Expert Systems with Applications 37 (2010) 3318–3325

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A non-parametric heuristic algorithm for convex and non-convex data clustering based on equipotential surfaces

Farhad Bayat [a,*], Ehsan Adeli Mosabbeb [b], Ali Akbar Jalali [a], Farshad Bayat [c]

[a] Electrical Engineering Department, Iran University of Science and Technology, Tehran, Iran
[b] Computer Engineering Department, Iran University of Science and Technology, Tehran, Iran
[c] Computer Engineering Department, Azad University of Zanjan, Zanjan, Iran

ARTICLE INFO

Keywords:
Clustering
Classification
Convex
Non-convex
Potential functions
Unsupervised and point location

ABSTRACT

In this paper, using the concepts of field theory and potential functions a sub-optimal non-parametric algorithm for clustering of convex and non-convex data is proposed. For this purpose, equipotential surfaces, created by interaction of the potential functions, are applied. Equipotential surfaces are the geometric location of the points in the space on which the potential is constant. It means all points in each surface were affected the same by the field. Regarding this concept and other characteristics of equipotential surfaces, the outcome of this method will be an optimal solution for the clustering problem. But with regard to the existence of several parameters requiring to be set in the algorithm, finding the global optimal solution leads to a high computational complexity and therefore is not practical. Thus by applying some considerations and approximations, the resulting outcome will be a sub-optimal solution, while appropriate setting of the parameters causes the result to be closer to the global optimal solution. The advantage of this method is that it does not need any external parameter setting, such as number of clusters. To this end, an automatic parameter setting algorithm is suggested based on an optimal clustering index. Simulation results for a number of standard datasets, illustrate the superb performance of this method, especially for non-convexly scattered data. All mentioned characteristics of this method are widely demanded in different scientific areas. In this case it has been utilized in the well-known Point Location Problem (PLP) to reduce computational complexity.

1. Introduction

Clustering, as an unsupervised pattern classification method, has an important role in data analysis and refinement. Classification, in general, has a wide domain of applications. It includes all from biology and biomedical imaging sciences to military and archeology applications. According to the sensitiveness and limitations of some mentioned applications, such as medical and military applications, data with certain classes for use in supervised classification algorithms, are not simply available. Therefore, obtaining powerful and reliable unsupervised classification algorithms are very important. So in the recent decades, clustering problem as a tool for pattern analysis, classification, decision making and information extraction and retrieval, has attracted the attention of many researchers. Several approaches and points of view are presented in the literature. Each of these approaches is based on a certain criterion, and has its own advantages and disadvantages. In general, a comprehensive method and criterion for optimal clustering of any kind of data does not exist.

In Dubes et al. (1976), the comparison of different clustering algorithms was done using the criteria presented in Fisher et al. (1971). A review of the results of applying some existing limitations on data sources to enhance the clustering process performance was done by Titterington et al. (1985). Limitations applied in this method are based on the combination of an unknown number of probability density functions with multivariable Gaussian functions, in which clustering tries to extract the probability density functions and their parameters. The development of clustering applications is presented for pattern recognition by Anderberg (1973), image processing by Jain et al. (1996) and information retrieval by Salton (1991) and Rasmussen (1992).

General requirements of a data clustering system are: scalability, ability to recognize different shape, size and density clusters, robustness versus noise and disturbance, least number of input parameters, etc. Based on these criteria, efficiency and performance of clustering algorithms are determined. As a result, old hierarchical clustering algorithms are of very high computational complexity and qualitatively weak (George & et al., 1999). For instance, Complete-Link method is biased for spherical clusters and Single-Link

* Corresponding author. Tel.: +98 2177240487; fax: +98 2177240486.
E-mail addresses: fbayat@iust.ac.ir (F. Bayat), eadeli@iust.ac.ir (E.A. Mosabbeb), ajalali@iust.ac.ir (A.A. Jalali), farshad.bayat@gmail.com (F. Bayat).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.10.019

undertakes chaining (Oyang, 2001), while newer clustering techniques, combining hierarchical and partitioning methods (Gan, 2003; Gan et al., 2003), result in more quality and less computational complexity. A hybrid genetic fuzzy k-modes algorithm has been proposed by Gan, Wu, and Yang (2009). They have optimized fuzzy k-modes clustering algorithm using GA, to avoid being stuck in local optima of the k-modes clustering. Hsu and Huang (2008) have presented a learning based algorithm to cluster data with either categorical or numeric values. They have used a modified version of an adaptive resonance theory (ART) unsupervised neural network for this purpose. Wang and et al. (2009) have also presented a clustering algorithm based on the extension theory and genetic algorithm, EGA.

One of the weak sides of most clustering algorithms is the challenging case of non-convexly scattered data and the discussion of how each algorithm behaves in such a circumstance.

Different metrics for clustering analysis have been proposed in different works, such as entropy, purity and mutual information. Park and Jun (2009) have proposed a k-means-like algorithm, known as k-medoid, which is efficient in the complexity point of view. The authors have proposed a Random Index to evaluate the performance of the clustering. Wei, Lee, and Hsu (2003) have done an empirical comparison of some partition-based clustering techniques. They have introduced some characteristics of the data such as data size, number of clusters, cluster distinctness, cluster asymmetry and data randomness. They have analyzed the effects of changes in each of these parameters in the clustering results. Also Wu et al. (2009) have introduced some cluster validation measures to be used to evaluate k-means clusters. Normalized variation of information (VI), van Dongen criterion (VD) and Mirkin metric (M) are the measures used as cluster quality quantifiers. They showed that using these metrics can avoid bias in the clustering process. All these metrics and cluster quality measure used in these works are for convex data sets. Some studies (Mitra, Pal, & Siddiqi, 2003; Pal, Ghosh, & Uma Shankar, 2000) have introduced metrics also used for non-convex data clusters. Such a same metric is used in this paper and will be explained more, later.

