A Modified K-Means Clustering with a Density-Sensitive Distance Metric

Ling Wang, Liefeng Bo, and Licheng Jiao

Institute of Intelligent Information Processing, Xidian University
Xi'an 710071, China
{wliiip, blf0218}@163.com, lchjiao@mail.xidian.edu.cn

Abstract. The K-Means clustering is by far the most widely used method for discovering clusters in data. It has a good performance on the data with compact super-sphere distributions, but tends to fail in the data organized in more complex and unknown shapes. In this paper, we analyze in detail the characteristic property of data clustering and propose a novel dissimilarity measure, named density-sensitive distance metric, which can describe the distribution characteristic of data clustering. By using this dissimilarity measure, a density-sensitive K-Means clustering algorithm is given, which has the ability to identify complex non-convex clusters compared with the original K-Means algorithm. The experimental results on both artificial data sets and real-world problems assess the validity of the algorithm.

Keywords: K-Means clustering, distance metric, dissimilarity measure.

1 Introduction

Data clustering has always been an active and challenging research area in machine learning and data mining. In its basic form the clustering problem is defined as the problem of finding homogeneous groups of data points in a given data set, each of which is referred to as a cluster. Numerous clustering algorithms are available in the literature. Extensive and good overviews of clustering algorithms can be found in the literature [1]. One of the earliest and most popular methods for finding clusters in data used in applications is the algorithm known as K-Means, which is a squared error-based clustering algorithm [2]. The K-Means algorithm is very simple and can be easily implemented in solving many practical problems. There exist a lot of extended versions of K-Means such as K-Median [3], adaptive K-Means [4],and global K-Means [5].

In order to mathematically identify clusters in a data set, it is usually necessary to first define a measure of dissimilarity which will establish a rule for assigning points to the domain of a particular cluster center. The most popular dissimilarity measure is the Euclidean distance. By using Euclidean distance as a measure of dissimilarity, the K-Means algorithm has a good performance on the data with compact super-sphere distributions, but tends to fail in the data organized in more complex and unknown shapes, which indicates that this dissimilarity measure is undesirable when clusters have random distributions.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 544–551, 2006.

As a result, it is necessary to design a more flexible dissimilarity measure for the K-Means algorithm. Su and Chou [6] proposed a nonmetric measure based on the concept of point symmetry, according to which a symmetry-based version of the K-Means algorithm is given. This algorithm assigns data points to a cluster center if they present a symmetrical structure with respect to the cluster center. Therefore, it is suitable to clustering data sets with clear symmetrical structure. Charalampidis [7] recently developed a dissimilarity measure for directional patterns represented by rotation-variant vectors and further introduced a circular K-Means algorithm to cluster vectors containing directional information, which is applicable for textural images clustering.

In this paper, through observing the characteristic property of data clustering, we design a novel data-dependent dissimilarity measure, namely, density-sensitive distance metric, which has the property of elongating the distance among points in different high density regions and simultaneously shortening that in the same high density region. Thus, this distance metric can reflect the characters of data clustering. Introducing the dissimilarity measure into the K-Means clustering, a density-sensitive K-Means clustering algorithm (DSKM) is proposed. Compared with the original K-Means clustering, DSKM can be used to group a given data set into a set of clusters of different geometrical structures.

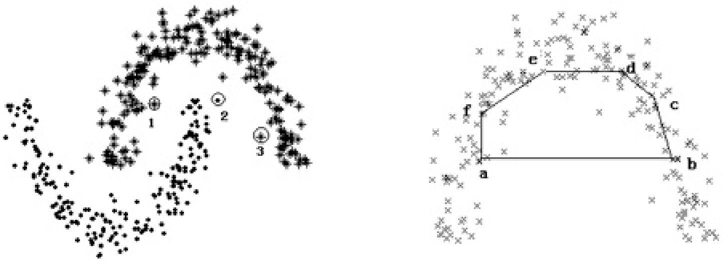
2 Density-Sensitive Distance Metric

As we all known, no meaningful cluster analysis is possible unless a meaningful measure of distance or proximity between pairs of data points has been established. Most of the clusters can be identified by their location or density characters. Through a large mount of observation, we have found the following two consistency characters of data clustering, which are coincident with the prior assumption of consistency in semi-supervised learning [8].

