2008 International Symposium on Nonlinear Theory and its Applications
NOLTA'08, Budapest, Hungary, September 7-10, 2008

On a New Dissimilarity of Projection Correlation

ENDO Yasunori[†], UCHIDA Fuyuki[‡], HAMASUNA Yukihiro[†]

†Department of Risk Engineering, Faculty of Systems and Information Engineering,
University of Tsukuba, Ibaraki 305-8573, Japan
{endo@, yhama@soft.}risk.tsukuba.ac.jp
‡Information Creative Laboratory Inc.
1-14-9, Gotanga, Shinagawa, Tokyo 141-0022, Japan

Abstract—This paper proposes projection correlation which is a new dissimilarity measure for clustering. This dissimilarity has both of the features of cosine correlation and norm-based dissimilarity. The details of the features are discussed by using dissimilarity level curve. Moreover, the effectiveness of the proposed dissimilarity is verified in comparison with the conventional dissimilarities, i.e., squared Euclidean metric and cosine correlation with applying to agglomerative hierarchical clustering.

1. Introduction

The more effective technique of data mining becomes to be needed [1, 2], as data to be handled by computers more increase.

Cluster analysis, or clustering, is one of the methods of the data mining. We can classify the given data into some groups called clusters without any external criteria by clustering. Measures of similarity or dissimilarity, which are defined between data, are used instead of the external criteria in clustering [1, 3]. Roughly speaking, there are two methods in clustering, one is hierarchical method and the other is non-hierarchical one. We consider the hierarchical method in this paper. In the method, a pair of data which has the maximum (minimum) value of similarity (dissimilarity) is regarded as one cluster. Here, we show the definition of the measure of dissimilarity as follows:

Definition 1 (Dissimilarity) *d is called dissimilarity iff the function $d : \mathcal{R}^p \times \mathcal{R}^p \to \mathcal{R}$ satisfies the following equations from* (1) *to* (2) *for all x, y.*

$$d(x,x) \leq d(x,y) \tag{1}$$

$$d(x,y) = d(y,x) \tag{2}$$

The explanation of dissimilarity is omitted owing to the limited space.

As the examples of similarity or dissimilarity, we can consider cosine correlation $\frac{x \cdot y}{\|x\| \cdot \|y\|}$ or squared Euclidean metric $\|x-y\|^2$ on $\mathcal{R}^p$, respectively. Particularly, the cosine correlation is effective in the field of the information retrieval [4].

However, the measure has a problem, that is, the measure is calculated only based on the angle between the vectors of data. In other words, even if the distance between two points is long or short, the similarity takes same value. From the property, the value of similarity between the data near the origin is strangely calculated. Of course, the cosine correlation is available in some field, e.g. information retrieval. But we believe that it is not available in clustering.

To solve the problem, we propose a new measure of dissimilarity which is called projection correlation. Moreover, we apply the measure to an agglomerative hierarchical clustering (AHC) and verify the effectiveness of the proposed dissimilarity through some numerical examples.

2. Projection Correlation

In this section, we propose a new proposed dissimilarity called projection correlation. First, we define projection correlation and derive some equations to use in clustering algorithms. Second, we plot dissimilarity level curves to show the shape of classification.

2.1. Definition of Projection Correlation

Let $x_k \in \mathcal{R}^p$ $(k = 1 \sim n)$ be each data. We assume that each data x_k is replaced to $\tilde{x}_k$ using arithmetic average $\bar{x} = \frac{\sum_{k=1}^{n} x_k}{n}$, that is, $\tilde{x}_k = x_k - \bar{x}$.

Here, we consider a new relation between x_a and x_b of projection correlation in the following Eq. (3).

$$\hat{d}(x_a, x_b) = \|\tilde{x}_a\| - \|\tilde{x}_b\| \cos\phi_{ab} \tag{3}$$

ϕ_{ab} is the angle between the vectors of $\tilde{x}_a$ and $\tilde{x}_b$. Fig. 1 and 2 shows the illustrative concept of projection correlation. In the cases like Fig. 2, $\hat{d}(x_b, x_a) < 0$.

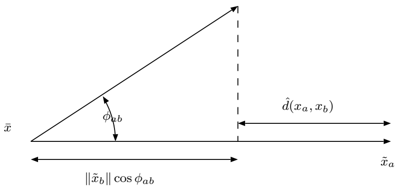

Figure 1: Projection Correlation $\hat{d}(x_a, x_b)$

We have to notice that $\hat{d}(x_a, x_b) \neq \hat{d}(x_b, x_a)$. Hence the function does not satisfies Eq. (2) and it is not a measure of dissimilarity. Therefore, we consider the following definitions which satisfy the definition of dissimilarity to use the projection correlation.

