Similarity Measure Design on High Dimensional Data

THEERA-UMPON Nipon[1], Sanghyuk Lee[2]
[1.]Department of Electrical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200 Thailand
[2.]Department of Electrical and Electronics Engineering, Xi' an Jiaotong-Liverpool University, Suzhou, China

Abstract Designing of similarity on high dimensional data was done. Similarity measure between high dimensional data was considered by analysing neighbor information with respect to data sets. Obtained result could be applied to big data, because big data has multiple characteristics compared to simple data set. Definitely, analysis of high dimensional data could be the pre-study of big data. High dimensional data analysis was also compared with the conventional similarity. Traditional similarity measure on overlapped data was illustrated, and application to non-overlapped data was carried out. Its usefulness was proved by way of mathematical proof, and verified by calculation of similarity for artificial data example.

• **Key Words** : Similarity measure, high dimensional data, big data, neighbor information.

1. INTRODUCTION

Similarity measure design for high dimensional data has been emphasized as one of research topic for big data [1 - 3]. Due to the complex characteristic of big data, it is usually represented as high dimension, or with different dimension data. As an emerging research topic, big data has been focused due to fast development of technology and becoming big amount data society. Therefore, it is essential to pay attention for big data processing [4 - 6]. One of key characteristic of such a big data, high dimensional data characteristics also has been emphasized through knowledge discovery on non-spatial data [10, 11]. Application of big data has been extended to every aspect of our society, including manufacturing, e-business services, life science. Scientific research has been revolutionized by research on big data recently [7]. It has the potential to extend to not just research, but also education [8]. Mentioned big data has its characteristics such as big amounts, atypical, fast spreading.

Similarity measure already has been provideduseful information to clustering, pattern recognition to data sets. However, in order to get proper result, multi-dimensional or high dimension data are also needed to deal with another methodology compared to the conventional approach. Square error clustering algorithm has been used from late of 1960's [1]. And it was later modified to create the cluster program [2]. Research topic has been extensively studied in many areas such as statistics [3], machine learning [4, 5], pattern recognition [6], and image processing.

Extended research on high-dimensional data can be applied to security business including fingerprint and

*교신저자 : Lee, S. (Xi' an Jiaotong-Liverpool University)
(Tel : +86-512-1886-1415 E-mail: Sanghyuk.Lee@xjtlu.edu.cn)
접수일 2013년 03월 12일 수정일 2013년 03월 20일 게재확정일 2013년 03월 22일

iris identification, and image processing enhancement, and big data application recently. In a broad range of application areas including engineering and business, data is being collected at unprecedented quantity.

Then, distance between vectors can be organized by norms such as 1-norm, Euclidean-norm, etc. Similarity measure is also designed with distance norm explicitly. Similarity measure design problem for high-dimension needs more considerate approach. By norm definition more analytical evaluation can be provided. Conventionally, similarity measure has been designed based on distance measure between two considered data. Whereas similarity measure design with distance measure was considered distance information between two membership functions. High-dimensional data is illustrated over two-dimensional configuration, which constitutes number of data and number of dimension. In this literature, high dimensional multi data were considered to analyze its similarity and correlation. Data distribution was illustrated over 2-dimensional plane with dimension and number of data as coordinates. In order to calculate their closeness measure was considered.

In the following chapter, preliminary results on similarity measure are introduced. Difference between designed measures were proposed and compared. In Chapter 3, similarity measure design for high-dimensional data was discussed and provided by way of norm structure. Usefulness was also verified with proof. Example with financial fraud with high-dimensional data was analyzed. Finally, conclusions are followed in Chapter 4.

2. Similarity Measure for Data

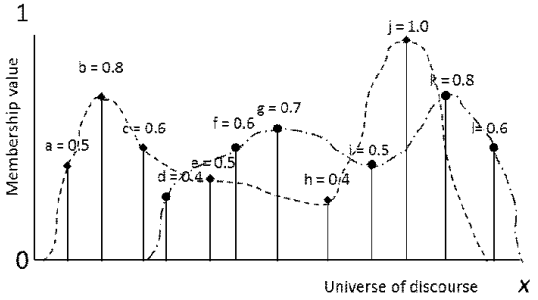
Properties of commutative, complementary, overlapped characteristics and triangular inequality feature are represented by the similarity measure. It was designed with distance measure such as Euclidian or Manhattan distance. Similarity measure can be represented as explicit structure with help of distance measure [10].

