Pattern Recognition 47 (2014) 3046–3059

Contents lists available at ScienceDirect

Pattern Recognition

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

Cross-entropy clustering

J. Tabor [1], P. Spurek [*,2]

Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland

ARTICLE INFO

Article history:
Received 6 December 2012
Received in revised form
7 February 2014
Accepted 1 March 2014
Available online 18 March 2014

Keywords:
Clustering
Cross-entropy
Memory compression

ABSTRACT

We build a general and easily applicable clustering theory, which we call cross-entropy clustering (shortly CEC), which joins the advantages of classical k-means (easy implementation and speed) with those of EM (affine invariance and ability to adapt to clusters of desired shapes). Moreover, contrary to k-means and EM, *CEC finds the optimal number of clusters by automatically removing groups which have negative information cost*.

Although CEC, like EM, can be built on an arbitrary family of densities, in the most important case of Gaussian CEC the division into clusters is affine invariant.

1. Introduction

Clustering plays a basic role in many parts of data engineering, pattern recognition and image analysis [1–5]. Thus it is not surprising that there are many methods of data clustering, which however often inherit the deficiencies of the first method called k-means [6,7]. Since k-means has the tendency to divide the data into spherical shaped clusters of similar sizes, it is not affine invariant and does not deal well with clusters of various sizes. This causes the so-called mouse-effect, see Fig. 1(b). Moreover, it does not find the right number of clusters, see Fig. 1(c), and consequently to apply it we usually need to use additional tools like gap statistics [8,9]. Since k-means has so many disadvantages, one can ask why it is so popular. One of the possible answers lies in the fact that k-means is simple to implement and very fast compared to more advanced clustering methods[3] like EM [12,13].

In our paper we construct a general cross-entropy clustering (CEC) theory which simultaneously joins the clustering advantages of classical k-means and EM. The motivation of CEC comes from the observation that it is often profitable to use various compression algorithms specialized in different data types. We apply this observation in reverse, namely *we group/cluster those data together which are compressed by one algorithm from the preselected set of compressing algorithms.* In development of this idea we were influenced by the classical Shannon Entropy Theory [14–17] and Minimum Description Length Principle [18,19]. In particular we were strongly inspired by the application of MDLP to image segmentation given in [20,21]. A close approach from the Bayesian perspective can also be found in [22,23].

The above approach allows us automatic reduction of unnecessary clusters: contrary to the case of classical k-means or EM, there is a cost of using each cluster. To visualize our theory let us look at the results of Gaussian CEC given in Fig. 2(c), where we started with $k = 10$ initial randomly chosen clusters which were reduced automatically by the algorithm. The step-by-step view at this process can be seen in Fig. 3, in which we illustrate the subsequent steps of the Spherical CEC on random data lying uniformly inside the circle, and divided initially at two almost equal parts.

The clustering limitations of CEC are similar to those of EM, namely we divide the data into clusters of shapes which are reminiscent of the level sets of the family of the densities used. In particular, contrary to the density clustering [24] with the use of Gaussian CEC we will not build clusters of complicated shapes. Moreover, in an analogy to k-means, CEC strongly depends on the initial choice of clusters. This is the reason why in the paper we always started CEC at least twenty times from randomly chosen initial conditions to avoid arriving at the local minimum of the cost function. Let us mention that there are clustering methods, see [25], which allow us to better minimize the global minimum, however at the cost of the fixed number of clusters. The advantage

[*] Corresponding author. Tel.: +48 12 664 7543.
E-mail addresses: jacek.tabor@ii.uj.edu.pl (J. Tabor), przemyslaw.spurek@ii.uj.edu.pl (P. Spurek).

[1] Research of this author was supported by the National Center of Science (Poland) grant no. 2011/01/B/ST6/01887.

[2] Research of this author was supported by the National Center of Science (Poland) grant no. 2013/09/N/ST6/01178.

[3] This is excellently summarized in the third paragraph of [10]: "[…] The weaknesses of k-MEANS result in poor quality clustering, and thus, more statistically sophisticated alternatives have been proposed. […] While these alternatives offer more statistical accuracy, robustness and less bias, they trade this for substantially more computational requirements and more detailed prior knowledge [11]."

http://dx.doi.org/10.1016/j.patcog.2014.03.006
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

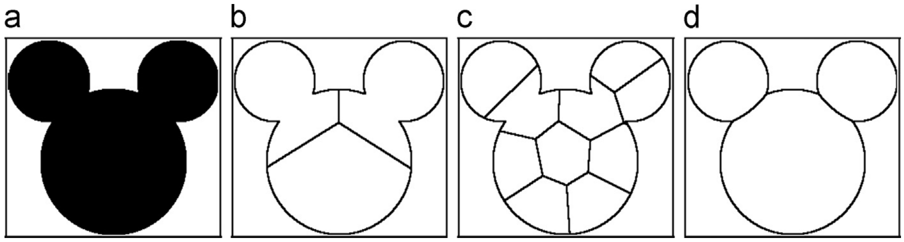

Fig. 1. Clustering of the uniform density on mouse-like set (a) by standard k-means algorithm with $k=3$ (b) and $k=10$ (c) compared with Spherical CEC (d) with initially 10 clusters (finished with 3). (a) Mouse-like set. (b) k-means with $k=3$. (c) k-means with $k=10$. (d) Spherical CEC.