Definition 2 (Maximum Projection Correlation) *$d_{\max} : \mathcal{R}^p \times \mathcal{R}^p \to \mathcal{R}$ is called maximum projection*

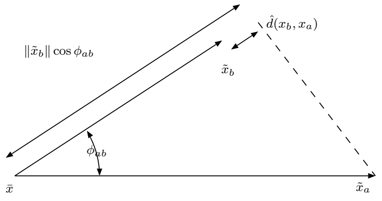

Figure 2: Projection Correlation $\hat{d}(x_b, x_a)$

correlation iff $d_{\max}$ satisfies the following equation:

$$d_{\max}(x_a, x_b) = \max\left(\hat{d}(x_a, x_b), \hat{d}(x_b, x_a)\right)$$

Here $\tilde{x}_k = x_k - \bar{x}$ $(k = a,\ b)$.

Definition 3 (Minimum Projection Correlation) *$d_{\min} : \mathcal{R}^p \times \mathcal{R}^p \to \mathcal{R}$ is called minimum projection correlation iff $d_{\min}$ satisfies the following equation:*

$$d_{\min}(x_a, x_b) = \min\left(\left|\hat{d}(x_a, x_b)\right|, \left|\hat{d}(x_b, x_a)\right|\right)$$

Definition 4 (Average Projection Correlation) *$d_{\text{avg}} : \mathcal{R}^p \times \mathcal{R}^p \to \mathcal{R}$ is called average projection correlation iff d_{avg} satisfies the following equation:*

$$d_{\text{avg}}(x_a, x_b) = \text{avg}\left(\hat{d}(x_a, x_b), \hat{d}(x_b, x_a)\right)$$

Definition 5 (Mean Square Projection Correlation) *$d_{\text{msq}} : \mathcal{R}^p \times \mathcal{R}^p \to \mathcal{R}$ is called mean square projection correlation iff d_{msq} satisfies the following equation:*

$$d_{\text{msq}}(x_a, x_b) = \text{avg}\left(\left(\hat{d}(x_a, x_b)\right)^2, \left(\hat{d}(x_b, x_a)\right)^2\right)$$

We show dissimilarity level curves of each correlation on Fig. 3, Fig. 4, Fig. 5 and Fig. 6.

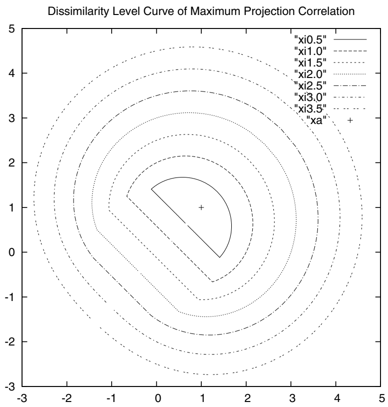

Figure 3: Dissimilarity Level Curve of Maximum Projection Correlation

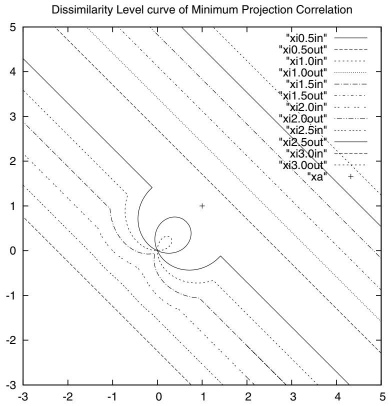

Figure 4: Dissimilarity Level Curve of Minimum Projection Correlation

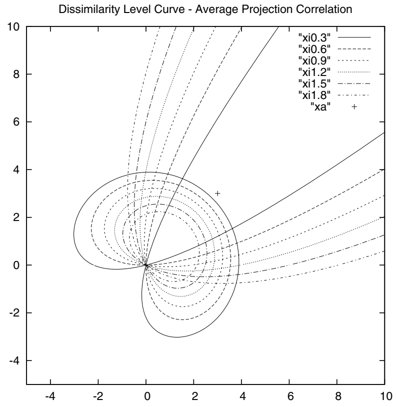

Figure 5: Dissimilarity Level Curve of Average Projection Correlation

3. Numerical Examples

In this section, we apply the proposed measures, Euclidean metric and cosine correlation to AHC Algorithms and verify the effectiveness of the measures.