- *L*ocal consistency refers that data points close in location will have a high affinity.
- *G*lobal consistency refers that data points locating in the same manifold structure will have a high affinity.

For real world problems, the distributions of data points take on a complex manifold structure, which results in the classical Euclidian distance metric can only describe the local consistency, but fails to describe the global consistency. We can illustrate this problem by the following example. As shown in Fig.1(a), we expect that the affinity between point 1 and point 3 is higher than that of point 1 and point 2. In other words, point 1 is much closer to point 3 than to point 2 according to some distance metric. In terms of Euclidian distance metric, however, point 1 is much closer to point 2, thus without reflecting the global consistency. Hence for complicated real world problems, simply using Euclidean distance metric as a dissimilarity measure can not fully reflect the characters of data clustering.

In the following, we will consider how to design a novel dissimilarity measure with the ability of reflecting both the local and global consistency. As an example,



(a) (b)

Fig. 1. (a) Looking for a distance metric according to which point 1 is closer to point 3 than to point 2; (b) $\overline{af} + \overline{fe} + \overline{ed} + \overline{dc} + \overline{cb} < \overline{ab}$

we can observe from the data distribution in Fig. 1(a) that data points in the same cluster tend to lie in a region of high density, and there exists a region of low density where there are a few data points. We can design a data-dependent dissimilarity measure in terms of that character of local data density.

At first, data points are taken as the nodes V of a weighted undirected graph $G = (V, E)$. Edges $E = W_{ij}$ reflect the affinity between each pair of data points. We expect to design a dissimilarity measure that ascribes high affinity to two points if they can be linked by a path running along a region of high density, and a low affinity if they cannot. This concept of dissimilarity measure has been shown in experiments to lead to significant improvement in classification accuracy when applied to semi-supervised learning [9], [10]. We can illustrate this concept in Fig. 1(a), that is, we are looking for a measure of dissimilarity according to which point 1 is closer to point 3 than to point 2. The aim of using this kind of measure is to elongate the paths cross low density regions, and simultaneously shorten those not cross.

To formalize this intuitive notion of dissimilarity, we need first define a so-called density adjusted length of line segment. We have found a property that a distance measure describing the global consistency of clustering does not always satisfy the triangle inequality under the Euclidean distance metric. In other words, a direct connected path between two points is not always the shortest one. As shown in Fig. 1(b), to describe the global consistency, it is required that the length of the path connected by shorter edges is smaller than that of the direct connected path, i.e. $\overline{af} + \overline{fe} + \overline{ed} + \overline{dc} + \overline{cb} < \overline{ab}$.

Enlightened by this property, we define a density adjusted length of line segment as follows.

Definition 1. Density adjusted length of line segment
A *density adjusted length of line segment* is defined as

$$L(x_i, x_j) = \rho^{dist(x_i, x_j)} - 1. \tag{1}$$

where $dist(x_i, x_j)$ is the Euclidean distance between x_i and x_j; $\rho > 1$ is the flexing factor.

Obviously, this formulation possesses the property mentioned above, thus can be utilized to describe the global consistency. In addition, the length of line segment between two points can be elongated or shortened by adjusting the flexing factor ρ.

According to the density adjusted length of line segment, we can further introduce a new distance metric, called density-sensitive distance metric, which measures the distance between a pair of points by searching for the shortest path in the graph.

