EAI Endorsed Transactions
on Context-Aware Systems and Applications Research Article ICST.ORG

Clustering the objective interestingness measures based on tendency of variation in statistical implications

Nghia Quoc Phan[1], Vinh Cong Phan[2], Hung Huu Huynh[3], Hiep Xuan Huynh[4]

[1] Travinh University, Nguyen Thien Thanh Street, Travinh City, Vietnam, nghiatvnt@gmail.com
[2] Nguyen Tat Thanh University, Nguyen Tat Thanh St., District 4, Ho chi Minh City, Vietnam, pcvinh@ntt.edu.vn
[3] Danang University of Science and Technology, Nguyen Luong Bang St, Danang City, Vietnam, hhhung@dut.udn.vn
[4] Cantho University, 3/2 Street, Ninh Kieu District, Cantho City, Vietnam, hxhiep@ctu.edu.vn

Abstract

In recent years, the research cluster of objective interestingness measures has rapidly developed in order to assist users to choose the appropriate measure for their application. Researchers in this field mainly focus on three main directions: clustering based on the properties of the measures, clustering based on the behavior of measures and clustering tendency of variation in statistical implications. In this paper we propose a new approach to cluster the objective interestingness measures based on tendency of variation in statistical implications. In this proposal, we built the statistical implication data of 31 objective interestingness measures based on the examination of the partial derivatives on four parameters. From this data, two distance matrices of interestingness measures are established based on Euclidean and Manhattan distance. The similarity trees are built based on distance matrix that gave results of 31 measures clustering with two different clustering thresholds.

Keywords: objective interestingness measures, tendency of variation in statistical implications, distance matrix, Similarity tree, Clustering objective interestingness measures.

Received on 28 December, 2015, accepted on 29 March, 2016, published on 02 May, 2016

doi: 10.4108/eai.2-5-2016.151212

1. Introduction

The objective interestingness measures play an important role in evaluating the quality of knowledge in the form of association rules, especially in the post-processing stage of the process of mining knowledge from databases. Researchers in this field are mainly concentrated in two main directions: (1) proposing new measures; (2) studying the properties, behaviors and trends of the variation of the measures in order to rank, cluster and classify them. This study aims to assist users to select appropriate measures for their particular application. Clustering of objective interestingness measures is one of the research areas that many researchers are concerning [9][21]. Clustering measures is the process of searching and discovery of clustering measures to match each application area [21]. Currently, there are many techniques that can be applied in the clustering measures: clustering based partitioning, clustering based on hierarchical [17][18] and clustering based on density. In general, these techniques are directed at two main goals: the first goal is to find the most appropriate measure for specific applications [21] and the second goal is to consider the relationship of a particular measure with the remaining measure in a set of the study measures [10]. In particular, the technique based on hierarchical clustering [19] mainly focused on the second objective.

The selection of an appropriate measure for applications is what many researchers and users have always wished. However, the list of the objective interestingness measures proposed is increasing [11] and has surpassed 100, the selection becomes a significant challenge for them. The research results of the variation of the objective interestingness measures based on partial derivatives have opened up some new researches such as classification measure based on the tendency of the variation of measures [15], the consideration of the variability of the measures with statistical implication parameters and the relationship of interdependence between the statistical implication parameters in every measures [8][16]. The list of the objective interestingness measures can be reduced based on partial derivative results to support users whether to choose better measure? This paper proposes a new approach using the hierarchical structure of similarity tree [3][4][5][7] to cluster the objective interestingness measures which

*Corresponding author. Email: nghiatvnt@gmail.com agreed asymmetrical properties. In this approach, we use the results of tendency of variation in statistical implications [15] based on partial derivatives of the calculated function of measures on each parameter to build distance matrix [13] of the measures. Clustering results of the measures are demonstrated through the structure of similarity tree. Each cluster is a group of measures that has proximity or similarity to each other. This is criteria for the use of researchers and users in order to choose an appropriate measure for their applications in a better way.

This paper is organized into six sections. Section 1 introduces the measures clustering technique and raises the research issue. Section 2 presents tendency of variation in statistical implications and builds data based on values of partial derivatives of measures. Section 3 outlines the method of measuring distance and distance matrix of the measures. Section 4 describes algorithm for clustering the measures and similarity tree of the measures. Section 5 presents experiments. The final section summarizes the important results achieved by the article.

2. Statistical implications

Statistical implications [8] study the implication relationship between variable data or attribute data that allows the detection of rules a → b asymmetrical in the form "if a, then nearly as b" or "consider to what extent that b will satisfy the implications of a".

