Pertanika J. Sci. & Technol. 25 (2): 619 - 630 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Review of Context-Based Similarity Measure for Categorical Data

Nurul Adzlyana, M. S.*, Rosma, M. D. and Nurazzah, A. R.

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 UiTM, Shah Alam, Selangor, Malaysia

ABSTRACT

Data mining processes such as clustering, classification, regression and outlier detection are developed based on similarity between two objects. Data mining processes of categorical data is found to be most challenging. Earlier similarity measures are context-free. In recent years, researchers have come up with context-sensitive similarity measure based on the relationships of objects. This paper provides an in-depth review of context-based similarity measures. Descriptions of algorithm for four context-based similarity measure, namely Association-based similarity measure, DILCA, CBDL and the hybrid context-based similarity measure, are described. Advantages and limitations of each context-based similarity measure are identified and explained. Context-based similarity measure is highly recommended for data-mining tasks for categorical data. The findings of this paper will help data miners in choosing appropriate similarity measures to achieve more accurate classification or clustering results.

Keywords: Categorical data, context-based, data mining, similarity measure

INTRODUCTION

Similarity measure is the measure of how much alike two data objects are. Similarity measure in data mining is usually described as a distance with dimensions representing features of the objects. A small distance means a high degree of similarity and vice versa. Similarity is very subjective and is highly dependent on the application domain (Yong, 2010).

Similarity between two objects plays an important role in data mining tasks such as clustering, classifying, regressing, or finding distance for outlier detection of various types of data (Desai et al., 2011) involving distance computations. The distance of similarity for integer-type data and ratio-scaled data are well defined and understood.

However, devising similarity or distance metrics for classification and clustering of categorical data is found to be more challenging (Alamuri et al., 2014). The

Article history:
Received: 27 May 2016
Accepted: 14 November 2016

E-mail addresses:
nurul_adzlyana@yahoo.com (Nurul Adzlyana, M. S.),
rosma@tmsk.uitm.edu.my (Rosma, M. D.),
nurazzah@tmsk.uitm.edu.my (Nurazzah, A. R.)
*Corresponding Author

ISSN: 0128-7680

usual similarity measures for categorical data are binary methods, where each bit indicates the presence or absence of a possible attribute value. The similarity between two objects is determined by the similarity between two corresponding binary vectors (Khorshidpour et al., 2010). Nevertheless, the alteration of data objects into binary vectors is the main problem, as the binary vectors will calculate the similarity between two values to be either 0 or 1 and in the process may eliminate important information about the data.

Earlier similarity measures are context-free. But recently researchers have come up with context-sensitive similarity measures. Hence, similarity measures can be divided into two main categories based on the way they utilise the context of the given attributes. Thus, similarity measures can be context-free or context-sensitive. Context-free similarity measure is used when the distance between two objects in the data is taken as a function of the objects only and does not depend on the relationship between those objects to other data objects. On the other hand, context-sensitive similarity measure is used when the similarity between two data objects depends not only on the two objects alone, but also on the relationship between the objects and the other data objects (Alamuri et al., 2014).

In more recent research, hybrid similarity measures have been introduced which combine two important elements. The first element is the context selection process followed by distance computation (Alamuri et al., 2014). Context-selection is observing the meta-attributes connected with the current attributes of the objects called context attributes. In order to determine the context of each attribute, a data driven method is employed and it is application specific while the distance computation is for the pair of values of an attribute based on context selection. Alamuri has suggested a hybrid similarity measure that combines learning algorithm for context selection and distance computation based on the learned context.

This paper will review three commonly used context-based similarity measures. The advantages and limitations of the methods are described for comparison purposes.

CONTEXT-BASED SIMILARITY MEASURE

There are four techniques in context-based similarity measure for categorical data, namely the Association-Based Similarity Measure, DILCA, CBDL and the hybrid context-based similarity measure. These techniques are described below.

Association-Based Similarity Measure

An association-based similarity measure was proposed by (Le & Ho, 2005). A novel indirect method to measure the dissimilarity for categorical data was introduced. The dissimilarity between two values of an attribute is indirectly estimated by using relations between other related attributes.

The efficiency of the proposed method is investigated in terms of theoretical proofs and the experiments with real data showed that attributes are typically correlated. However, this method is found to be unsuitable for data sets with independent attributes (Le & Ho, 2005). The

Association-Based Similarity Measure comprises two steps which are finding the dissimilarity between two values of attribute followed by finding the dissimilarity between two data objects.

