Pattern Recognition 48 (2015) 941–952

Contents lists available at ScienceDirect

Pattern Recognition

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

Estimating the number of clusters in a numerical data set via quantization error modeling

Alexander Kolesnikov [a,*], Elena Trichina [b], Tuomo Kauranne [c]

[a] Arbonaut Ltd., Joensuu, Finland
[b] University of Eastern Finland, Joensuu, Finland
[c] Lappeenranta University of Technology, Lappeenranta, Finland

ARTICLE INFO

Article history:
Received 13 April 2014
Received in revised form
10 August 2014
Accepted 15 September 2014
Available online 30 September 2014

Keywords:
Clustering
Number of clusters
Vector quantization
Color quantization
Dominant colors
Fractal dimensions

ABSTRACT

In this paper, we consider the problem of unsupervised clustering (vector quantization) of multi-dimensional numerical data. We propose a new method for determining an optimal number of clusters in the data set. The method is based on parametric modeling of the quantization error. The model parameter can be treated as the effective dimensionality of the data set. The proposed method was tested with artificial and real numerical data sets and the results of the experiments demonstrate empirically not only the effectiveness of the method but its ability to cope with difficult cases where other known methods fail.

1. Introduction

Cluster analysis can be characterized as an attempt to represent a large population (data set) by a smaller number of points (centroids) thus sacrificing some of the information in favor of a more economic representation and more efficient processing of data. A large body of work has been dedicated to clustering validity criteria; the work of Milligan and Cooper [1] provides an extensive collection of references, which have been complemented by more recent surveys [2–4].

Estimation of the number of clusters in a data set is an important theoretical and practical problem in cluster analysis. With too few clusters, one cannot preserve the most relevant information about the structure of the data set **X**. On the other hand, with too many clusters, resources are wasted by processing non-relevant data. Although many algorithms have been suggested, there does not appear to be one most reliable method.

Our paper contributes to the quest for an effective method to establish the optimal number of clusters. The rest of the paper is organized as follows. The problem is formulated in Section 2 and several known methods reviewed in Section 3. Section 4 presents our solution to the problem. A new parametric model of the quantization error is introduced and a validity criterion derived from this model. In Section 5, we discuss the results of experiments with a number of data sets and provide comparisons with other methods. Section 6 presents conclusions.

2. Problem formulation

The problem comprises a data set $\mathbf{X}=\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ that consists of N points in a d-dimensional space: $\mathbf{x}_i=(x_{i,1}, \ldots, x_{i,d})$. The data are clustered into M clusters $\{C_1, \ldots, C_M\}$. A cluster C_j is defined by the centroid $c_j=(c_{j,1}, \ldots, c_{j,d})$ and the indices of data points in the cluster. The mean value $\bar{x}$ and the variance σ_x^2 of **X** are defined as follows:

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N} x_i, \tag{1a}$$

$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{N} \|x_i-\bar{x}\|^2. \tag{1b}$$

The error caused by clustering (quantization error) of the continuous numerical data can be defined as *within-class variance* W_m (Mean Square Error, MSE):

$$W_m=\frac{1}{N}\sum_{j=1}^{M}\sum_{x_i\in C_j} \|x_i-c_j\|^2. \tag{2}$$

* Corresponding author.
E-mail addresses: alexander.kolesnikov@arbonaut.com (A. Kolesnikov), elena.trichina@laposte.net (E. Trichina), tuomo.kauranne@lut.fi (T. Kauranne).

http://dx.doi.org/10.1016/j.patcog.2014.09.017
0031-3203/

Thus, the objective of our studies is to find an optimal balance between the clustering (quantization) error W_m and the number of clusters M.

3. Known solutions

The problem can be solved by introducing a cost function that incorporates the number of clusters and the error in clustering [1–4]. A number of approaches have been proposed as solutions and these approaches can be categorized as being based on the use of *optimization-like criteria*, *difference-like criteria*, and other methods [1,3]. When using *optimization-like criteria*, the optimal number of clusters is given by the minimization (maximization) of a certain cost function. With *difference-like criteria*, sharp changes in the values of the cost function are found by analysis of the difference or ratio of chosen cost function value for sequential points.

3.1. Optimization-like criteria

One of the most popular optimization-like criteria, CH [5] is based on the ratio of between-class B_m and within-class variances W_m

$$CH(m)=\frac{B_m\cdot(N-m)}{W_m\cdot(m-1)}, \tag{3a}$$

$$M=\arg\max\{CH(m)\}. \tag{3b}$$

Here *between-class variance* B_m is defined as follows:

$$B_m=\sigma_x^2-W_m=\frac{1}{N}\sum_{j=1}^{m} n_j\cdot\|c_j-\bar{x}\|^2, \tag{4}$$

where n_j is the number of points in the cluster C_j.