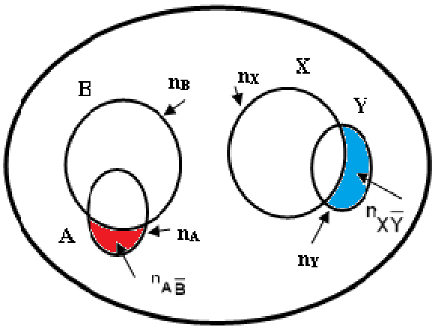


Figure 1: The model represents a statistical implication rule a → b

2.1. Tendency of variation in statistical implications

The tendency of variation in statistical implications is a researching direction to examine the stability of the implication intensity to observe small variations of measures in the surrounding space of parameters n, n_A, n_B and $n_{A\bar{B}}$ [8]. Identifying trends of variation in statistical implications of the measures shows some possibilities for application in the study of the interestingness measures and practical application: study the variability of the measure, dependent relationship between the variable parameters $n, n_A, n_B, n_{A\bar{B}}$ [8], classification of the objective interestingness measures [15]. To clarify the tendency of variation in statistical implications, we examine the Implication index measures [[8]] under 4 parameters $n, n_A, n_B, n_{A\bar{B}}$ with formula defined:

$$q(a,\bar{b}) = \frac{n_{A\bar{B}} - \frac{n_A(\mathrm{n}-\mathrm{n}_B)}{n}}{\sqrt{\frac{n_A(\mathrm{n}-\mathrm{n}_B)}{n}}}$$

To observe the variation of q from the variability of the parameters $n, n_A, n_B, n_{A\bar{B}}$, Let us consider the parameters $n, n_A, n_B, n_{A\bar{B}}$ as real numbers satisfying the following inequalities:

$$n_{A\bar{B}} \leq \inf(n_A, n_B) \text{ and } \sup(n_A, n_B) \leq n$$

In this case, q can be considered as a continuous differentiable function:

$$dq = \frac{\partial q}{\partial n}dn + \frac{\partial q}{\partial n_A}dn_A + \frac{\partial q}{\partial n_B}dn_B + \frac{\partial q}{\partial n_{A\bar{B}}}dn_{A\bar{B}}$$

The function $q(n, n_A, n_B, n_{A\bar{B}})$ is a scalar function variations on the surface represent 4 parameters. To observe the variation of q according to the parameters, we calculated the partial derivative for each parameter. In fact, this variation is estimated rising of the function q with variation according to the variation of Δq corresponding components $\Delta n, \Delta n_A$, Δn_B and $\Delta n_{A\bar{B}}$. Therefore, we have the formula:

$$\Delta\,\mathrm{q} = \frac{\partial q}{\partial n}\Delta n + \frac{\partial q}{\partial n_A}\Delta n_A + \frac{\partial q}{\partial n_B}\Delta n_B + \frac{\partial q}{\partial n_{A\bar{B}}}\Delta n_{A\bar{B}}$$

Let us take the partial derivatives of q under n we have the following formula:

$$\frac{\partial \mathrm{q}}{\partial \mathrm{n}} = \frac{1}{2\sqrt{\mathrm{n}}}\left(\mathrm{n}_{A\bar{B}} + \frac{\mathrm{n}_A(\mathrm{n}-\mathrm{n}_B)}{\mathrm{n}}\right)$$

According to the formula, if the parameters $n_A, n_B, n_{A\bar{B}}$ is constant, the implication index decreases under parameter n. For example, with $n = 100, n_A = 20, n_B = 40, n_{A\bar{B}} = 4$ then q = -2.309401 and $\frac{\partial \mathrm{q}}{\partial \mathrm{n}} = 0.8$, when n increases from 100 to 120 and the rest parameters do not change the value of q and $\frac{\partial \mathrm{q}}{\partial \mathrm{n}}$ decrease (q = -2.556039 and $\frac{\partial \mathrm{q}}{\partial \mathrm{n}}$ =0.7911548).

