Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 410–422
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Faster and more robust point symmetry-based K-means algorithm

Kuo-Liang Chung*,[1], Jhin-Sian Lin

Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, No. 43, Section 4, Keelung Road, Taipei, Taiwan 10672, ROC

Received 6 August 2004; received in revised form 20 May 2005; accepted 21 September 2005

Abstract

Based on the recently published point symmetry distance (PSD) measure, this paper presents a novel PSD measure, namely symmetry similarity level (SSL) operator for K-means algorithm. Our proposed modified point symmetry-based K-means (MPSK) algorithm is more robust than the previous PSK algorithm by Su and Chou. Not only the proposed MPSK algorithm is suitable for the symmetrical intra-clusters as the PSK algorithm does, the proposed MPSK algorithm is also suitable for the symmetrical inter-clusters. In addition, two speedup strategies are presented to reduce the time required in the proposed MPSK algorithm. Experimental results demonstrate the significant execution-time improvement and the extension to the symmetrical inter-clusters of the proposed MPSK algorithm when compared to the previous PSK algorithm.
© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Clustering; Inter-cluster; Intra-cluster; K-means algorithm; Point symmetry; Robustness; Speedup

1. Introduction

Clustering plays an important role in data analysis and pattern classification. It has many applications in codebook design [1,2], data compression [3], data mining [4], image segmentation [5], and so on. Clustering aims to partition a set of data points into some nonoverlapping subsets [6]. In the past three decades, many efficient clustering algorithms [1,7–15] have been developed. Among these developed clustering algorithms, the K-means algorithm is the oldest and the most popular one due to its simplicity and effectiveness.

In order to improve the performance of the K-means algorithm, several improved K-means algorithms have been developed in the past several years. In Ref. [1], instead of initially assigning each point to the closest center, Kövesi et al. presented a stochastic K-means algorithm to improve the clustering result. Based on the kd-tree data structure [16], Kanungo et al. [11] presented an improved K-means algorithm which can speed up the time performance while preserving the same clustering result as in the K-means algorithm. Based on code vector activity detection approach, Kaukoranta et al. [17] presented a faster K-means algorithm, which can be used to speed up the codebook construction by using the generalized Lloyd algorithm, and has the same clustering result as in the K-means algorithm. Considering the distribution of points for the case of symmetrical inter-clusters, Su and Chou [14] adopted the idea of symmetry feature [18–21] and presented an efficient point symmetry-based K-means (PSK) algorithm based on their proposed point symmetry distance (PSD) measure. Simulation results show that their proposed PSK algorithm has a better clustering result when compared to the K-means algorithm for symmetrical intra-cluster case. The motivation of this research are twofold: (1) presenting speedup strategies to reduce the execution time required in the previous PSK algorithm significantly and (2) presenting a new symmetry similarity level (SSL) operator to handle both the intra-cluster case and the inter-cluster case.

This paper first surveys the previous PSD measure [14] and explains why the PSD measure cannot handle the case

* Corresponding author. Tel.: +886 2 27376771; fax: +886 2 27376777.
E-mail address: klchung@cs.ntust.edu.tw (K.-L. Chung).
[1] Supported in part by the National Science Council of ROC under contracts NSC92-2213-E-011-079 and NSC93-2213-E-011-023.

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.09.015

of symmetrical inter-clusters well. Next, a novel SSL operator is presented to calculate the symmetry level between the data point p_i and the data point p_j relative to the cluster centroid c_k. When compared to the previous PSD measure, the proposed SSL operator not only can measure the orientation symmetry between p_i and p_j with respect to c_k as in the PSD measure, but also can measure the distance symmetry between the line segment $\overline{p_i c_k}$ and the line segment $\overline{c_k p_j}$. In addition, a simple constraint is suggested to enhance in the proposed SSL operator to handle both the case of symmetrical intra-clusters and the case of symmetrical inter-clusters. Further, two speedup strategies are presented to reduce the computation time required in the proposed modified PSK (MPSK) algorithm. In order to speed up the computation of the proposed SSL operator, a two-phase speedup strategy is presented. Since the proposed MPSK clustering algorithm includes the coarse-tuning step, which is realized by the K-means algorithm, a speedup strategy is also presented to improve the code vector activity detection approach [17] such that the coarse-tuning step can be performed in a faster way. Experimental results demonstrate the significant execution-time improvement and the extension to the symmetrical inter-clusters of the proposed MPSK algorithm when compared to the previous PSK algorithm by Su and Chou.

