On Clustering Algorithm Based on Probabilistic Similarity

YAMASHIRO Makito¹, ENDO Yasunori², HAMASUNA Yukihiro¹

¹ Graduate school of Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan
E-Mail:[yamashiro,yhama]@soft.risk.tsukuba.ac.jp

² Faculty of Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan
E-Mail:endo@risk.tsukuba.ac.jp

Abstract—In this paper, we propose a new clustering algorithm based on probabilistic similarity. The probabilistic similarity is formed by introducing the concept of probability into conventional similarity. First, we define the probabilistic similarity. Next, we show some examples probabilistic similarity functions. Third, we consider a objective function with the probabilistic similarity. Furthermore, we construct a new clustering algorithm based on the probabilistic similarity by using the optimal solutions which maximize the objective function. Last, we show some numerical examples to verify the effectiveness of the proposed algorithm.

Index Terms—Probabilistic Similarity, Clustering, Optimization.

I. INTRODUCTION

In recent years, many studies to make the computer be able to recognize the complicated data like as human have been developed. Moreover, retrieval speed of information becomes faster by using the data mining, e.g., pattern recognition, data classification and so on. Data classification which is one of the typical technique of data mining, is often used in the field of engineering, medical treatment and so on. The methodology of data classification consists of the two methods, one is with external criterion and the other is without it. The classification method with external criterion is called clustering. Clustering consists of two method, one is hierarchical and the other is non-hierarchical. Agglomerative Hierarchical Clustering (AHC), which is the most typical hierarchical one, classifies data into clusters by combing data. The clustering result is obtained as dendrogram. On the other hand, non-hierarchical clustering classifies the data into c clusters by optimizing a objective function. The value of c is given in advance. The concept of cluster center plays important role in non-hierarchical clustering. The most typical method of non-hierarchical clustering is fuzzy c-means (FCM), which is constructed by introducing the concept of membership grade into hard c-means (HCM)[1], [2]. In both the clustering, the concept of similarity is extremely important. The selection of the measure strongly influence not only the clustering results but also structure of the algorithm. This paper proposes a new clustering algorithm in which the measure of similarity is defined on probabilistic spaces. We call the measure probabilistic similarity. The probabilistic similarity is very unique from the viewpoint that the similarity between object pairs is represented by probability.

First, probabilistic similarity is defined based on the conventional similarity and some examples are shown. Next, an objective function is proposed using the probabilistic similarity. Third, the optimal solutions which are maximize the objective function are derived and a new clustering algorithm is constructed by the solutions. Moreover, the effectiveness of the proposed algorithm is verified through some numerical examples. Last, the conclusion is described.

II. PRELIMINARIES

In this section, we show the conventional similarity and propose probabilistic similarity. Next, probabilistic similarity is defined.

A. Probabilistic Similarity

In this subsection, we show the definitions of similarity and probabilistic similarity.

1) Axiom of Similarity: Axiom of similarity is shown as follows. X is arbitrary sets and \( ^{i}p,q \in X \). Any function \( s \) satisfying (1)~(2) is a similarity on \( X \).

\[
s(p, p) \geq s(p, q) \quad (1)
\]

\[
s(p, q) = s(q, p) \quad (2)
\]

2) Probabilistic Similarity: The definition of probabilistic similarity \( P \) based on probabilistic metric [3] is shown as follows. \( X \) is arbitrary sets and satisfied \( ^{i}p,q,r \in X, \; ^{i}x,y \in [0, +\infty) \). Any function \( P \), satisfying (3)~(7) is probabilistic similarity on \( X \).

\[
s(p, p) \geq s(p, q) \quad (1)
\]

\[
s(p, q) = s(q, p) \quad (2)
\]

\[
s^{i}(p, x) = s^{i}(p, y) \quad (3)
\]

\[
s^{i}(p, x) = s^{i}(q, y) \quad (4)
\]

\[
s^{i}(p, x) = s^{i}(x, q) \quad (5)
\]

\[
s^{i}(p, x) = s^{i}(y, q) \quad (6)
\]

\[
s^{i}(p, x) = s^{i}(x, y) \quad (7)
\]
Probabilistic Similarity Function

In this section, we propose new clustering algorithm based on probabilistic similarity. First, we propose a new objective function. Next, we construct new clustering algorithm by maximizing the objective function. For convenience, the new clustering algorithm is called CPS (Clustering based on Probabilistic Similarity).