Let us take the partial derivatives of q under n_A we have the following formula:

$$\frac{\partial q}{\partial n_A} = -\frac{1 \quad n_{A\bar{B}}}{2\sqrt{\frac{n-n_B}{n}}}\left(\frac{n}{n_A}\right)^{\frac{3}{2}} - \frac{1}{2}\sqrt{\frac{n-n_B}{n_A}}$$

From the formula, if the parameter n_A changes from 0 to n_B, the Implication index always reduces to n_A and reaches the lowest value when $n_A = n_B$. For example, with $n = 100, n_A = 20, n_B = 40, n_{A\bar{B}} = 4$ then q = -2.309401 and $\frac{\partial \mathrm{q}}{\partial n_A} = -29.73354$, when n_A increases from 20 to 30 and the rest parameters do not change the value of q decrease and $\frac{\partial \mathrm{q}}{\partial n_A}$ increase (q = -3.299832 and $\frac{\partial \mathrm{q}}{\partial n_A} = -16.42059$).

Let us take the partial derivatives of q under n_B we have the following formula:

$$\frac{\partial q}{\partial n_B} = \frac{1}{2}n_{A\bar{B}}\left(\frac{n_A}{n}\right)^{-\frac{1}{2}}(n-n_B)^{-\frac{3}{2}} + \frac{1}{2}\left(\frac{n_A}{n}\right)^{\frac{1}{2}}(n-n_B)^{-\frac{1}{2}}$$

Let us take the partial derivatives of q under $n_{A\bar{B}}$ we have the following formula:

$$\frac{\partial q}{\partial n_{A\bar{B}}} = \frac{1}{\sqrt{\frac{n_A(n-n_B)}{n}}}$$

From two formulas above, if Δn_B and $\Delta n_{A\bar{B}}$ increase, the Implication index increase. For example, with $n = 100, n_A = 20, n_B = 40, n_{A\bar{B}} = 4$ then q = -2.309401, $\frac{\partial \text{q}}{\partial n_B} =$ 0.03849002 and $\frac{\partial \text{q}}{\partial n_{A\bar{B}}} =$ 0.2886751, when n_B increases from 40 to 50 and the rest parameters do not change the value of q and $\frac{\partial \text{q}}{\partial n_B}$ increase (q = -1.897367 and $\frac{\partial \text{q}}{\partial n_B} =$ 0.04427819), when $n_{A\bar{B}}$ increases from 4 to 8 and the rest parameters do not change, the value of q increases and $\frac{\partial \text{q}}{\partial n_B}$ does not change(q = -1.154701 and $\frac{\partial \text{q}}{\partial n_{A\bar{B}}}$ = 0.2886751).

2.2. Building the statistical implication data of measures

From the examining results of the Implication index measures, the tendency of the variation in statistical implications or partial derivatives for each parameter $n, n_A, n_B, n_{A\bar{B}}$ reflects relatively accurate trends and rate of change of the measures [8][16][15]. However, the value variation of the partial derivatives disagrees with the variation of measures. It means that they only reflect on the meaning of the derivative mathematically. If partial derivative value is positive, the measures variably increase; if partial derivative value is negative, the measures variably decrease; if partial derivative values are zero, the measures are independent with the corresponding parameters [15]. Based on the commented above, this paper builds the statistical implication data of measures based on the partial derivative values under 4 parameters by 3 principles as follows:

Principles 1: If the partial derivative values of corresponding parameter are positive, the property of measures in the corresponding parameter is set to 1 (The measures variably increase with the corresponding parameter).

Principle 2: If the partial derivative values of corresponding parameter are negative, the property of measures in the corresponding parameter is set to -1 (The measures variably decrease with the corresponding parameter).

Principle 3: If the partial derivative values of corresponding parameter are zero, the property of measures in the corresponding parameter is set to 0 (The measures are independent on corresponding parameter).

In these three principles, each measure is considered as a vector in 4-dimensional space under the form: m(v_1, v_2, v_3, v_4). For example, Recall measures with $n = 100, n_A = 20, n_B = 40, n_{A\bar{B}} = 4$, the partial derivative values of 4 parameters are determined as follows: $\frac{\partial \text{q}}{\partial \text{n}}$ = 0, $\frac{\partial \text{q}}{\partial n_A}$ = 0.025, $\frac{\partial \text{q}}{\partial n_B}$ = -0.01, $\frac{\partial \text{q}}{\partial n_{A\bar{B}}}$ = -0.025. From this partial derivative value, we define the statistical implication data of Recall measures: Recall(0, 1, -1, -1).