Definition 2. Density-sensitive distance metric. Let data points be the nodes of graph $G = (V, E)$, and $p \in V^l$ be a path of length $l =: |p|$ connecting the nodes p_1 and $p_{|p|}$, in which $(p_k, p_{k+1}) \in E$, $1 \leq k < |p|$. Let $P_{i,j}$ denote the set of all paths connecting nodes x_i and x_j. The *density-sensitive distance metric* between two points is defined to be

$$D_{ij} = \min_{p \in P_{i,j}} \sum_{k=1}^{|p|-1} L(p_k, p_{k+1}). \tag{2}$$

Thus D_{ij} satisfies the four conditions for a metric, i.e. $D_{ij} = D_{ji}$; $D_{ij} \geq 0$; $D_{ij} \leq D_{ik} + D_{kj}$ for all x_i, x_j, x_k; and $D_{ij} = 0$ *iff* $x_i = x_j$.

As a result, the density-sensitive distance metric can measure the geodesic distance along the manifold, which results in any two points in the same region of high density being connected by a lot of shorter edges while any two points in different regions of high density are connected by a longer edge through a region of low density. This achieves the aim of elongating the distance among data points in different regions of high density and simultaneously shortening that in the same region of high density. Hence, this distance metric is data-dependent, and can reflect the data character of local density, namely, what is called density-sensitive.

3 Density-Sensitive K-Means Algorithm

According to the analysis in the previous section, we can conclude that the choice of dissimilarity measure will greatly influence the clustering results. It is natural to consider utilizing the density-sensitive distance metric as a dissimilarity measure in the original K-Means algorithm and expect to have better performance. Consequently, we have a modified K-Means algorithm, called density-sensitive K-Means algorithm (DSKM), whose detailed procedure is summarized in Alg. 1. DSKM is a trade-off of flexibility in clustering data with computational complexity. The main computational cost for the flexibility in detecting clusters lies in searching for the shortest path between each pair of data points.

4 Simulations

In order to validate the clustering performance of DSKM, here we give the experimental results on artificial data sets and real-world problems. The results

Algorithm 1. Density-Sensitive K-Means Algorithm

```
Input  : $n$ data points $\{x_i\}_{i=1}^n$; cluster number $k$; maximum iteration number
         $tmax$; stop threshold $e$.
Output: Partition of the data set $C_1, \ldots, C_k$.
  1. Initialization. Randomly choose $k$ data points from the data set to initialize $k$
     cluster centers;
  2. For any two points $x_i, x_j$, compute the density-sensitive distance in terms of
     $D_{ij} = \min_{p \in P_{i,j}} \sum_{k=1}^{|p|-1} L(p_k, p_{k+1})$;
  3. Each point is assigned to the cluster which the density-sensitive distance of
     its center to the point is minimum;
  4. Recalculate the center of each cluster;
  5. Continuation. If no points change categories or the number of iterations has
     reached the maximum number $tmax$, then stop. Otherwise, go to step 2.
```

will be compared with the original K-Means algorithm. In all the problems, the desired clusters number is set to be known in advance, and the maximum iterative number is set to 500, the stop threshold 10^{-5}. Both algorithms are run 10 times for each of the candidate parameters and the average result is finally output.