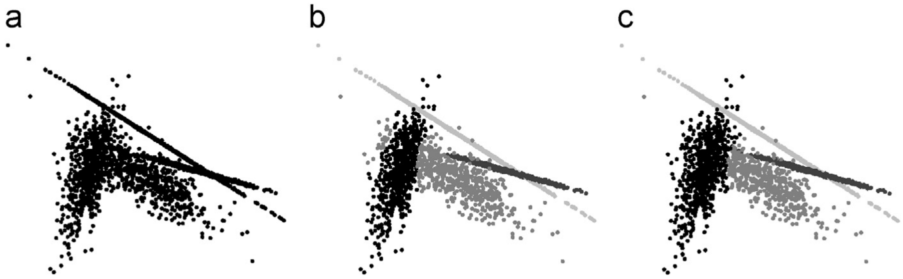

Fig. 2. Comparison of clustering of mixture of 4 Gaussians by EM (with 4 Gaussian densities) and Gaussian CEC starting from 10 initial clusters. (a) Data coming from mixture of 4 Gaussians. (b) EM clustering with 4 Gaussians. (c) CEC clustering with initially 10 Gaussians.

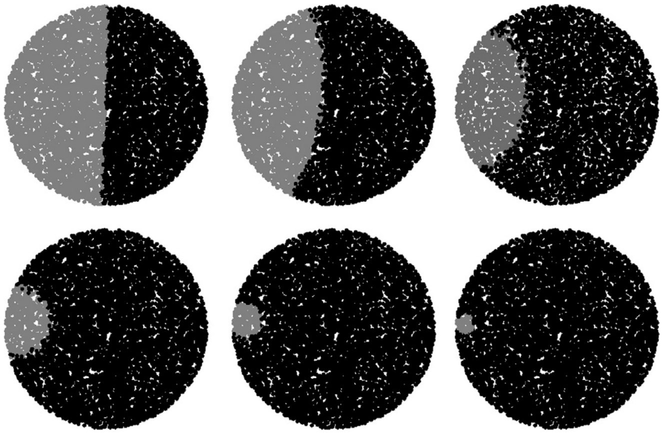

Fig. 3. Reduction of cluster by the spherical CEC.

comparing to most classical clustering methods [26] lies in the fact that we need only the maximal number of clusters, while we keep the same complexity as k-means.

There are a few probabilistic methods which try to estimate the correct number of clusters. For example in [27] the authors use the generalized distance between Gaussian mixture models with different components number by using the Kullback–Leibler divergence [14,16]. A similar approach is presented in [28] (Competitive Expectation Maximization) which uses a Minimum Message Length criterion [29]. In practice one can also directly use the MDLP in clustering [30]. The basic ideological difference lies in the fact that in MDLP we want to take into account the total memory cost of building the model, while in our case, like in EM, we use the classical entropy approach, and therefore assume that the memory cost of remembering the Gaussian (or in general density) parameters is zero.

Another modern clustering method worth mentioning is clearly support vector clustering [31]. In its basic form SVM allows separating the data with the use of hyperplanes while CEC (similarly as EM) allows the quadratic discriminant functions [32]. However, in its general form with the use of kernel functions support vector clustering will allow us to cluster the date into more complicated sets than CEC, usually at a cost larger numerical complexity. Consequently the CEC framework presented in the paper cannot cluster sufficiently well datasets presented in [32], since they are not well divided into Gaussian-shaped clusters.

For the convenience of the reader we now briefly summarize the contents of the paper. The next section is devoted to the gentle introduction to the basic properties of CEC. In particular we show that if the data comes from the known number of Gaussian densities, the basic results of CEC and EM clustering are similar. At the end of this section we discuss applications of CEC on real data-sets. In the following section we introduce notation and recall the necessary information concerning relative entropy. In the fourth section we provide a detailed motivation and explanation of our main idea which allows us to reinterpret the cross-entropy for the case of many “coding densities”. We also show how to apply classical Lloyds and Hartigan approaches to cross-entropy minimizations.

The last section contains applications of our theory to clustering with respect to various Gaussian subfamilies. We put a special attention on the question whether the given group of data should be divided into two separate clusters.

First we investigate the most important case of Gaussian CEC and show that it reduces to the search for the partition $(U_i)_{i=1}^k$ of the given data-set U which minimizes the objective cost function:

$$\frac{N}{2}\ln(2\pi e)+\sum_{i=1}^{k} p(U_i)\cdot\left[-\ln(p(U_i))+\frac{1}{2}\ln\det(\Sigma_{U_i})\right],$$

where $p(V)$ denotes the probability of choosing set V and Σ_V denotes the covariance matrix of the set V.

Then we study clustering based on the Spherical Gaussians, that is those with covariance proportional to identity. Comparing Spherical CEC to classical k-means we obtain that clustering is scale and translation invariant and clusters do not tend to be of fixed size. Consequently we do not obtain the mouse effect, see Fig. 1(d). To apply Spherical clustering we need the same information as in the classical k-means: in the case of k-means we seek the splitting of the data $U\subset\mathbb{R}^N$ into k sets $(U_i)_{i=1}^k$ such that the value of $\sum_{i=1}^k p(U_i)\cdot D_{U_i}$ is minimal, where $D_V=(1/\mathrm{card}(V))\sum_{v\in V}\|v-\mathrm{m}_V\|^2$ denotes the mean within cluster V sum of squares (and m_V is the mean of V). It occurs that the Gaussian spherical clustering in $\mathbb{R}^N$ reduces