dimensionality reduction is a problem intensively studied in both supervised and unsupervised clustering. The main goal is to reduce the size of the representation of data items in order to decrease the computational cost of subsequent steps, with minimal alterations in terms of descriptive accuracy. Dimensionality reduction is approached in two distinct ways: feature selection (FS) and feature extraction (FE). The feature selection approach searches for irrelevant original features (attributes) and excludes them; additionally, feature weighting may be performed. Feature extraction methods create new features from the original ones. The points in the original D-dimensional feature space are mapped into new points in a d-dimensional feature space, $d<D$. Compared to FS methods, FE methods provide an improved lower-dimensional representation for the full data set; however, an important drawback of FE methods is that the relationship between the original and the reduced space is more difficult to interpret.

Feature selection plays different roles in the supervised and, respectively, the unsupervised scenario. In both situations, in a pre-processing step redundant features may be eliminated by means of statistical analysis. Further, in classification feature selection aims at identifying those features that predict with highest accuracy the

* Correspondence to author. Facultatea de Informatica, Universitatea “Al. I. Cuza”, General Berthelot, 16, Iasi 700483, Romania. Tel.: +40 744 821303; fax: +40 232 201490.

E-mail addresses: pmihaela@infoiasi.ro (M. Breaban), hluchian@infoiasi.ro (H. Luchian).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.10.006

appropriate class labels, while in clustering feature selection aims at identifying the features which generate good partitions.

Unsupervised feature selection is performed by means of:

- filter approaches, which compute some entropy measure in order to asses the grouping tendency of data items in different feature subspaces. The subsequent unsupervised learning method is completely ignored.
- wrapper approaches, which actually search for partitions in different feature subspaces using a clustering algorithm. These approaches give better results but at higher computational costs.

In view of the definition of clustering, feature selection can be stated as an optimization problem:

find

$$w^* = \underset{w}{\operatorname{argmax}} Q(S')$$

where

- $w = \{w_1, w_2, \ldots, w_m\} \in \{0,1\}^m$ is a binary string;
- S' is the data set constructed from the original set S and the string w as follows: $S' = \{d_1', d_2', \ldots, d_n'\}$, $d_i' = \{w_1 \cdot f_{i1}, w_2 \cdot f_{i2}, \ldots, w_m \cdot f_{im}\}$, $\forall i = \overline{1..n}$;
- $Q(S')$ is a function which measures the tendency of data items in set S' to group into well-separated clusters; it can be expressed by means of the entropy (filter approaches) or of a fitness function which measures the quality of a partition detected by a clustering algorithm (wrapper approaches). In the latter case feature weighting is akin to solving the clustering problem in different feature spaces.

Our study approaches unsupervised feature selection in a wrapper manner. In this regard, a new optimization criterion largely unbiased with respect to the number of clusters is introduced in Section 2. Section 3 discusses the normalization of the clustering criterion with respect to the number of features. Section 4 presents a framework for performing unsupervised feature selection in conjunction with unsupervised clustering and summarizes the experimental results. Section 5 draws conclusions and points to future work.

2. Unsupervised clustering: searching for the optimal number of clusters

Classical clustering methods, such as k-Means and hierarchical algorithms, are designed to use prior knowledge on the number of clusters. In k-Means, an iterative process reallocates data items to the clusters of a k-class partition in order to minimize the within-cluster variance. Hierarchical clustering adopts a greedy strategy constructing trees/dendrograms based on the similarity between data items; each level in these dendrograms corresponds to partitions with a specific number of clusters and the method offers no guidance regarding the level where the optimal partition is represented (hence, the optimal number of clusters).

The algorithms mentioned above are local optimizers. In order to design a global optimizer for the clustering problem, a criterion for ranking all partitions, irrespective of the number of clusters, is needed. The problem is far for being trivial: with no hint on the number of clusters, common-sense clustering criteria like minimizing the variance within clusters and/or maximizing the distance between clusters guide the search towards the extreme solution—the n-class partition with each class containing exactly one point.

Existing studies in the literature propose and experiment with various clustering criteria [3,22,24,13]. The main concern is the bias these criteria introduce towards either lower or higher numbers of clusters. Since this bias proved to be hard to eliminate, multi-objective algorithms were proposed [9], which evaluate the quality of a partition against several criteria. The main drawback remains the fact that identifying the optimal solution within the final Pareto front is not straightforward.

The clustering criterion used in the present work originates in the analogy with the Huygens' theorem from mechanics, analogy introduced in [6] and used further in [19]. Considering the data set S in the above definitions, the following notations are used:

$W = \sum_{i=1}^{k} \sum_{d \in C_i} \delta(c_i, d)$ is the within-cluster inertia computed as the sum of the distances between all data items d in cluster C_i and their cluster center c_i;

$B = \sum_{i=1}^{k} |C_i| \cdot \delta(c_i, g)$ is the between-cluster inertia computed as the sum of the distances between the cluster centers c_i and the center of the entire data set g weighted with the size of each cluster $|C_i|$.

$T = \sum_{i=1}^{n} \delta(d_i, g)$ is the total inertia of the data set computed as the sum of the distances between the data items and the center g of the data set.

In the above *center* is the gravity center.

The above-mentioned analogy with mechanics can only be applied as an approximation. The simplest approximation of the Huygens theorem is then

$$W + B \approx T$$

According to the above formula, for any partition of the data set, regardless the number of clusters, the sum $W+B$ is merely constant. Fig. 1 illustrates this for the case of a data set with 10 random Gaussian features/variables: W, B, and $W+B$ are computed for locally optimal partitions of the data set obtained by the k-Means algorithm with the number of clusters varying between 2 and 50.

In view of the Huygens theorem, if the number of clusters is fixed, minimizing W or maximizing B are equivalent clustering criteria which can be used in general heuristics [6]. Note that the within-cluster variance is a widely used clustering criterion in supervised clustering. The Huygens theorem provides an equivalent clustering criterion (namely B), at a lower computational cost, which can be used in a nearest-neighbor assignment scenario [19].

When the number of clusters is unknown both these criteria are useless: they direct the search towards the extreme n-class partition. However, a corollary of the Huygens theorem in conjunction with penalties against the increase of the number of clusters proved to work in unsupervised clustering: $(B/T)^k$ is used in [21]; an equivalent (in view of the Huygens' theorem) function $(1/(1+W/B))^k$ is used in [20] in order to use local Mahalanobis distances. Unfortunately, extensive experiments we conducted recently with these fitness functions, showed that they

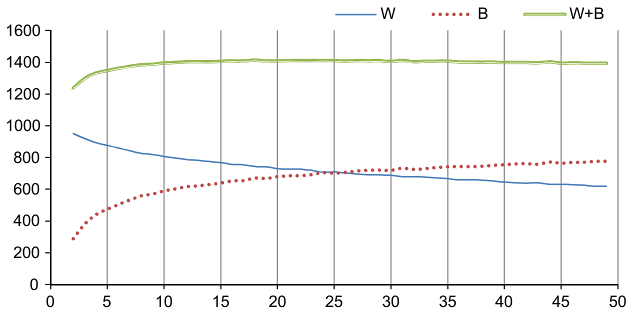

Fig. 1. The within-cluster inertia W, between-cluster inertia B and their sum plotted for locally optimal partitions obtained with k-Means over different numbers of clusters.