The algorithm of the method is as below:

Step 1: The dissimilarity between two values of attribute. The dissimilarity between two values v_i and v'_i of attribute A_i, denoted by $\emptyset_{A_i}(v_i, v'_i)$, is the sum of dissimilarities between conditional probability distributions of other attributes given that attribute A_i holds values v_i and v'_i:

$$\emptyset_{A_i}(v_i, v'_i) = \sum_{j, j \neq i} \psi\left(cpd\left(A_j | A_i = v_i\right), cpd\left(A_j | A_i = v'_i\right)\right) \quad (1)$$

where $\psi(.,.)$ is a dissimilarity function for two probability distributions.

The dissimilarity between two values v_i and v'_i of attributes is directly proportional to dissimilarities between the conditional probability distributions of other attributes. Thus, the smaller the dissimilarities between these conditional probability distributions, the smaller the dissimilarity between v_i and v'_i.

Le and Ho (2005) used the popular dissimilarity measure, which is the KullbackLeibler method, (Kullback & Leibler, 1951) that is given by:

$$KL(P, P') \sum_x \left(p(x) lg \frac{p(x)}{p'(x)} + p'(x) lg \frac{p'(x)}{p(x)}\right) \quad (2)$$

where is a logarithm of base 2.

Step 2: The dissimilarity between two data objects. The dissimilarity between two data objects x and y, denoted by $\emptyset(x, y)$, is the sum of dissimilarities of their attribute value pairs:

$$KL(P, P') \sum_x \left(p(x) lg \frac{p(x)}{p'(x)} + p'(x) lg \frac{p'(x)}{p(x)}\right) \quad (3)$$

If the dissimilarities of attribute value pairs of x and y is smaller, then the dissimilarities between x and y is also smaller.

Distance Learning for Categorical Attributes (DILCA)

Ienco et al. (2012) proposed a context-based similarity measure called Distance Learning for Categorical Attributes (DILCA) to compute the distance between any pair of values of a specific categorical attribute. The method consists of two steps: context selection and distance computation. The context selection step is the process of selecting the relevant subset of the whole attributes set while the distance computation is the process of computing the distance between pair of values of the same attribute using the context defined in the context selection.

Step 1: Context selection. The aim in this step is to select a subset of relevant and not overlapped features. Ienco et al. (2012) proposed several approaches for measuring the correlation between two variables. One of them is the Symmetric Uncertainty. This context selection is a correlation based measure inspired by information theory. Symmetric Uncertainty is derived from entropy as it is a measure of the uncertainty of a random variable. The entropy of a random variable is defined as:

$$H(X) = -\sum_{i} P(x_i)\, log_2\big(P(x_i)\big) \tag{4}$$

where $P(x_i)$ is the probability of the value x_i of the X. The entropy of X after having observed the values of another variable Y is defined as:

$$H(X|Y) = -\sum_{j} P\left(y_j\right) \sum_{i} P(x_i|y_i)\, log_2\big(P(x_i|y_i)\big) \tag{5}$$

where $P(x_i|y_i)$ is the probability that $X = x_i$ after observing that $Y = y_i$. The information about X provided by Y is given by the information gain, which is defined as follows:

$$IG(X|Y) = H(X) - H(X|Y) \tag{6}$$

When $IG(X|Y) > IG(Z|Y)$ then the feature X is more correlated to Y than Z. Moreover, the information gain is symmetrical for two random variables X and Y. The symmetrical uncertainty is then defined as:

$$SU(X,Y) = 2 \cdot \frac{IG(X|Y)}{H(X) + H(Y)} \tag{7}$$

This measure varies between 0 and 1 where 1 indicates that knowledge of the value of either X or Y and completely predicts the value of the other variable while 0 indicates that X and Y are independent. The advantage of symmetrical uncertainty is that the Information Gain that it measures is not biased by the number of values of an attribute.

Step 2: Distance Computation. The goal of this step is to compute the distance between x_i and x_j where $x_i \in X, x_j \in X$, using this given formulation:

$$d\big(x_i, x_j\big) = \sqrt{\sum_{Y \in context(X)} \sum_{y_k \in Y} \Big(P(x_i|y_k) - P\big(x_j|y_k\big)\Big)^2} \tag{8}$$

The conditional probability for both values x_i and x_j in each context attribute Y is given by the values $y_k \in Y$. Then, the Euclidean distance is applied.

