The 6th International Conference on Construction Engineering and Project Management (ICCEPM 2015)
Oct. 11 (Sun) ~ 14 (Wed) 2015 • Paradise Hotel Busan • Busan, Korea
www.iccepm2015.org

Comparative Study on Similarity Measurement Methods in CBR Cost Estimation

Joseph Ahn[1]; Moonseo Park[2]; Hyun-Soo Lee[3]; Sung Jin Ahn[4]; Sae-Hyun Ji[5]; Sooyoung Kim[6]; Kwonsik Song[7]; and Jeong Hoon Lee[8]

***Abstract:** In order to improve the reliability of cost estimation results using CBR, there has been a continuous issue on similarity measurement to accurately compute the distance among attributes and cases to retrieve the most similar singular or plural cases. However, these existing similarity measures have limitations in taking the covariance among attributes into consideration and reflecting the effects of covariance in computation of distances among attributes. To deal with this challenging issue, this research examines the weighted Mahalanobis distance based similarity measure applied to CBR cost estimation and carries out the comparative study on the existing distance measurement methods of CBR. To validate the suggest CBR cost model, leave-one-out cross validation (LOOCV) using two different sets of simulation data are carried out. Consequently, this research is expected to provide an analysis of covariance effects in similarity measurement and a basis for further research on the fundamentals of case retrieval.*

***Keywords:** case-based reasoning, cost estimation, similarity measurement, Mahalanobis distance*

I. INTRODUCTION

To achieve a success of a construction project, construction cost, one of key decision making elements, should be properly planned and managed throughout the construction life cycle. The critical issue is how entire or partial construction cost can be accurately decided in early design stages as it can improve the stability of cost planning and reduce the risk of cost overruns (Schuette and Liska 1994; Kim 2005).

CBR method, traced its roots to the area of psychology and theories about how human memory works, have been progressively more applied to cost estimating (Yau and Yang 1998; Karshenas and Tse 2002; An et al. 2007; Chou 2009; Ji et al. 2011). This method utilizes the experience of past cases to work out new ones (Schirmer 2000) and can support various decision-makings (Pal and Shiu 2004).

Many research have been carried out to examine the specific issues among them. It takes notice that a large number of the studies are focused on retrieval to improve accuracy of the results in early phase. Because it can enhance of reliability of the estimation process in reuse, revise, and retain phases.

However, these existing similarity measures have limitations in taking the covariance among attributes into consideration and reflecting the effects of covariance in computation of distances among attributes. To deal with this challenging issue, this research examines the weighted Mahalanobis distance based similarity measure applied to CBR cost estimation and carries out the comparative study on the existing distance measurement methods of CBR. To validate the suggest CBR cost model, leave-one-out cross validation (LOOCV) using two different sets of simulation data are carried out. Consequently, this research is expected to provide an analysis of covariance effects in similarity measurement and a basis for further research on the fundamentals of case retrieval.

II. LITERATURE REVIEW: MAHALANOBIS DISTANCE

The Mahalanobis distance is a distance measure between a point P and a distribution D, introduced by P. C. Mahalanobis in 1936 and is widely used in cluster analysis and classification techniques (McLachlan 1992). This distance measure takes into account of correlations between attributes by which different patterns can be identified and analyzed as it is computed using the inverse of variance-covariance matrix of the data set (Maesschalck et al. 2000). It is considered to be an appropriate measure as it eliminates the unnecessary influence of covariance between attributes (Mahalanobis 1936; Du and Bormann 2014). Also, it is useful to determine similarity of an unknown sample set to a known one.

The Mahalanobis distance of a multivariate vector $x=(x\_1,x\_2,x\_3,\ldots,x\_N)^T$ from a group values with mean $\mu=(\mu\_1,\mu\_2,\mu\_3,\ldots,\mu\_N)^T$ and covariance matrix S is defined as:

$$D_M(x) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T S^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \qquad \text{(Eq. 1)}$$

A crucial difference from Euclidean distance which is a widely adopted distance measure is that it considers the correlations of the data set. Mahalanobis distance is affected by both variance and correlation. If the covariance