4.1 Artificial Data Sets

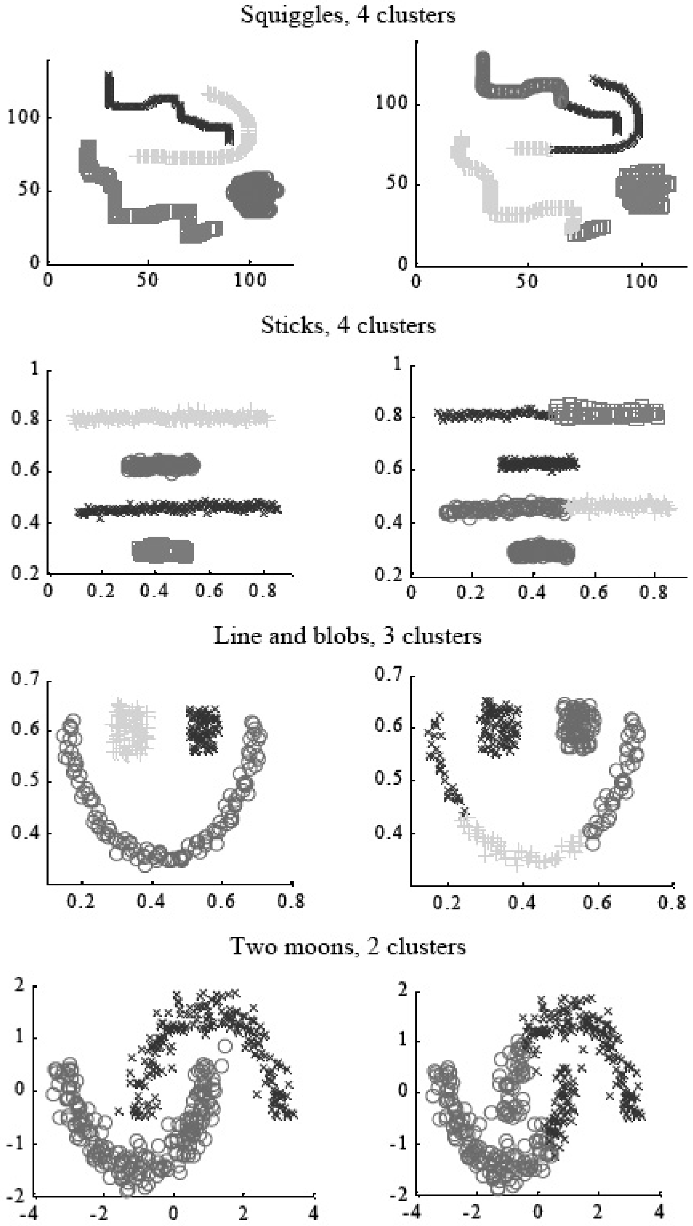
In this section, we evaluate the performance of DSKM on some artificial data sets. Here, we construct four ”challenge problems” with different distributions of data points. Clustering results obtained by DSKM and KM are shown in Fig. 2. We can see clearly that KM fails in obtaining the correct clusters for all the problems. This is due to the complex structure of data points, which does not satisfy convex distribution. On the other hand, DSKM can successfully recognize these complex clusters, which indicates the density-sensitive distance metric is very suitable to measure the complicated clustering structure.

We need to emphasize that the correct clusters are achieved by DSKM in a wider range of parameters. We choose the two moons problem as an example. With any flexing factor satisfying $1 < \rho < e^{18}$, DSKM can obtain the desired clusters. Therefore, DSKM is not sensitive to the choice of free parameter.

4.2 Real-World Data Sets

We have conducted experiments on USPS handwritten digit data set and three real data sets from UCI machine learning repository, i.e. Iris, Breast Cancer, and Heart [11]. USPS data set contains 9298 16×16 gray images of handwritten digits (7291 for training and 2007 for testing). The test set is taken as the clustering data, and we perform experiments recognizing three groups of digits, i.e. $0, 8$; $3, 5, 8$ and $0, 2, 4, 6, 7$.

Now that the ”true” clustering is available, we can use the class message to evaluate the clustering performance of the algorithms. Let the true clustering be $\Delta^{true} = \{C_1^{true}, C_2^{true}, \ldots, C_{k_{true}}^{true}\}$ and the clustering produced be



Squiggles, 4 clusters

Sticks, 4 clusters

Line and blobs, 3 clusters

Two moons, 2 clusters

Fig. 2. Four "challenge problems" are successfully clustered by DSKM (left). Cluster membership is indicated by different marker symbol and colors. KM fails in all the problems (right).