A. Objective Function

In this subsection, we consider the following objective function.

\[ J(V, L) = -\sum_{k=1}^{n} \sum_{i=1}^{c} P_{s}(x_k, v_i, l_{ki}) \log P_{s}(x_k, v_i, l_{ki}) - \lambda \sum_{k=1}^{n} \sum_{i=1}^{c} l_{ki} \]  

(13)

Let \( x_k \) be data and \( G_i \) and \( v_i \) be cluster and cluster center, respectively. \( V = v_i \) and \( L = l_{ki} \). \( l_{ki} \geq 0 \) are constant. As above, we use \((10)\) as probabilistic similarity. Thus, \((13)\) can be rewritten as follows.

\[ J(V, L) = \sum_{k=1}^{n} \sum_{i=1}^{c} \left[ \frac{||x_k - v_i||^2}{P_{s}} \right] e^{-\frac{||x_k - v_i||^2}{P_{s}}} - \lambda \sum_{k=1}^{n} \sum_{i=1}^{c} l_{ki} \]  

(14)

We will construct a new algorithm which maximize \((14)\). Hence, we have to introduce the optimal solutions of \( v_i \) and \( l_{ki} \). In next subsection, we will define the solutions.

B. Derivation of The Optimal Solution

In this section, we derive the optimal solutions \( v_i \) and \( l_{ki} \) which maximize \((14)\). Here, \( J = J(V, L) \).

1) Optimal Solution about \( v_i \): We differentiate partially \((14)\) by \( v_i \).

\[ \frac{\partial J}{\partial v_i} = \sum_{k=1}^{n} \left[ \left( \frac{-2(x_k - v_i)}{P_{s}^{2}} \right) e^{-\frac{||x_k - v_i||^2}{P_{s}}} + \frac{2(x_k - v_i)}{P_{s}^{2}} \left( \frac{||x_k - v_i||^2}{P_{s}} e^{-\frac{||x_k - v_i||^2}{P_{s}}} \right) \right] = 0 \]  

(12)

In order to simplify the notation, \( P_s = e^{-\frac{||x_k - v_i||^2}{P_{s}}} \)

\[ \sum_{k=1}^{n} \left[ \frac{x_k - v_i}{P_{s}^{2}} (1 - s_{ki}) P_s \right] = 0 \]

From the above formulation, we can get optimal solution of \( v_i \) as follows.

\[ v'_i = \sum_{k=1}^{n} \left( \frac{P_{s}(1 - s_{ki})}{P_{s}^2} \right) x_k \]  

(15)

2) Optimal Solution about \( l_{ki} \): We differentiate partially \((14)\) by \( l_{ki} \).

\[ \frac{\partial J}{\partial l_{ki}} = -\frac{2||x_k - v_i||^2}{l_{ki}^2} e^{-s_{ki}} + \frac{||x_k - v_i||^2}{l_{ki}^2} e^{-s_{ki}} \left( \frac{2||x_k - v_i||^2}{l_{ki}^2} \right) - \lambda \]  

\[ = 0 \]

\[ \Rightarrow \left( \frac{2s_{ki} - 1}{l_{ki}^2} \right) ||x_k - v_i||^2 = \lambda \]

Therefore, from the above formulation, we can get optimal solution of \( l_{ki} \) as follows.

\[ l_{ki} = \left[ \frac{2||x_k - v_i||^2 P_s}{\lambda P_{s}^2 + 2||x_k - v_i||^2 P_s} \right]^{\frac{1}{2}} \]  

(16)

In this paper, we use \((10)\) as probabilistic similarity function.
C. CPS Algorithm

In this subsection, we propose CPS algorithm to maximize the objective function (14). The framework of proposed algorithm is alternative optimization [2] using (15) and (16). The CPS algorithm is shown as follows.

CPS Algorithm

Step. 1: Give the values of $\lambda$ and initial values of $\bar{V} = V$ and $\bar{L} = L$.

Step. 2: Set $\bar{V}$, then calculate the $V$ from (15). In addition, $\bar{V} = V$.