A similar criterion *TWH* was mentioned in Ref. [6] as a possible modification of the criterion *CH*:

$$TWH(m)=\frac{B_m\cdot(N-m)}{W_m\cdot m}, \tag{5a}$$

$$M=\arg\max\{TWH(m)\}. \tag{5b}$$

To compare the criterion *CH* with other criteria, we will use the inverted criterion *CH**:

$$CH^*(m)=\frac{W_m\cdot(m-1)}{B_m\cdot(N-m)}, \tag{6a}$$

$$M=\arg\max\{CH^*(m)\}. \tag{6b}$$

The following slight modification of the inverted *CH** criterion was presented in Refs. [7,8] and later studied in Ref. [9]:

$$ZXF(m)=\frac{W_m\cdot m}{B_m}, \tag{7a}$$

$$M=\arg\max\{ZXF(m)\}. \tag{7b}$$

The so-called *heuristic mean-square-error* was used in Ref. [10]

$$LA(m)=W_m\cdot(m+1)^{1/2}, \tag{8a}$$

$$M=\arg\max\{LA(m)\}. \tag{8b}$$

A more general criterion for D-dimensional data clustering was proposed in Ref. [11]

$$Xu(m)=W_m\cdot m^{2/D}, \tag{9a}$$

$$M=\arg\max\{Xu(m)\}. \tag{9b}$$

With *optimization-like* criteria, the number of clusters in $\mathbf{X}$ is given by the minimum of the corresponding cost function *LA*, *ZHF*, *Xu*, *and CH**.

3.2. Difference-like criteria

The *difference-like criterion KL* is close to the criterion *Xu*, but search for the number of clusters is based on the ratio of differences of the criterion [12]

$$KL(m)=\left|\frac{W_{m-1}\cdot(m-1)^{2/D}-W_m\cdot m^{2/D}}{W_m\cdot m^{2/D}-W_{m+1}\cdot(m+1)^{2/D}}\right|, \tag{10a}$$

$$M=\arg\max\{KL(m)\}. \tag{10b}$$

Based on the Information Theory approach, the following *difference-like criterion SJ* was proposed in Ref. [13]:

$$SJ(m)=W_{m-1}^{-2/D}-W_m^{-2/D}, \tag{11a}$$

$$M=\arg\max\{SJ(m)\}. \tag{11b}$$

The *difference-like criteria KL* and *SJ* are usually less accurate than algorithms based on *optimization-like criteria* [1]. “Knee” and “elbow” detection algorithms, and other *difference-like criteria* with 1st and 2nd difference calculation are too sensitive to unavoidable variations of the cost function caused by noise in the calculated quantization error [1].

In addition to the above-mentioned criteria based only on the clustering error, methods exist with cost functions that include a heuristic penalty function [14–21]. The main drawback with these criteria is that their results depend on the heuristic function in use.

4. Novel parametric criterion

To overcome such undesirable cost function dependence as displayed by the methods listed above, we introduce a broad parameterized family of cost functions. Moreover, the parameter that identifies particular members of the cost function family is closely related to the data at hand, yet can be determined automatically.

4.1. Parametric modeling of the quantization error

The clustering error as a function of the number m of clusters is called *the rate–distortion* (*R–D*) *curve*. In our case, this is a *within-class variance* (Mean Square Error) W_m. Examples of *R–D* curves for two data sets are given in Fig. 1. In theory, the *R–D* function is always a monotone decreasing function: adding more clusters causes the distortion to decrease. When the number of centroids is small, the error first decreases dramatically, then more slowly, until it flattens to almost zero as the number of clusters approaches the number of data points. The actual shape of the *R–D* curve, or the behavior of the clusterization error as a function of the number of clusters, depends on the distribution of the data points in the D-dimensional space.

A model of clusterization (quantization) error is required to be able to find the number of clusters. The quantization error for uniformly distributed D-dimensional data forms the starting point for the modeling. In such cases, the quantization error with error measure L_2 for a uniform vector quantizer with m clusters is

$$W_m=\frac{Const}{m^{2/D}}. \tag{12}$$

Based on (12), the following parameterized model of the quantization error in the general case can be introduced:

$$\hat{W}_m=\frac{Const}{m^{2/a}}. \tag{13}$$

where a is the parameter of the model. The quantization error W_m can then be approximated by the following log-linear model with