3. The distance and distance matrix

3.1. The distance between two measures

For clustering the measures based on the statistical implication data, the calculation of gap between two measures is an important step. Currently, there are many ways to calculate the distance between two measures in n-dimensional vector space. To calculate the distance between two measures based on the statistical implication data, we apply two calculating methods: Euclidean distance and Manhattan distance [12]. These methods have been used to calculate distance for data clustering applications since they are simple and effective. For the statistical implication data, we suppose that we need to calculate the distance between the measures with the vector form as follows: $m_1(x_1, x_2, x_3, x_4)$ and $m_2(y_1, y_2, y_3, y_4)$. We determined formula distance between two objective interestingness measures based on the statistical implication data as follows:

The Euclidean distance between m_1 and m_2 is determined by the following formula:

$$d_{Euclide}(m_1, m_2) = \left(\sum_{j=1}^{4} \left|x_j - y_j\right|^2\right)^{\frac{1}{2}} \qquad (1)$$

Example: To calculate the Euclidean distance between Confidence(0, 1, 0, -1) and Zhang (1, 1 -1, -1) as follows:

$$d_{Euclide}(\text{C}, \text{Z}) = (|0-1|^2 + |1-1|^2 + |0+1|^2 + |-1+1|^2)^{\frac{1}{2}}$$

$$d_{Euclide}(\text{C}, \text{Z}) = 1.414$$

The Manhattan distance between m_1 and m_2 is determined by the following formula:

$$d_{Manhattan}(m_1, m_2) = \sum_{j=1}^{4} \left|x_j - y_j\right| \qquad (2)$$

Example: To calculate the Manhattan distance between Coverage (-1, 1, 0, 0) and Laplace (0, 1, 0, -1) as follows:

$$d_{Manhattan}(\text{C}, \text{L}) = |0+1| + |1-1| + |0-0| + |0+1|$$

$$d_{Manhattan}(\text{C}, \text{L}) = 2$$

3.2. Distance matrix of the measures

Distance matrix of the measures is a symmetric matrix with structure: line and column of the matrix are the measures, the cells of the matrix (intersection of rows and columns) is worth the distance between two measures on corresponding line and column. Given a set of measures determined M = {m_1, m_2, ..., m_n}, each measures is described by 4-dimensional vector m_i(v_1, v_2, v_3, v_4), Distance matrix of the measures is defined as follows:

$$Matrix_{dist}(M) = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \ldots & d_{2n} \\ \vdots & . & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

$d_{ij} = d(m_i, m_j)$ is value of the distance between two measures: m_i and m_j, and is calculated by formula (1), (2) in Section 3.1. **Example**: with set M={m_1(0,1,-1,0), m_2(1,-1,0,0), m_3(-1,0,1,1), m_4(1,-1,0,1), m_5(1,0,1,0)}, then we have the Euclidean distance matrix:

$$Matrix_{dist}(M) = \begin{pmatrix} 0 & 2.45 & 2.65 & 2.65 & 2.45 \\ 2.45 & 0 & 2.65 & 1.00 & 1.41 \\ 2.65 & 2.65 & 0 & 2.45 & 2.24 \\ 2.65 & 1.00 & 2.45 & 0 & 1.73 \\ 2.45 & 1.41 & 2.24 & 1.73 & 0 \end{pmatrix}$$

4. Clustering of measures based on the hierarchical structure

4.1. Clustering algorithms for measures

Hierarchical clustering for measures is a method of clustering analysis that seeks for building a hierarchy of clusters of measures [1][2][14]. For the process of clustering, we assign each measure a cluster. Then we group two clusters with the closest distance into one cluster. This process is repeated until all measures are grouped into the same cluster.

Clustering algorithm includes the following steps:

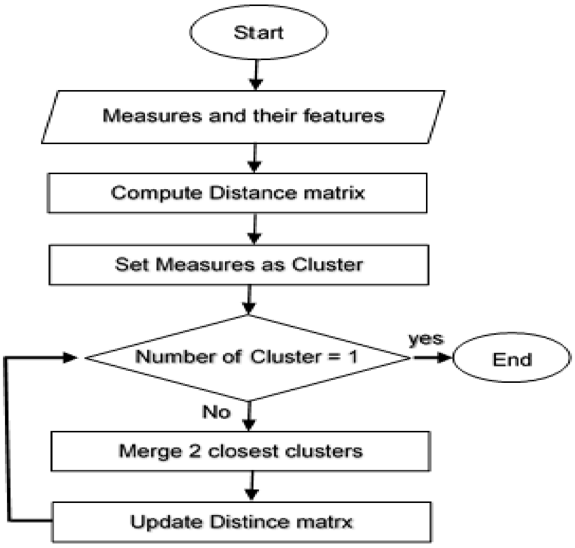
Step1: Change the properties of the measures variation in distance matrix.

Step2: Put each measure into one cluster (if we have 5 measures, we will have 5 clusters).