The remainder of this paper is organized as follows: In Section 2, the previous PSD measure is surveyed. In addition, one example is given to demonstrate the clustering power of the previous PSD measure for the case of symmetrical intra-clusters. In Section 3, the related problems that the PSD measure may occur are pointed out. In Section 4, the proposed SSL operator and the two-phase speedup strategy are presented. In Section 5, the proposed whole MPSK clustering algorithm is presented. In addition, a speedup strategy is described to speed up the coarse-tuning step in the MPSK algorithm. In Section 6, some experimental results are demonstrated to show the computational and robust advantages of the proposed MPSK clustering algorithm. In Section 7, some concluding remarks are addressed.

2. The past PSD measure

In this section, first the PSD measure by Su and Chou [14] is surveyed. Next, an example of symmetrical intra-clusters demonstrates the excellent applicability of the PSD measure.

In natural scenes, symmetry is an important feature [21,22]. Since the K-means algorithm cannot handle the case of intra-clusters well, recently, Su and Chou [14] presents a PSD measure and plugs it into the K-means algorithm to handle the case of intra-clusters efficiently.

Given N data points, say $\{p_i \mid \text{for} 1 \leqslant i \leqslant N\}$, using the K-means algorithm, let the temporary obtained K cluster centroids be denoted by $\{c_k \mid \text{for} 1 \leqslant k \leqslant K\}$. The PSD measure between the data point p_i and the data point p_j relative to the cluster centroid c_k is defined as

$$d_s(p_j, c_k) = \min_{\forall i \neq j \text{ and } 1 \leqslant i \leqslant N} \frac{\|(p_j - c_k) + (p_i - c_k)\|}{\|p_j - c_k\| + \|p_i - c_k\|}, \quad (1)$$

where $\|\cdot\|$ denotes the 2-norm distance.

An example is used to demonstrate how the PSD measure works well for the case of symmetrical intra-clusters. Fig. 1(a) illustrates two symmetrical intra-clusters, C_1 and C_2, where the data points are denoted by black dots and c_1 and c_2 are two centroids of the cluster C_1 and the cluster C_2, respectively. The positions of c_1 and c_2 are $c_1 = (5, 8)$ and $c_2 = (9.5, 8)$. p_1, p_2, and p_3 are three data points and their positions are $p_1 = (8, 7)$, $p_2 = (2, 9)$, and $p_3 = (12.5, 9.5)$, respectively. After running the K-means algorithm in Fig. 1(a), the data point p_1 in Fig. 1(a) would be assigned to the cluster C_2 because the data point p_1 is closer to c_2 than c_1. Fig. 1(b) shows the unsatisfactory clustering result by running the K-means algorithm in Fig. 1(a). In Fig. 1(b), the first unsatisfactory clustering result C_1 is denoted by squares and the second unsatisfactory clustering result C_2 is denoted by triangles. According to the visual inspection, the data point p_1 should be assigned to the cluster C_1 due to the symmetrical distribution of data points in C_1. The efficient PSD measure proposed by Su and Chou can indeed handle the case of symmetrical intra-clusters. By Eq. (1), for the data point p_1, it yields

$$d_s(p_1, c_1) = \frac{\|(p_1 - c_1) + (p_2 - c_1)\|}{\|(p_1 - c_1)\| + \|(p_2 - c_1)\|} = \frac{0}{\sqrt{10} + \sqrt{10}} = 0$$

and

$$d_s(p_1, c_2) = \frac{\|(p_1 - c_2) + (p_3 - c_2)\|}{\|(p_1 - c_2)\| + \|(p_3 - c_2)\|} = \frac{\sqrt{2.5}}{\sqrt{3.25} + \sqrt{11.25}} = 0.31.$$

Because $d_s(p_1, c_1) < d_s(p_1, c_2)$ and $d_s(p_1, c_1)$ is less than the specified threshold θ, e.g. $\theta = 0.18$ [14], the data point p_2 is said to be the most symmetrical point of p_1 relative to c_1, thus we have

$$p_2 = Arg\, d_s(p_1, c_1).$$

Consequently, assigning the data point p_1 to the cluster C_1 is a good decision. Fig. 1(c) depicts two satisfactory resulting clusters when applying the PSD measure to Fig. 1(a).

3. Possible problems occurred in the PSD measure

In this section, three observations are given to point out the three problems that the PSD measure may occur. The three possible problems existed in the PSD measure are (1) lacking the distance difference symmetry property, (2) leading to an unsatisfactory clustering result for the case of symmetrical

Download English Version:

https://daneshyari.com/en/article/531633

Download Persian Version:

https://daneshyari.com/article/531633

Daneshyari.com