Context-Based Distance Learning (CBDL)

In 2011, Khorshidpour et al. (2010) proposed a method to measure the dissimilarity of categorical data, CBDL. This method consists of two steps. In the first step, a relevant subset of the whole attributes is selected. Then, the dissimilarity between pair of the values of the same attribute is computed using the context defined in the first step. The two steps are described as below:

Step 1: Context Selection. Supervised feature selection is employed in this step. The goal is to select a subset of correlated features based on the given one. The outcome of feature selection is a subset of input variables by eliminating features with given class attribute. Feature selection process can improve the comprehensibility of the resulting classifier models. It often builds a model that generalises better to unseen points. The steps are defined as below:

Entropy can be used to derive Mutual Information. The entropy of a random variable A_i is defined as:

$$H(A_i) = -\sum_k p(a_k^i) log_2 \left(p(a_k^i)\right)$$

$$H(A_i|A_j) = -\sum_k p\left(a_k^i\right) \sum_l p(a_l^i|a_k^j)\, log_2 p(a_l^i|a_k^j) \tag{9}$$

where $p(a_k^i)$ is the probability of the value a_k of A_i and $p(a_k^i|a_k^j)$ is the probability that $A_i = a_l$ after observing $A_j = a_k$. The mutual information is related to the conditional entropy through

$$MI(A_i|A_j) = H(A_i) - H(A_i|A_j) \tag{10}$$

The redundancy, R, is a more useful and symmetric scaled information measure, where:

$$R = \frac{MI(A_i; A_j)}{H(A_i) + H(A_j)} \tag{11}$$

The symmetric uncertainty is another alternative of the symmetrical measure given by:

$$DS(A_i; A_j) = SU(A_i; A_j) = 2R = 2\frac{MI(A_i; A_j)}{H(A_i) + H(A_j)} \tag{12}$$

This measure varies between 0 and 1 (1 indicates that knowledge of the value of either A_i or A_j completely predicts the value of the other variable while 0 indicates that A_i and A_j are independent).

Finally, the relevance score for each feature A_i, $RS(A_i)$, is computed as the average dependence score between A_i and the rest of the feature:

$$RS(A_i) = \frac{1}{m-1} \sum_{A_j \in F \setminus A_i} DS(A_i; A_j) \tag{13}$$

where m denotes number of features and F is feature set. The lower the value of $RS(A_i)$, the lesser relevant of A_i. The following inequality is used to determine the context of an attribute A_i:

$$context(A_i) = \{A_j \in F \setminus A_i$$

$$DS(A_i; A_j) \geq 0.5RS(A_i)\}$$

Step 2: Distance Computation. The sum of distance of their attribute value pairs reflects distance between two data objects X and Y denoted by $D(X,Y)$, is:

$$D(X,Y) = \sum_{i=1}^{m} d_{A_i}(x_i, y_i) \tag{14}$$

where $D(x_i, y_i)$ denoted between two values x_i and y_i of attribute A_i:

$$d_{A_i}(x_i, y_i) = \sum_{A_j \in context(A_i)} KL\left(cpd(A_j|A_i = x_i)\right), \left(cpd(A_j|A_i = x_i)\right) \tag{15}$$

Dissimilarity between two values x_i and y_i of attribute A_i is directly proportional to dissimilarities of context's attributes given these values.

KullbackLeibler divergence method is used to compute dissimilarity between probability distributions.

$$KL(p,q) = \sum_i \left(p_i log \frac{p_i}{q_i} + q_i log \frac{q_i}{p_i}\right) \tag{16}$$

For each context attribute A_j, the conditional probability is computed for both values x_i and y_i given the values $v_j \in A_j$. Then, KullbackLeibler divergence method is applied. The dissimilarity between x_i and y_i equals to 0 if and only if the conditional probability distributions of other attributes when A_i holds values x_i and y_i are identical since KullbackLeibler dissimilarity between two probability distributions is non-negative, and equal to 0 if and only if the distributions are identical.

HYBRID SIMILARITY MEASURE

Alamuri et al. (2014) introduced a two-step hybrid similarity measure using context selection and distance computation. Context selection considers the meta attributes related to the current attributes called "context attributes". The determination of the context of every attribute is data-driven and data-specific. Distance computation is made for the pair of values of an attribute. The context selection describes the steps adopted.