Step3: Repeat the two following operations until the cluster is equal to 1:

- Group 2 clusters with the closest distance into one cluster.
- Recalculate the distance matrix.

Clustering algorithm is presented on the following diagram:



4.2. Similarity tree of the measures

Similarity tree [4][5][7] of the measures is a graphical hierarchical structure which is used to express the relationship of similarity between the measures. In similarity tree of the measures, the ordering of the leaf nodes expresses similarities of one measure compared to the rest measures of the tree [22]. Two nearly leaf nodes at the same level in the tree represent the similarity of two measures. The height of the tree showing the difference between the measures is represented in the tree. Two leaf nodes spaced larger height, showing differences between the two measures are represented in the tree. Similarity tree of the measures is built up the distance matrix of the measures. The nodes of the tree will be represented in the ordered distance value following the principle: if the distance value between two measures is smaller, they are represented closer together according to the hierarchical structure.

4.3. Threshold for clustering of measures

The threshold for clustering is the smallest distance between two clustered measures. On similarity tree of measures, threshold for clustering the measures is determined based on the height of the tree. In the process of creating a similarity tree, clustering threshold is determined based on the distance matrix between the measures. At starting, the threshold for clustering of measures has the value equal to the smallest gap in the measures. This threshold will be updated after each step to build the tree and recalculate distance matrix. Clustering threshold of the measures tends to increase and reach a maximum value when all measures are mixed into one cluster.

Example: For distance matrix in the example above, we apply clustering algorithms and build similarity tree of the measures. The result is presented in Figure 2.

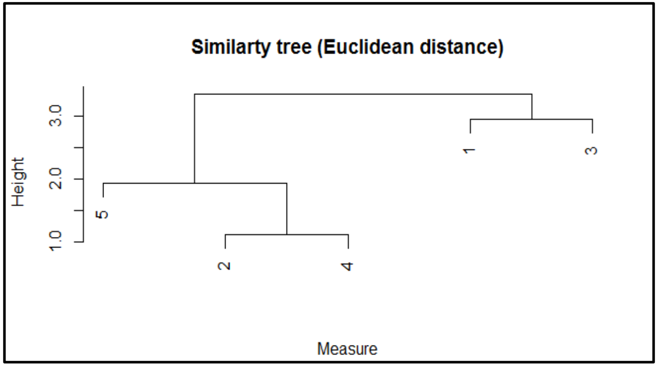


Figure 2: The similarity tree of the measures

With threshold clustering h=3, the measures are classified into two clusters: L_1={2,4,5}; L_2={1,3}.

5. Experiment

5.1. Data description

The objective of this research is to cluster the objective interestingness measures based on the tendency of variation in statistical implications, in this experiment we selected 39 objective interestingness measures agreed with asymmetric nature in order to examine partial derivative value according to 4 parameters by three principles that were defined in Section 2.2 [15]. However, in the process of examining partial derivative value of the measures we found 8 measures that their partial derivatives value is not always positive, negative or zero, where they change sign according to variations of the parameters $n, n_A, n_B, n_{A\bar{B}}$. Thus, the list of the measures agreed with three principles and only remained in this experiment is 31 measures. The results of examning on tendency of variation in statistical implications of 31 measures with ARQAT tool are presented in Table 1.

Table 1: The statistical implication data of the measures according to parameters $n, n_A, n_B, n_{A\bar{B}}$.

N^0	Measures	n	n_A	n_B	$n_{A\bar{B}}$
1	One-way Support	1	1	-1	-1
2	Added value	1	1	-1	-1
3	Bayes factor	1	1	-1	-1
4	Causal-Confidence	1	1	-1	-1
5	Causal-Confirmed confidence	1	1	-1	-1
6	Loevinger	1	1	-1	-1
7	Confidence	0	1	0	-1
8	Causal Confirm	1	1	-1	-1
9	Conviction	1	1	-1	-1
10	Coverage	-1	1	0	0
11	Descriptive Confirmed-Confidence	0	1	0	-1
12	Descriptive-Confirm	-1	1	0	-1
13	Entropic Implication Intensity 1	1	1	-1	-1
14	Entropic Implication Intensity 2	1	1	-1	-1
15	Examples and counter-examples rate	0	1	0	-1
16	Gain	-1	1	0	-1
17	Implication index	-1	-1	1	1
18	Implication Intensity	1	1	-1	-1
19	IPEE	0	-1	0	0
20	Klosgen	1	1	-1	-1
21	K-measure	1	1	-1	-1
22	Kulczynski index	0	1	-1	-1
23	Laplace	0	1	0	-1
24	Least contradiction	0	1	-1	-1
25	Leverage	1	-1	-1	-1
26	Prevalence	-1	0	1	0
27	Putative Causal Dependency	-1	1	-1	-1
28	Recall	0	1	-1	-1
29	Sebag and Schoenauer	0	1	0	-1
30	Negative Reliability	1	1	-1	-1
31	Zhang	1	1	-1	-1