3.1. AHC Algorithms

In this section, we show AHC algorithm and some methods to update the values of similarity. We don't show the case of dissimilarity.

Algorithm 1 (AHC Algorithm)

AHC-1 *Initialization*

$$G_i := \{x_i\}$$
$$d(G_a, G_b) := s(x_a, x_b)$$

AHC-2 *Merging*

$$\max\ s(G_a, G_b) \longrightarrow G' := G_a \cup G_b$$
$$C := C - 1$$

AHC-3 *Update Similarity or Dissimilarity*

Calculate $d(G', G_i)$ for all $G_i : G_i \neq G'$

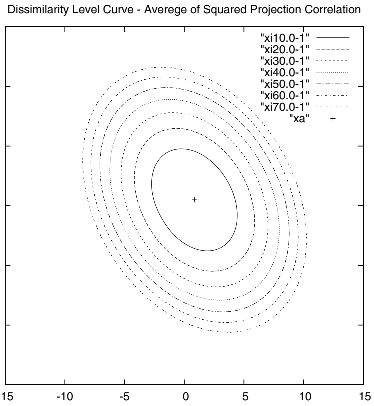

Figure 6: Dissimilarity Level Curve of Mean Square Projection Correlation

AHC-4 *Convergence Criterion*

If $C = 1$*, stop this algorithm, otherwise go back to* **AHC-2**.

3.2. Update Option

Here, we show three methods to update, that is, *single linkage method*, *complete linkage method* and *unweighted pair-group method using arithmetic averages* (it is called also *average linkage between the merged group*). The explanation is omitted owing to the limited space.

3.3. Numerical Examples

We show some numerical examples in this section.

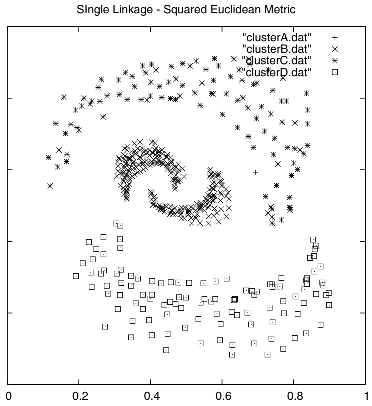

Figure 7: Squared Euclidean Metric using Single Linkage Method

4. Discussion

4.1. Maximum Projection Correlation and Mean Square Projection Correlation

When we apply AHC by using squared Euclidean metric to data set 1, we can not classify the data near the average of all data $\bar{x}$ (Fig. 7). By contrast, when we apply AHC by using maximum projection correlation to data set 1, we can classify data set 1 into four clusters (Fig. 8). Likewise, when we use mean square projection correlation, we can classify data set 1 into four cluster (Fig. 9).

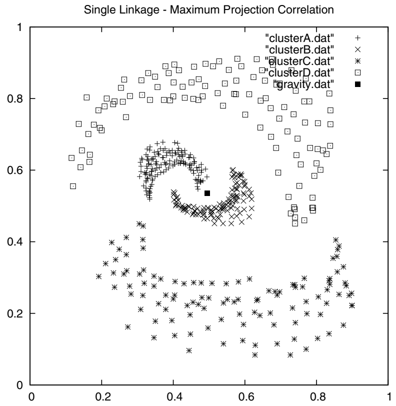

Figure 8: Maximum Projection Correlation using Single Linkage Method

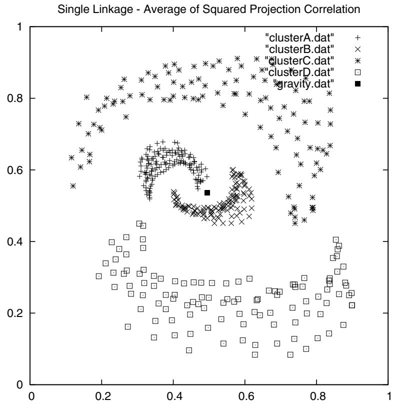

Figure 9: Mean Square Projection Correlation using Single Linkage Method

These results from the classification depend on configuration using by maximum projection correlation and mean square projection correlation. The data near $\bar{x}$ is classified sensitively. On the contrary, the data far $\bar{x}$ is classified loosely.

4.2. Average Projection Correlation

When we apply AHC by using average projection correlation and cosine correlation to data set 2, the results of Fig. 10 and Fig. 11 are obtained. The data near $\bar{x}$ are strangely classified in case of using both of the measures.

