On Hard $c$-Means Using Quadratic Penalty-Vector Regularization for Uncertain Data

Yasunori Endo$^1$, Arisa Taniguchi$^2$, Aoi Takahashi$^2$, and Yukihiro Hamasuna$^3$

$^1$Department of Risk Engineering, University of Tsukuba, Tennodai 1-1-1, Tsukuba, Ibaraki, 305-8573, Japan
endo@risk.tsukuba.ac.jp

$^2$Graduate School of Systems and Information Engineering, University of Tsukuba, Tennodai 1-1-1, Tsukuba, Ibaraki, 305-8573, Japan
{taniguchi, aoi}@soft.risk.tsukuba.ac.jp

$^3$Department of Informatics, Kinki University, 3-4-1, Kowakae, Higashiosaka, Osaka 577-8502, Japan
yhama@info.kindai.ac.jp

Abstract. Clustering is one of the unsupervised classification techniques of the data analysis. Data are transformed from a real space into a pattern space to apply clustering methods. However, the data cannot be often represented by a point because of uncertainty of the data, e.g., measurement error margin and missing values in data. In this paper, we introduce quadratic penalty-vector regularization to handle such uncertain data into hard $c$-means (HCM) which is one of the most typical clustering algorithms. First, we propose a new clustering algorithm called hard $c$-means using quadratic penalty-vector regularization for uncertain data (HCMP). Second, we propose sequential extraction hard $c$-means using quadratic penalty-vector regularization (SHCMP) to handle datasets whose cluster number is unknown. Moreover, we verify the effectiveness of our propose algorithms through some numerical examples.

1 Introduction

Clustering methods are known as very useful tools in many fields for data mining and we can find the construction of datasets through the clustering methods.

As computers have become sophisticated, the more studies for uncertainty are done. In the past, each datum handled by the computers was approximately represented as one point or value because of poor ability of the computers. However, the ability is now enough to handle the data with uncertainty called uncertain data and a lot of researchers have tried to handle original data from the viewpoint that the datum should be represented as not one point approximately but certain distribution exactly in a data space.
Whenever we construct the clustering methods for the uncertain data, we have one problem, that is, how should we represent the uncertainty of data?

To solve the above problems, we have proposed “tolerance” as a convenient tool to handle uncertain data and applied some of clustering algorithms [3–8]. In our proposed tolerance, tolerance vectors [3] and penalty ones [8, 2] play main role. Each uncertain datum is allowed to allocate any position by those vectors as far as the constraints for those vectors are satisfied and the position is derived as an optimal solution of a given objective function. Hence, we can say that this concept is in the framework of methodology of soft computing. Penalty vectors are similar to tolerance ones and the methods using penalty vectors become more flexible than tolerance vectors because no constraint for the vectors is needed. Then, we consider the penalty vectors in this paper.

By the way, sequential extraction hard c-means is proposed in Ref [9] which is based on noise clustering in Ref [11, 12]. The clustering does not need the initial number of clusters. The whole dataset is classified into one cluster and one noise dataset and data of the cluster are removed. The sequential extraction HCM classifies the dataset by repeating the above procedure and it can thus handle datasets whose cluster number is unknown.

The goal of this paper is to propose two new clustering algorithms for uncertain data based on hard c-means (HCM) [10], that is, hard c-means using quadratic penalty-vector regularization (HCMP), and sequential extraction hard c-means using quadratic penalty-vector regularization (SHCMP). We believe the proposed algorithms can classify the datasets which consists of uncertain data and whose cluster number is unknown.

2 Preliminaries

In this section, we explain the basic concept of tolerance and penalty, and hard c-means (HCM) clustering. First of all, we define some symbols. Each data is denoted \( x_k = (x_{k1}, \ldots, x_{kp})^T \in \mathbb{R}^p \) and the dataset \( X = \{x_1, \ldots, x_n\} \) is given. Each cluster \( C_i (i = 1, \ldots, c) \) has a cluster center \( v_i = (v_{i1}, \ldots, v_{ip})^T \in \mathbb{R}^p \). \( V \) means a set of cluster centers \( \{v_1, \ldots, v_c\} \). A membership grade for \( x_k \) to \( C_i \) which means belongingness of \( x_k \) to \( C_i \) is denoted by \( u_{ki} \). \( U \) means a partition matrix \( (u_{ki})_{1\leq k \leq n, 1 \leq i \leq c} \).

2.1 Tolerance and Penalty Vectors

In this paragraph, we explain two basic concepts, tolerance and penalty as the tools to handle uncertain data in the framework of optimization.

First, we describe the basic concept of tolerance. In general, a datum \( x \in \mathbb{R}^p \) with uncertainty is presented by some interval, i.e.,

\[
[x, \overline{x}] = [(x_1, \ldots, x_p)^T, (\overline{x}_1, \ldots, \overline{x}_p)^T] \subset \mathbb{R}^p.
\]