$\Delta = \{C_1, C_2, \ldots, C_k\}$. $\forall i \in [1, \ldots, k_{true}]$, $j \in [1, \ldots, k]$, *Confusion*(i, j) denotes the number of same data points both in the true cluster C_i^{true} and in the cluster C_j produced. Then, the clustering error (CE) is defined as

$$CE(\Delta, \Delta^{true}) = \frac{1}{n} \sum_{i=1}^{k_{true}} \sum_{j=1 i \neq j}^{k} \mathit{Confusion(i,j)}. \tag{3}$$

where n is the total number of data pints. Note that there exists a renumbering problem. For example, cluster 1 in the true clustering might be assigned cluster 3 in the clustering produced and so on. To counter that, the CE is computed for all possible renumbering of the clustering produced, and the minimum of all those is taken.

The best clustering performance, i.e. the smallest CE achieved by DSKM and KM on the four data sets is reported in Table 1, from which we can see that DSKM has a dominant performance on these real world data sets compared with the original KM.

Table 1. Performance Comparisons of DSKM and KM

Problem	Best CE	
	DSKM	KM
Iris	**0.106**	0.147
Breast Cancer	**0.235**	0.267
Heart	**0.142**	0.163
0,8	**0.025**	0.191
3,5,8	**0.146**	0.252
0,2,4,6,7	**0.113**	0.202

Finally, we can conclude from the simulations that DSKM not only has a significant improvement on the clustering performance compared with the original K-Means clustering algorithm, but also can be applied in the case where the distributions of data points are not compact super-spheres. And furthermore, the experimental results also indicate the general applicability of density-sensitive dissimilarity measure.

5 Conclusions

This paper presents a modified K-Means clustering based on a novel dissimilarity measure, namely, density-sensitive distance metric. The density-sensitive K-Means algorithm can identify non-convex clustering structures, thus generalizing the application area of the original K-Means algorithm. The experimental results on both artificial and real world data sets validate the efficiency of the modified algorithm.

Acknowledgement

Our research has been supported by the National Natural Science Foundation of China under Grant No. 60372050 and the National Grand Fundamental Research 973 Program of China under Grant No. 2001CB309403.

References

1. Xu R., Wunsch, D.: Survey of Clustering Algorithms. *IEEE Trans. Neural Networks.* 16 (2005) 645-678.
2. Hartigan, J.A., Wong, M.A.: A K-means clustering algorithm. *Applied Statistics.* 28 (1979) 100-108.
3. Bradley, P.S., Mangasarian, O.L., Street, W.N.: Clustering via concave minimization. In: *Advances in Neural Information Processing Systems 9.*MIT Press, Cambridge, MA (1997) 368-374.
4. Chinrungrueng, C., Sequin, C.H.: Optimal adaptive K-means algorithm with dynamic adjustment of learning rate. *IEEE Trans Neural Network.* 1 (1995) 157-169.
5. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. *Pattern Recognition.* 36 (2003) 451-461.
6. Su, M-C., Chou, C-H.: A modified version of the K-Means algorithm with a distance based on cluster symmetry. *IEEE Transactions on Pattern Anal. Machine Intell..* 23 (2001) 674-680.
7. Charalampidis, D.: A modified K-Means algorithm for circular invariant clustering. *IEEE Transactions on Pattern Anal. Machine Intell..* 27 (2005) 1856-1865.
8. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Scholkopf, B: Learning with Local and Global Consistency. In: Thrun, S., Saul, L., Scholkopf B, Eds., *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA, USA (2004) 321-328.
9. Bousquet, O., Chapelle,O., Hein,M.: Measure based regularization. In: Thrun, S., Saul, L., Scholkopf B, Eds., *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA, USA (2004).
10. Blum, A, Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of the Eighteenth International Conference on Machine Learning(ICML) 18, (2001) 19-26.
11. Blake, C.L., Merz., C.J.: UCI repository of machine learning databases. Technical report, University of California, Department of Information and Computer Science, Ir-vine, CA (1998).