These results by the average projection correlation has the similar classification configuration to cosine correlation.

4.3. Minimum Projection Correlation

When we apply AHC by using squared Euclidean metric and the minimum projection correlation to data set 3, the results of Fig. 12 and Fig. 13 are obtained.

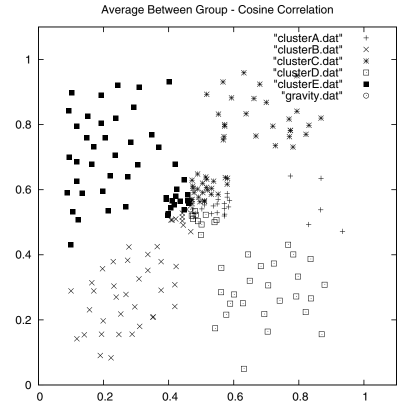

Figure 10: Cosine Correlation using Average Linkage Between the Merged Group

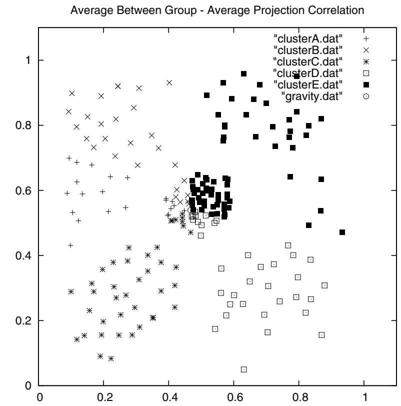

Figure 11: Average Projection Correlation using Average Linkage Between the Merged Group

By contrast to the case of squared Euclidean metric, minimum projection correlation has strange classification border.

5. Conclusion

In this paper, we have shown behavior of some dissimilarities derived from projection correlation. From these results, we have obtained conclusion: (1) We can anticipate that we apply maximum projection correlation to the data set that density of data, cluster configuration or cluster size are nonuniform. (2) Average projection correlation is not an effective method by contrast with cosine correlation because the classification configuration of average projection correlation is similar to the case of cosine correlation. In addition, the calculation of average projection correlation is complicate. (3) We can consider that minimum projection correlation is unsuitable for clustering.

In future works, we will discuss the following problems: (1) Changing the definition of $\bar{x}$, (2) Applying projection correlation to other clustering algorithms, (3)Constructing some clustering algorithms based on projection correlation.

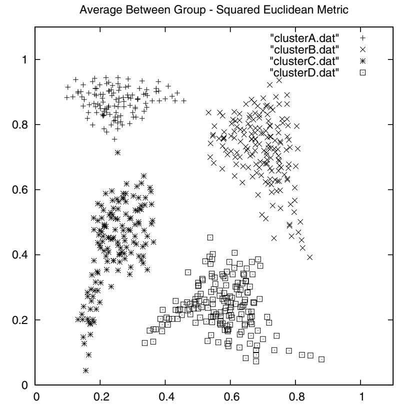

Figure 12: Squared Euclidean Metric using Average Linkage Between the Merged Group

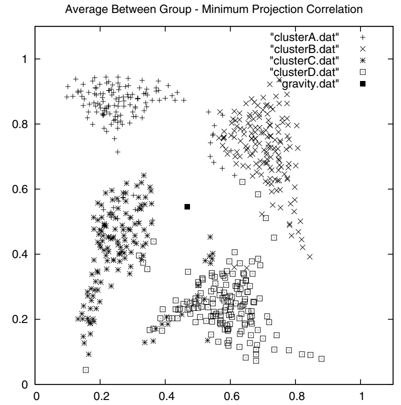

Figure 13: Minimum Projection Correlation using Average Linkage Between the Merged Group

Acknowledgment

This study is partly supported by the Grant-in-Aid for Scientific Research (C) (Project No.18500170) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] Fukuda Takeshi, Morimoto Yasuhiko, Tokuyama Takeshi : ‘Data Mining’, Kyouritsu Shuppan, 2001.

[2] P. Adriaans, D. Zantige : ‘Data Mining’, Addison Wesley Longman, 1998.

[3] J. A. Hartigan : ‘Clustering Algorithms ’, John Wiley & Sons, 1983.

[4] Kishida kazuaki, “Techniques of Document Clustering: A Review”, `http://wwwsoc.nii.ac.jp/mslis`, 2003.

[5] H. Charles Romesburg, ‘Cluster Analysis for Researchers’, 1989.