In this paper, using the idea of potential functions and equipotential surfaces arising from fields' interaction a clustering method for both convex and non-convex data is presented. In this method, a potential function is assigned to each data sample. Then, by the fields' interactions in feature space and extraction of the equipotential surfaces the clustering procedure can be conducted. It is very important to know that most of the unsupervised classification techniques are based on the degree of similarity in data sample feature vector, such that members of a data class generally have the most similarity. Regarding this and the fundamental concepts of fields' theory in physic, applying the equipotential surfaces as the clusters discriminant boundaries the optimal solution would be achieved. This is because the equipotential surfaces introduce the geometric locations in the space on which all points have the same average membership or similarity to the class inside and outside the boundary. Thus, by proper setting of the reference potential level as the decision boundary, optimal solution could be obtained. Finally, using cluster measures a simple algorithm is presented which can set the reference potential level automatically. This approach determines the number of clusters itself. So, no parameters are left to be set externally. If one wants to determine the number of the clusters, he/she can change the reference potential by trial and error. Moreover, the proposed method enables us to construct a hierarchical non-parametric data clustering which is widely demanded in different areas. As an application, this method has been used to reduce computational complexity of the Point Location Problem (PLP) which is the most time consuming part in the Explicit Model Predictive Control (Bayat et al., 2009).

This paper is organized as follows: in Section 2 concepts and some definitions used in the subsequent sections are brought up. Then, in Section 3 potential functions are presented. After that using the potential functions concept, the clustering algorithm is demonstrated in Section 4. Finally, Section 5 illustrates some examples pondering the performance of the proposed algorithm.

2. Mathematical background and definitions

As described above, the method presented in this paper is based on the concept of field, which is one of the basics in Physics and has a vast number of applications. Different kinds of fields on Physics include magnetic, electric, gravity, and nucleus power fields. Although each of these instances has own different definitions, the common concept relating them is that instead of studying the mutual interaction between components (electric particles, for example), we can use the influence of the field on the components in that working set. In what follows, we utilize this simple concept to extract an optimal method for data clustering. The following definitions are evident and used in the algorithm definition.

Definition 1. Space $(D, \|\cdot\|)$ in linear vectors set $D \in \Re^n$ and real function $\|\cdot\| : D \to \Re_+$ are called a norm space if all following conditions fulfill:

(i) $\|\mathbf{X}\| \geq 0,\ \forall \mathbf{X} \in D$,
(ii) $\|\mathbf{X}\| = 0 \Longleftrightarrow \mathbf{X} = \mathbf{0},\ \forall \mathbf{X} \in D$,
(iii) $\|\alpha\mathbf{X}\| = |\alpha| \cdot \|\mathbf{X}\|,\ \forall \mathbf{X} \in D,\ \alpha \in \Re$,
(iv) $\|\mathbf{X} + \mathbf{Y}\| \leq \|\mathbf{X}\| + \|\mathbf{Y}\|,\ \forall \mathbf{X},\ \mathbf{Y} \in D$.

Definition 2. Assume $i = 1, \ldots, N,\ j = 1, \ldots, n,\ D \in \Re^n$ and $D = \{\mathbf{X}_i \in \Re^n | \mathbf{X}_i = (x_{i1}, \ldots, x_{in}),\ x_{ij} \in \Re\}$, then the norm space $(D, \|\cdot\|_2)$ is called a limited norm space relative to $\|\cdot\|_2$ if and only if there exists a scalar $0 < h < \infty$ such that:

$$\|\mathbf{X}_i\|_2 \leq h \text{ for all } \mathbf{X}_i \in D \text{ where } \|\mathbf{X}_i\|_2^2 = \sum_{j=1}^{n} x_{ij}^2.$$

Definition 3. Scalar function $V(\mathbf{X}) : \Re^n \to \Re$ is called a potential function if:

(i) $V(\mathbf{X})$ is a continuous smooth function in the given limited norm space (later we will find out that this space is in fact the feature vector space).
(ii) $V(\mathbf{X})$ is isotopic, i.e., it has symmetric behavior and characteristics in all dimensions.
(iii) If $V_i(\mathbf{X})$ is the potential function for component $\mathbf{X}_i$, increasing $\|\mathbf{X} - \mathbf{X}_i\|_2$ should cause $V_i(\mathbf{X})$ to decrease and $V_i(\mathbf{X}) \to 0$ for each $\|\mathbf{X} - \mathbf{X}_i\|_2 \to \infty$.

In what follows, assuming that the space regarding the feature vector in the clustering problem is a limited norm space, clustering algorithm based on the potential function could be extracted.

3. Establishing the proper potential function

Once more assume a vector space with limited norm for the feature vector:

$$D = \{\mathbf{X}_i \in \Re^n | \mathbf{X}_i = (x_{i1}, \ldots, x_{in}),\ x_{ij} \in \Re\}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, n \tag{1}$$

where $\mathbf{X}_i \in \Re^n$ is the feature vector for the ith sample, and $x_{ij} \in \Re$ is the jth feature in the ith feature vector. Also 'N' is the number of patterns and 'n' the number of features for each pattern.