Alamuri et al. (2014) proposed a hybrid method based on entropy (Cover & Thomas, 1991) and mutual information (Shannon, 1948). This method is described below:

Let D be the data set with feature set $F = \{A_1, A_2, \dots, A_M\}$ of n data points, where each data object is of m attributes which are categorical. The entropy of a random variable A_i is defined as:

$$H(A_i) = -\sum_{k \in A_i} p\left(a_k^i\right) log_2 p\left(a_k^i\right) \tag{17}$$

where $p\left(a_k^i\right)$ is the probability of value a_k of attribute A_i. The entropy of random variable can be conditioned on other variables. The conditional entropy A_i given A_j is:

$$H\left(A_i|A_j\right) = -\sum_{k \in A_j} p\left(a_k^j\right) \sum_{l \in A_i} p\left(a_l^i|a_k^j\right) log_2 p\left(a_l^i|a_k^j\right) \tag{18}$$

where $p\left(a_l^i|a_k^j\right)$is the probability that $A_i = a_k$. This means the amount of uncertainty present in A_i after observing the variable A_j. The amount of information shared between A_i and A_j (mutual information) is defined as:

$$I\left(A_i; A_j\right) = H(A_i) - H\left(A_i|A_j\right) \tag{19}$$

This is the difference between two entropies which can be interpreted as the amount of uncertainty in A_i which is removed by knowing A_j. After observing another variable A_z, the mutual information can also be conditioned as the amount of information still shared between A_i and A_j. The conditional mutual information is:

$$I\left(A_i; A_j|A_z\right) = H(A_i|A_z) - H\left(A_i|A_j A_z\right)$$

$$I\left(A_i; A_j|A_z\right) = \sum_{m \in A_z} p(a_m^z) \sum_{k \in A_i} \sum_{l \in A_j} p\left(a_k^i a_l^j|a_m^z\right) log \frac{p\left(a_k^i a_l^j|a_m^z\right)}{p\left(a_k^i|a_m^z\right) p\left(a_l^j|a_m^z\right)} \tag{20}$$

KullbackLeibler divergence method is applied to calculate the distance between pair of values of an attribute. The formula is given by:

$$d_{A_i}(x_i, y_i) = \sum_{A_i \in context(A_i)} \sum_{v_j \in A_j} \left(p(v_j|x_i) log \frac{p(v_j|x_i)}{p(v_j|y_i)} + p(v_j|y_i) log \frac{p(v_j|y_i)}{p(v_j|x_i)} \right) \quad (21)$$

DISCUSSION & CONCLUSION

A review of several context-based similarity measures was conducted. First, components of four context-based similarity measures, including one hybrid similarity measure, were identified. The context-based similarity measures include Association Based Similarity Measure, DILCA, CBDL and the Hybrid Similarity Measure proposed by (Alamuri et al., 2014). The components for each context-based similarity measures and context-free similarity measure may be different in terms of their ability to compute similarity of attributes, similarity of objects, and distance computation. Furthermore, the concepts used to develop the similarity measures are found to be different between one similarity measure and the other. Specific components are shown in Table 1 below. Context-free similarity is based on a very simple concept which uses mainly distance computation using overlap measure. This measure does not take into consideration the relationship between data features. On the other hand, context-based similarity measures compute similarity between attributes and/or between objects. Above all, hybrid similarity measure is found to be the best since it measures similarity by considering all the three components.

Table 1
Components of the context-free in comparison to context-based similarity measure

	Association-based Similarity Measure	**DILCA**	**CBDL**	**Hybrid Similarity Measure**	**Context-Free Similarity Measure**
Similarity of attributes	Kullback Leibler	Entropy & Mutual Information	Entropy & Mutual Information	Entropy & Mutual Information	
Similarity of objects	Sum of dissimilarities of attributes value pair			Entropy & Mutual Information	
Distance computation		Euclidean distance	Kullback Leibler	Kullback Leibler	Overlap Measure

Second, a thorough comparison of four commonly used context-based similarity measures, including one hybrid similarity measure, was done. The context-based similarity measures include Association Based Similarity Measure, DILCA, CBDL and the Hybrid Similarity Measure proposed by (Alamuri et al., 2014). Four characteristics are discussed, namely algorithm, concepts, strengths and limitations of each method . A description of each characteristic for respective context-based similarity measure is provided in Table 2.