5.2. Implementation tools (ARQAT)

We use ARQAT package to deploy the experimental cluster measures on language R. In this package, we have quite fully updated objective interestingness measures functions for association rules based on 4 parameters $n, n_A, n_B, n_{A\bar{B}}$, partial derivative functions of the measures, functional distance matrix calculation, integrated hierarchical clustering functions and drawing structure of similarity tree functions of stats package on R [20].

5.3. Experimental results

Based on the statistical implication data of 31 measures presented in Table 1, we set up the Euclidean distance and Manhattan distance matrix of the measures on ARQAT tools. From these distance matrices, we apply clustering functionality for measures and drawing structure of similarity tree functions on ARQAT tools to cluster the measures based on distance matrices and drawing structure of similarity tree of 31 measures. The similar trees representation corresponding to each distance calculation method is presented in Figure 3.

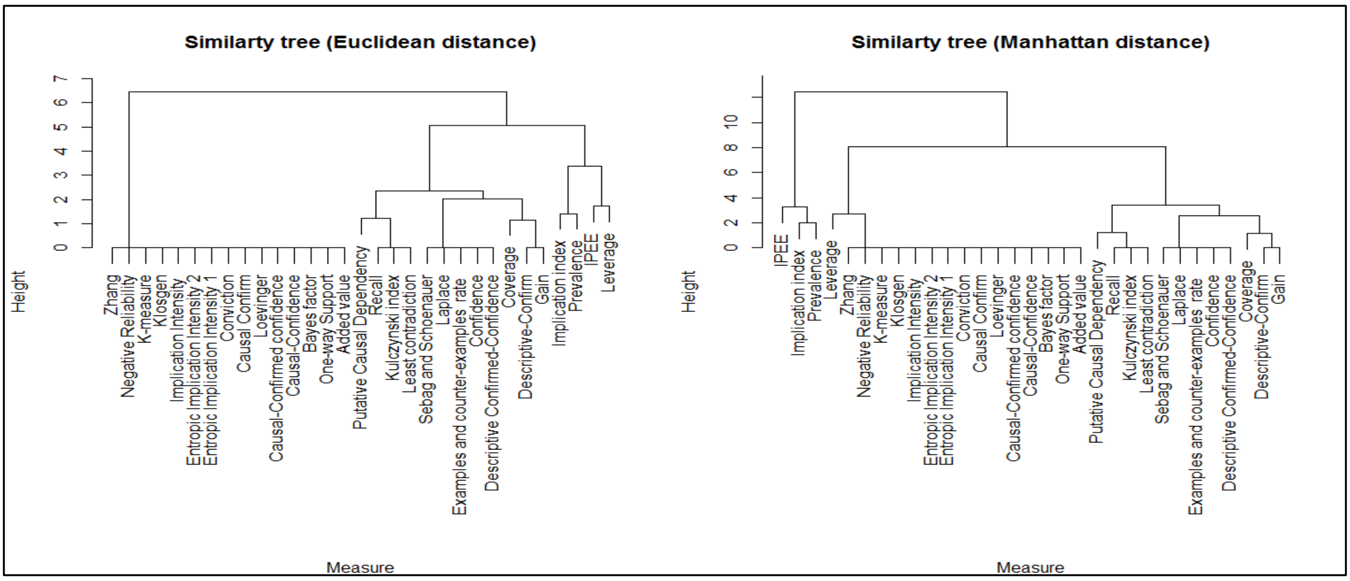


Figure 3: Similarity trees based on the statistical implication data.