Among multiple similarity measures following measures are considered. Here, for any set $A, B \in F(X)$, if d satisfies Hamming distance measure, then

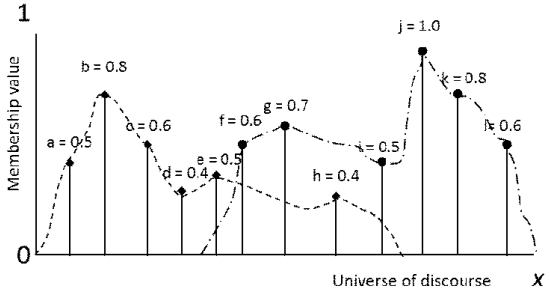
$$s(A,B) = 1 - d\big((A \cap B), (A \cup B)\big) \tag{1}$$

is the similarity measure between set A and B. Proofs are found in previous results [11,12].

Besides similarity measures of (1), numerous similarity measures are also possible [11 - 13]. For data distribution with Fig 1,



[Fig. 1] (a) Data distribution between circle and diamond



[Fig. 1] (b) Data distribution between circle and diamond

Two data pairs that constitute different distributions are considered in Fig. 1. Twelve data with six diamonds (♦) and six circles (●) are illustrated.Similarity degree between circles and

diamonds must be different between two figures because of different distribution. However, (1) represents

$$s(A,B) = 1 - d\big((A \cap B),(A \cup B)\big)$$

where, $A \cap B = \min(A,B) = 0$,and $A \cup B = \max(A,B) = A \; or \; B$. Hence,

$$s(A,B) = 1 - \frac{1}{N}\sum \max\,(A,B)$$

$$= 1 - \frac{1}{12}\sum (0.5 + 0.8 + 0.6 + 0.4 + 0.5 + 0.6 + 0.7 + 0.4 + 0.5 + 1 + 0.8 + 0.6)$$

$$= 0.38$$

the same result is obtained for Fig 1. (a) and (b).

Hence, similarity measure (1) is not proper for non-overlapped data distribution. Therefore, it is required to design similarity measure for non-overlapping data distribution. Consider the following similarity measure for non-overlapped data.

Theorem 2.1 For singletons or discrete data $a, b \in P(X)$, if d satisfied Hamming distance measure, then

$$s(A,B) = 1 - \sum_{a\in A, b\in B} |s_a - s_b| \qquad (2)$$

is similarity measure between singleton a and b. s_a and $s_b$satisfy $\mathrm{maxd}\,(a, adjacent\ of\ a\ in\ B)$ and $\mathrm{maxd}\,(b, adjacent\ of\ b\ in\ A)$, respectively. Proof is clear, hence we provide its concepts in here.

Where, $\mathrm{maxd}\,(a, adjacent\ of\ a\ in\ B$ indicates the maximum distance between $a \in A$ and its two adjacent in B. Whereas $\mathrm{maxd}\,(b, adjacent\ of\ b\ in\ A)$ shows the maximum distance between $b \in B$ and its two adjacent in A. Furthermore, min value is also possible for the conservative measure.

Now, similarity measure (2) is applied to Fig. 1 to calculate the similarity measure between circle and diamond.

For Fig. 1(a),

$$s(\blacklozenge,\bullet) = 1 - 1/6\,|(0.2 + 0.2 + 0.3 + 0.5 + 0.2 + 0.4) - (0.1 + 0.4 + 0.2 + 0.1 + 0.3 + 0.5)| + 0.3 + 0.5)|$$

$$= 1 - 1/6\,|1.8 - 1.6| = 0.967$$

is satisfied.

For calculation of Fig. 1(b),

$$s(\blacklozenge,\bullet) = 1 - 1/6\,|(0.1 + 0.2 + 0 + 0.2 + 0.1 + 0.3) - (0.4 + 0.3 + 0.1 + 0.6 + 0.4 + 0.2)|$$

$$= 1 - 1/6\,|0.9 - 2.0| = 0.812$$

Calculation result shows that the proposed similarity measure is possible to evaluate degree of similarity for non-overlapped distributions. By comparison, Fig. 1 (a) shows more similar than Fig. 1 (b).

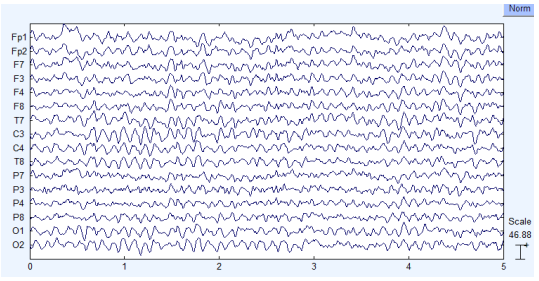
3. High-Dimensional Similarity

Ordinary high dimensional data are illustrated by the following example.