1 Ph.D. Student, Dept. of Architecture, Seoul National Univ., Gwanak-ro 1, Gwanak-gu, Seoul, 151-742, Korea, E-mail: josephahn@snu.ac.kr
2 Prof., Dept. of Architecture, Seoul National Univ., Gwanak-ro 1, Gwanak-gu, Seoul, 151-742, Korea, E-mail: mspark@snu.ac.kr
3 Prof., Dept. of Architecture, Seoul National Univ., Gwanak-ro 1, Gwanak-gu, Seoul, 151-742, Korea, E-mail: hyunslee@snu.ac.kr
4 Prof., Dept. of Information Statistics, Gyeongsang National Univ., 501, Jinju-daero, Jinju, Gyeongnam, 660-701, Korea, E-mail: ahnsj@gnu.ac.kr
5 Research Assistant, Ph.D., Defense Acquisition Program Administration, Seoul 140-701, Korea, E-mail: oldclock@snu.ac.kr (corresponding author)
6 Ph.D. Student, Dept. of Architecture, Seoul National Univ., Gwanak-ro 1, Gwanak-gu, Seoul, 151-742, Korea, E-mail: wing195@snu.ac.kr
7 Ph.D. Student, Dept. of Architecture, Seoul National Univ., Gwanak-ro 1, Gwanak-gu, Seoul, 151-742, Korea, E-mail: woihj@snu.ac.kr
8 Ph.D. Student, Dept. of Architecture, Seoul National Univ., Gwanak-ro 1, Gwanak-gu, Seoul, 151-742, Korea,E-mail: di5555@snu.ac.kr

matrix is the identity matrix then it is the same as Euclidean distance. If covariance matrix is diagonal, then it is called normalized Euclidean distance. However, calculating the variance-covariance matrix using over a large number of attributes still remains as limitation since they can contain much redundant or correlated information (Maesschalck et al. 2000). Also, the Mahalanobis distance concept can be used where the data set is multivariate normally distributed; in other words, each attribute should be normally distributed (Johnson and Wichern 1998; Robinson et al. 2005).

Ⅲ. CBR COST MODEL AND EXPERIMENTAL DESIGN

As far as here, we have conducted preliminary studies on various similarity measures and elaborated their common limitations and consequential effects that might be resulted from not accounting for the covariance among the attributes. To manage this challenging issue, this research have proposed the weighted Mahalanobis distance to accurately measure the similarities among attributes where covariance exists. In order to examine how the suggested weighted Mahalanobis distance based similarity measure is different from other previously adopted measures in terms of accuracy, stability, and propriety, an experiment for comparative analysis is designed as illustrated in Figure 1.

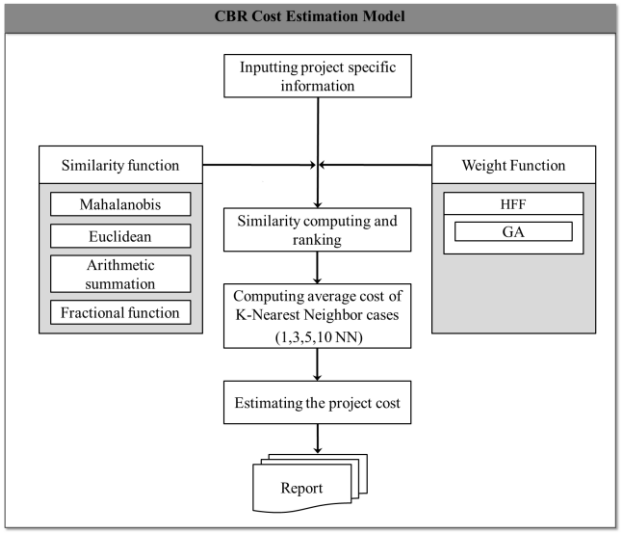

Figure 1. CBR Cost Estimation Model

Simulation Data Condition 1 is Parameters of multi-variate normal distribution: mean vector: μ = [132.1 259.5 77.5 346.6 542.4]T Also, Simulation Data Condition 2 is Parameters of multi-variate normal distribution: mean vector: $\mu = [132.1\ 259.5\ 77.5\ 346.6\ 542.4]^T$

Ⅳ. CONCLUSIONS

Resultantly, there are high possibilities of an inaccurate similarity measure among attributes which might mislead to the results of case retrievals. To deal with this challenging issue, this research proposed the weighted Mahalanobis distance based similarity measure applied to CBR cost estimation and carried out the comparative study on each distance measurement. Ultimately, this research demonstrated that the Mahalanobis distance measurement can be an effective CBR based cost estimating during the initial project stages in practical and theory. However, this research applied only to public apartment buildings in Korea, and further research should validate the use of the Mahalanobis measurement method with other building types for greater generalization.

REFERENCES

[1] S. D. Schuette, R. W. Liska, “Building Construction Estimating”, McGraw-Hill, New York, 1994.
[2] S. K. Pal, S. C. K. Shiu, “Foundations of Soft Case-Based Reasoning”, Wiley Interscience, 2004.
[3] G. J. McLachlan, “Discriminant Analysis and Statistical Pattern Recognition”, Wiley Interscience, 1992.
[4] R. A. Johnson, D. W. Wichern, “Applied Multivariate Statistical Analysis”, Prentice-Hall, Englewood Cliffs, N.J, 1998