Clustering results based on Euclidean distance

From similarity tree presented in Figure 3, if we consider the threshold h=2, the clustering result of 31 measures is 6 clusters. This result is presented specifically in Table 2. The first cluster includes 15 measures with the same characteristics as increasing variation with two parameters n, n_A and reducing variation with two parameters $n_B, n_{A\overline{B}}$. The second cluster includes 5 measures that have the same nature as variable increasing with parameter n_A, decreasing with parameter $n_{A\overline{B}}$ and independent with both parameters n, n_B. The third cluster includes three measures (Coverage, Descriptive Confirmed Confidence, Gain) that have similar variability on two parameters n_A, n_B and different on two parameters $\mathrm{n}, n_{A\overline{B}}$. The fourth cluster and Fifth cluster are similar in the number of measures (2 measures) but they have different distance between the measures. The measures in these clusters have the same nature varying on statistical implication parameters. The sixth cluster includes 4 measures, a special cluster created from Putative Causal Dependency and one subcluster of three measures: Kulczynski index, Least contradiction and Recall having the same nature varying on statistical implication parameters.

Table 2: The clustering results of measures based on Euclidean distance (h=2).

Clustering	List of measures of clusters
Cluster 1 (15)	Added value, One-way Support, Bayes factor, Causal-Confidence, Causal-Confirmed confidence, Loevinger, Causal Confirm, Conviction, Entropic Implication Intensity 1, Entropic Implication Intensity 2, Implication Intensity, Klosgen, K-measure, Negative Reliability, Zhang
Cluster 2 (5)	Descriptive Confirmed-Confidence, Confidence, Examples and counter-examples rate, Laplace, Sebag and Schoenauer
Cluster 3 (3)	Coverage, Descriptive-Confirm, Gain
Cluster 4 (2)	Implication index, Prevalence
Cluster 5 (2)	IPEE, Leverage
Cluster 6 (4)	Least contradiction, Kulczynski index, Recall, Putative Causal Dependency

Clustering results based on Manhattan distance

Basing on the structure of similar tree built in Manhattan distance matrix, with the threshold h = 2, the measures are divided into 7 clusters (Table 3). The first cluster consists of 15 measures that they have the same natures as increasing variation with two parameters n, n_A and reducing variation with two parameters $n_B, n_{A\overline{B}}$. The second cluster includes 5 measures that have the same characteristics as variable increasing with parameter n_A, decreasing with parameter $n_{A\overline{B}}$ and independent with both parameters n, n_B. The third cluster is a particular cluster since it formed from a cluster consisting of two measure with the same varied properties (Descriptive-Confirm, Gain) and one measure of variability is not of the same natures (Coverage). The fourth cluster formed by two measures is not of property variability in both parameters $n_A, n_{A\overline{B}}$ but their distance is equal to the level of clustering. The fifth cluster has only IPEE measures. This measures have special properties varied independent on all three parameters $\mathrm{n}, n_B, n_{A\overline{B}}$ and decreasing variability on parameter n_A. The sixth cluster includes 04 measures, a special cluster created from Putative Causal Dependency and one subcluster of three measures: Kulczynski index, Least contradiction and Recall. They have the same nature varying on statistical implication parameters. The seventh cluster also has only one measure (Leverage). This measure reduces variability with all three parameters $\mathrm{n}_A, n_B, n_{A\overline{B}}$ and increases variable parameter n.

Table 3: The clustering results of measures based on Manhattan distance (h=2).

Clustering	List of measures of clusters
Cluster 1 (15)	Added value, One-way Support, Bayes factor, Causal-Confidence, Causal-Confirmed confidence, Loevinger, Causal Confirm, Conviction, Entropic Implication Intensity 1, Entropic Implication Intensity 2, Implication Intensity, Klosgen, K-measure, Negative Reliability, Zhang
Cluster 2 (5)	Descriptive Confirmed-Confidence, Confidence, Examples and counter-examples rate, Laplace, Sebag and Schoenauer
Cluster 3 (3)	Coverage, Descriptive-Confirm, Gain
Cluster 4 (2)	Implication index, Prevalence
Cluster 5 (1)	IPEE
Cluster 6 (4)	Least contradiction, Kulczynski index, Recall, Putative Causal Dependency
Cluster 7 (1)	Leverage

Comparison of the Clustering results

Based on the similarity trees in Figure 3, overall, clustering results in two measurement distances are relatively similar. When considering the threshold h=2, both clustering results have the list of measures in the majority of clusters that are similar as cluster 1, cluster 2 and clusters 3, cluster 4 and cluster 6. However, with clustering results based on Euclidean distance matrix, IPEE and Leverage are classified in the same clusters but with clustering results based on Manhattan distance matrix, IPEE and Leverage are classified in different clusters. As clustering thresholds increasing h=4, both similarity trees have 3 clusters for two distance measurement methods (Table 4, 5). In particular, the second cluster of both trees has the same list of measures. The first cluster and third cluster have the list of measures that differ on both similarity trees. The reason of deviation is that Leverage measures are located in two different clusters on two similarity trees. In the tree according to Euclidean distance, Leverage measures are classified into third cluster while the other tree, Leverage measures is classified in the first cluster.