Step. 3: Set $\bar{L}$, then calculate the $L$ from (16). In addition, $\bar{L} = L$.

Step. 4: If $(\bar{V}, \bar{L})$ is convergence, end of algorithm. stop this algorithm. Otherwise, go back to Step. 2.

Convergence criterion of the algorithm is $||J_n - J_{n-1}|| < \epsilon$.

1) Clustering Using CPS Algorithm: This algorithm does not calculate the membership values of $x_k$ to $G_i$. Therefore, it is not easy to decide which cluster each data belongs to. However, instead of the membership values, the values of probabilistic similarity $P_s$ is obtained by the algorithm. Hence, we can assign each data to an adequate cluster by $P_s$ such as

$$x_k \in G_i \iff P_s(x_k, v_i, I_{ki}) = \max P_s(x_k, v_i, l_{ki})$$

However, the definition of probabilistic similarity is not sufficient to find the maximum of $P_s$. Thus, we add the following condition.

$$P_s(p, q, x) = P_s(p, r, y) \quad (||p - q|| > ||p - r||)$$

$$\Rightarrow P_s(p, q, x + y) < P_s(p, r, x + y) \quad (17)$$

We should notice that the condition is defined on the norm space. However, we can assume Euclidean space as data space in many cases like as the numerical example so that the condition is natural.

IV. NUMERICAL EXAMPLES

In this section, we show some numerical examples to verify the effectiveness of our CPS algorithm.

A. Artificial Datasets

We use an artificial datasets in Fig. 2 and Fig. 3. The Dataset 1 consists of 88 objects in the two dimensional Euclidean space. The Dataset 2 consists of 314 objects and the boundary between two clusters is nonlinear. We set random numbers in $[0, 1]$ on each initial center clusters $v_i$. Furthermore, the probabilistic similarity $l_{ki}$ is set random numbers in $[1, 10]$.

B. Result

We show a result using CPS algorithm for the datasets in Fig 4 ~ 7. CPS algorithm are repeated until the values of the objective function becomes maximum. We obtained the maximum value of the objective function through the alternative optimization with (15) and (16).

Each data $x_k$ is assigned to the cluster $G_i$, of which the probabilistic similarity $P_s$ between $x_k$ and $v_i$ is maximum.

We set $\lambda = 0.05$ and $\epsilon = 0.0001$ in Dataset 1. We can show that the data is classified into four clusters in Fig 4. Next, we set $\lambda = 0.0010$ and $\epsilon = 0.0001$ in Dataset 2. We can show that the data is classified into two clusters in Fig 5 and Fig 6. We illustrate Fig 7 with z-coordinate which means the variable of probabilistic similarity between each object and the optimal cluster.

C. Discussion

In this subsection, we consider the result by CPS algorithm. We know that the value of probabilistic similarity $P_s(x_k, v_i, I_{ki})$ comes to be close to 1 as the value $\lambda$ decreases. Moreover, we know the relation between $l_{ki}$ and $\lambda$ from (14), that is, the less the value $\lambda$ decreases, the more the value $l_{ki}$ increases and also $P_s(x_k, v_i, I_{ki})$. When we did not use (17), the object of which did not assigned the optimal cluster the same value of $P_s$ between different cluster. The result is shown in Fig 4. When we used (17), the one assigned optimal cluster.
Fig. 4. Result of Dataset 1 A.

Fig. 5. Result of Dataset 1 B.

The result is shown in Fig 5. We may say that to adapt (17) was a success. Another result is shown in Fig 6 and Fig 7. It is found from the result that CPS algorithm can classify the dataset of which the boundary is nonlinear. The values of $P_s$ near the classification boundary with larger the values of $l_{ki}$ is large. The classification boundary in Fig 7 is shown as overlapping of normal distribution.

V. CONCLUSION

In this paper, we proposed probabilistic similarity, and constructed new clustering algorithm based on probabilistic similarity. Moreover, we showed some numerical examples to discuss effectiveness of the proposed CPS algorithm. In future works, there are two issues to be developed in this study. First, we will apply for proposed algorithm to nonlinear data. Second, we will considering probabilistic similarity based on other norm spaces. e.g. $L_1$-norm.

REFERENCES