Table 2
Characteristic descriptions of four context-based similarity measures

	Association-based Dissimilarity Measure	**DILCA**	**CBDL**	**Hybrid Context-Based Similarity Measure**
Algorithm	Dissimilarities between two values of an attribute is found as the sum of the dissimilarities between conditional probability distributions of other related attributes. Finding the dissimilarity between two data objects.	Feature Selection is applied to select the relevant subset of the whole attributes in respect to the given one. Euclidean distance is applied to compute the distance between values of the same attribute.	Context Extraction Component is used to extract the relevant subset of feature set for a given attribute. Distance Learning Component is applied to learn distance between each pair of values of an attribute based on extracted context.	Context selection process looks at meta attributes associated with the current attributes. Distance computation is done for the pair of values of an attribute using the context defined in context selection.
Strengths	Experiments show that attributes are typically correlated. Lead to the idea of replacing each of the attribute groups by one or a few attributes that can have more discriminating power. Boost the accuracy of neural network when applied to real data.	Good clustering result is obtained when applied into clustering algorithm. A new methodology to compute a matrix of distances between any pair of values of a specific categorical attribute X The method is independent from the specific clustering algorithm. DILCA is considered a a simple way to compute distances for categorical attribute. Attributes that introduce noise are ignored in the value distance computation step.	No sign of degradation when the number of irrelevant attributes increased. Accuracy was significantly higher when compared with the other popular similarity measure called Value Difference Metric (VDM). Improve the classification accuracy by reducing the effects of irrelevant attributes. Can be applied to any data mining task that involves categorical data.	Context selection is done by taking into consideration the meta attributes associated with the current attributes called context attributes.

Table 2 (*continue*)

	Association-based Dissimilarity Measure	**DILCA**	**CBDL**	**Hybrid Context-Based Similarity Measure**
Limitation	Cannot be applied to databases whose attributes are absolutely independent.	When the size of the dataset is small with respect to the number of attributes, it is expected that the results are biased by the weak representativeness of the samples. In some cases, performances are low for any clustering algorithm. The partitions determined by the class labels are not supported by data.		The context selection algorithm has the tendency to select the complete set of attributes as relevant context for the given attribute.

In summary, all the three context-based similarity measures reviewed above provide high accuracy when applied to clustering tasks. The Association based similarity measure can boost the accuracy of NN in clustering tasks. However, it cannot be applied when attributes are absolutely independent from one another. The strength of DILCA lies in the fact that the method uses simple distance computation and at the same time ignores the noise that exists in the attributes. However, it may produce biased results when dealing with small data set. The CBDL method showed no sign of degradation even as the number of irrelevant attributes increased. Alamuri's hybrid similarity measure has not been fully experimented. Its strength is based on its ability to take into consideration the meta attributes associated with the current attributes called context attributes. However, its context selection algorithm has the tendency to select the complete set of attributes as relevant context for the given attribute. In conclusion, context-based similarity measure is found to be highly recommended for data-mining tasks for categorical data. The findings of this paper will help data miners to choose the most appropriate similarity measure in achieving a more accurate classification or clustering result.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Teknologi MARA for its financial assistance.

REFERENCES

Alamuri, M., Surampudi, B. R., & Negi, A. (2014, September). A survey of distance/similarity measures for categorical data. In *Conference on Neural Networks (IJCNN), 2014 International Joint* (pp. 1907-1914). IEEE.

Cover T. M. & Thomas J. A. (1991). *Elements of Information Theory*. United States of America, USA: John Wiley & Sons.

Desai, A., Singh, H., & Pudi, V. (2011, May). Disc: Data-intensive similarity measure for categorical data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 469-481). Springer Berlin Heidelberg.

Ienco D., Pensa. R. G. & Meo R. (2012). From Context to Distance: Learning Dissimilarity for Categorical Data Clustering. *ACM Transaction on Knowledge Discovery from Data, 6*(1), 1-25.

Khorshidpour, Z., Hashemi, S., & Hamzeh, A. (2010, October). Distance learning for categorical attribute based on context information. In *2010 2nd International Conference on Software Technology and Engineering (ICSTE),* (Vol. 2, pp. V2-296). IEEE.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79-86.

Le, S. Q., & Ho, T. B. (2005). An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters, 26*(16), 2549-2557.

Shannon C. E. (1948). A Mathematical Theory of Communication. *Bell Systems Technical Journal, 27*(3), 379-423.

Yong J. B. (2010). *Data Mining Portfolio: Similarity Measure.* Retrieved from: *humanoriented.com/classes/2010/fall/csci568/portfolio_exports/bhoenes/similarity.htm.*