- Biomedical data such as DNA sequence or EEG data: DNA sequence and EEG signal are the latest breakthroughs in experimental molecular biology. Analysis of such data is becoming one of the major bottlenecks in the utilization of the technology.
- E-commerce: Recommendation systems and target marketing are important applications in the E-commerce area. In these applications, sets of customers/clients with similar behavior need to be identified so that we can predict customers'interest

and make proper recommendations.

- EV-station scheduling: EV station scheduling problem provides cost and electrical energy saving by optimal clustering in big data.



EEG Signals

Then the high dimensional data can be illustrated as

$$D_i = (d_{i1}, d_{i2}, d_{i3}, \dots, d_{ij}),$$

where $i = 1, \dots, n. j = 1, \dots, m$

n and m denote number of data and dimension, respectively.

3.1 Similarity Measure on high-dimension data

Direct data comparison is applicable to overlapped data with norm definition including Euclidean norm such as

$$\mathbf{d}_1(D_i, D_j) = \sum_{k=1}^{m} | d_{ik} - d_{jk}|$$

$$\mathbf{d}_2(D_i, D_j) = \sqrt{\sum_{k=1}^{m} (d_{ik} - d_{jk})^2}$$

$$\mathbf{d}_p(D_i, D_j) = (\sum_{k=1}^{m} |d_{ik} - d_{jk}|^p)^{1/p}$$

m representsthe number of characteristics or attributes. Hence, analysis and comparison with each attributes provide explicit importance of each data. For example, which element is the most decisive among from

$$D_i = (d_{i1}, d_{i2}, d_{i3}, \dots, d_{ij}), j = 1, \dots, m.$$

Such a solution can be provided from analysis of similarity measure such as

$$s(d_{i1}, f),\ s(d_{i2}, f),\ \cdots,\ s(d_{im}, f),\ \forall i = 1, \dots, n$$

By analyzing similarity measure, it is possible to provide which data is decisive. Furthermore, it also possible to decide the specific data D_i that is closely to f, that is

$$s(d_i, f) = \alpha_1 s(d_{i1}, f) + \alpha_2 s(d_{i2}, f) + \cdots \alpha_m s(d_{im}, f)$$

where α_1, α_2, $\cdots, \alpha_m$ be weighting factors of combination.

3.2 Illustrative Example

Personal data related with financial fraud was considered to analyze their relation. Personal information constitutes age A_i, gender G_i, qualification Q_i, and their related job J_i for 142 persons, that is, $i = 1, \dots, 142$.Hence, personal information can be represented by multi-dimensional information such as

$$D_i = (A_i, G_i, Q_i, J_i), \text{ where } i = 1, \dots, 142.$$

Whole informationis included in Table 1. And among 142 personal data 7 frauds were also expressed as

$$F_j = (A_{F_j}, G_{F_j}, Q_{F_j}, J_{F_j}), \text{ where } j = 1, \dots, 7.$$

Information of 7 frauds are also illustrated in Table 2.

[Table 1] Demographic information of respondents

Variables		Count	Percentage
Age	< 20 Years	0	0
	21-30 Years	74	52.1
	31-40 Years	53	37.3
	41-50 Years	10	7.0
	51-60 Years	5	3.5
	61-70 Years	0	0
	> 71 Years	0	0
Total		142	100
Gender	Male	74	52.1
	Female	68	47.9
Total		142	100
Qualification	No formal qualification	4	2.8
	GCSE / 0 LEVEL	3	2.1
	A level	12	8.5
	BSc/BA	92	64.8
	Further qualification	29	20.4
	Not known	2	1.4
Total		142	100
IT/Finance related	Neither of them	56	39.4
	IT related	9	6.3
	Finance related	61	43.0
	Both of them	16	11.3
Total		142	100
Incidents of actual financial fraud		7	4.9
Total		142	

[Table 2] Fraud information(statistics)

age		gender		highest qualification		IT related or finance related	
21-30	2	M	3	BSc BA	5	Neither	1
31-40	5	F	4	further	2	Finance	5
						Both	1

Considering the data, it is obvious that data is overlapped. Hence it is clear from similarity measures (1) satisfying similarity measure on overlapped data.

Similarity measure between age and fraud is expressed as

$$s\left(A_i, A_{F_j}\right) = 1 - d\left(\left(A_i \cap A_{F_j}\right), \left(A_i \cup A_{F_j}\right)\right) \quad (5)$$

also job,

$$s\left(J_i, J_{F_j}\right) = 1 - d\left(\left(J_i \cap J_{F_j}\right), \left(J_i \cup J_{F_j}\right)\right). \quad (6)$$

Normalized similarity calculation results are illustrated in Table 3.

[Table 3] Similarity measure between data and fraud

Similarity measure	Values
$s\left(A_i, A_{F_j}\right)$	0.544
$s\left(G_i, G_{F_j}\right)$	0.0878
$s\left(Q_i, Q_{F_j}\right)$	0.755
$s\left(J_i, J_{F_j}\right)$	0.438

With results it is obvious that

$$s\left(Q_i, Q_{F_j}\right) > s\left(A_i, A_{F_j}\right) > s\left(J_i, J_{F_j}\right) > s\left(G_i, G_{F_j}\right).$$

It means that fraud distribution is closely related with qualification. Here, we can decide the proper weighting factors $\alpha_1, \alpha_2, \cdots, \alpha_m$ that help to provide knowledge to bankers.

$$s(d_i, f) = \alpha_1 0.544 + \alpha_2 0.0878 + \alpha_3 0.755 + \alpha_4 0.438$$

Subject to maximize $s(d_i, f) \leq 1$

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$$

After we solve the sub-optimization problem, meaningful weighting factors can provided. Hence, with the mentioned weighting factors, we can decide the similarity between each person and fraud.

4. Conclusions

Similarity measure on high-dimensional data was designed. At first, overlapped and non-overlapped types of data were considered, and its similarity measures also introduced. With the conventional similarity measure, calculation of similarity on non-overlapped data was carried out. Calculation results showed similarity on each data were not consistency because conventional similarity measure was designed based on different ways of approaches,

distance measure and neighbor information.

High-dimension data was considered, comparison result ondata related fraud cases showed that conventional similarity measure is still useful for overlapped data. Similarity measure calculation provides that the priority relation between data and fraud.

REFERENCES

[1] Fisher D.H., Knowledge acquisition via incremental conceptual clustering, In Machine Learning, 1987.

[2] Jain A.K. and R.C. Dubes, Algorithms for Clustering Data. Prentice-Hall, 1988.

[3] Murtagh F., A survey of recent hierarchical clustering algorithms, In the Computer Journal, 1983.

[4] Michalski R. S. and R.E. Stepp, "Learning from observation: conceptual clustering, In Machine Learning: An artificial intelligence approaches", pp. 331-363, 1983.

[5] Friedman H.P. and J. Rubin, "On Some Invariant Criteria for Grouping Data", J. Am. Statistical Assoc., pp. 1159-1178, 1967.

[6] Fukunaga K., Introduction to Statistical Pattern Recognition, Academic Press, 1990.

[7] Advancing Discovery in Science and Engineering. Computing Community Consortium, Spring 2011.

[8] Advancing Personalized Education. Computing Community Consortium, Spring 2011.

[9] Smart Health and Wellbeing. Computing Community Consortium, Spring 2011.

[10] Liu Xuecheng, "Entropy, distance measure and similarity measure of fuzzy sets and their relations", Fuzzy Sets and Systems, pp. 305-318, 1992.

[11] Lee S.H., W. Pedrycz, and Gyoyong Sohn, "Design of Similarity and Dissimilarity Measures for Fuzzy Sets on the Basis of Distance Measure", International Journal of Fuzzy Systems, pp. 67-72, 2009.

[12] Lee S.H., K.H. Ryu, G.Y. Sohn, "Study on Entropy and Similarity Measure for Fuzzy Set", IEICE Trans. Inf. & Syst., pp. 1783-1786, 2009.

[13] Lee S.H., S. J. Kim, N. Y. Jang, "Design of Fuzzy Entropy for Non Convex Membership Function", CCIS, pp. 55-60, 2008.

[14] Cheng Y. and G. Church, "Biclustering of expression data", In Proc. of 8th international conference on intelligent system for molecular biology, 2000.

저자소개

Nipon Theera-Umpon

- 1993. Feb. : Chiang Mai University, Thailand, Electrical Engineering, (B. Eng. (Hons.))
- 1996. May: University of Southern California, U.S.A., Electrical Engineering (M.S.)
- 2000. May: University of Missouri-Columbia, U.S.A., Electrical Engineering (Ph.D.)
- 1993. Apr.~Present: Associate Professor of Dept. of Electrical Engineering, Chiang Mai University, Thailand
- E-Mail : nipon@ieee.org

Sanghyuk Lee

- 1988. Feb. : Chungbuk National University, Korea, Electrical Engineering, (B. Eng.)
- 1991. Feb: Seoul National University, Korea, Electrical Engineering, (M.S.)
- 1998. Feb: Seoul National University, Korea, Electrical Engineering, (Ph.D.)
- 2011. Aug.~Present: Professor of Dept. of Electrical Engineering, XI'anJiaotong-Liverpool University, China
- E-Mail : Sangyuk.Lee@xjtlu.edu.cn