Clustering results show that the group of objective interestingness measures agreed with asymmetric properties is mainly classified into three large clusters: Cluster 1, Cluster 2 and Cluster 6. Each cluster consists of the measures with the same characteristics following the tendency of variation in statistical implications. For example, Cluster 2 is a group of measures variable increasing with parameter n_A, decreasing with parameter $n_{A\overline{B}}$ and independent with both parameters n, n_B. The results also show the distance of the intensity variation of the measures under statistical implication parameters. This is a useful basis for further study about the relationship between the measures based on an examinination of their partial derivatives.

Table 4: The clustering results of measures based on Euclidean distance (h=4).

Clustering	List of measures of clusters
Cluster 1 (15)	Added value, One-way Support, Bayes factor, Causal-Confidence, Causal-Confirmed confidence, Loevinger, Causal Confirm, Conviction, Entropic Implication Intensity 1, Entropic Implication Intensity 2, Implication Intensity, Klosgen, K-measure, Negative Reliability, Zhang
Cluster 2 (12)	Descriptive Confirmed-Confidence, Confidence, Examples and counter-examples rate, Laplace, Sebag and Schoenauer, Coverage, Descriptive-Confirm, Gain, Least contradiction, Kulczynski index, Recall, Putative Causal Dependency
Cluster 3 (4)	Implication index, Prevalence, IPEE, Leverage

Table 5: The clustering results of measures based on Manhattan distance (h=4).

Clustering	List of measures of clusters
Cluster 1 (16)	Added value, One-way Support, Bayes factor, Causal-Confidence, Causal-Confirmed confidence, Loevinger, Causal Confirm, Conviction, Entropic Implication Intensity 1, Entropic Implication Intensity 2, Implication Intensity, Klosgen, K-measure, Negative Reliability, Zhang, Leverage
Cluster 2 (12)	Descriptive Confirmed-Confidence, Confidence, Examples and counter-examples rate, Laplace, Sebag and Schoenauer, Coverage, Descriptive-Confirm, Gain, Least contradiction, Kulczynski index, Recall, Putative Causal Dependency
Cluster 3 (3)	Implication index, Prevalence, IPEE

6. Conclusion

Clustering the objective interestingness measures is attracted to many researchers in the field of data mining. The study for clustering measures is primarily based on three main techniques: clustering based on partition, clustering based on hierarchical and clustering based on density. In this article we propose clustering method to cluster the objective interestingness measures based on the tendency of variation in statistical implications by hierarchical clustering techniques. From the statistical implication data, distance matrices of measures are built on two distance calculation methods of Euclidean and Manhattan. After calculating the distance matrices, we use our tools to build similarity trees for clustering 31 measures. The similarity trees show that the measures are classified with two clustering thresholds h=2 and h=4. This result can be used to support the choice of the appropriate measure of researchers and users for their specific applications and is also useful basis for further study about the objective interestingness measures bassed on the tendency of variation in statistical implications.

References

[1] Abdolreza Mirzaei, Mohammad Rahmati and Majid Ahmadi, (2008) A new method for hierarchical clustering combination, Volume 12 Issue 6, 549-571.

[2] Cheng-Hsien Tang, Meng-Feng Tsai, Shan-Hao Chuang, Jen-Jung Cheng, Wei-Jen Wang, (2014) Shortest-linkage-based parallel hierarchical clustering on main-belt moving objects of the solar system, Future Generation Computer Systems archive Volume 34, 26-46.

[3] Couturier, (2008) CHIC: Cohesive Hierarchical Implicative Classification, Studies in Computational Intelligence (SCI) 127, 41–53.

[4] Espinoza et al., (2011) Using Hierarchical Clustering and Dendrograms to Quantify the Clustering of Membrane Proteins, A Journal Devoted to Research at the Junction of Computational, Theoretical and Experimental Biology Official Journal of The Society for Mathematical Biology ISSN 0092-8240.

[5] Flor A. Espinoza, Janet M. Oliver, Bridget S. Wilson and Stanly L. Steinberg, (2011) Using Hierarchical Clustering and Dendrograms to Quantify the Clustering of Membrane Proteins, Springer, 1-24.

[6] Geng and Hamilton, (2006) Interestingness measures for data mining: A survey, ACM Computing Surveys (Volume 38), 1-32.

[7] Gleb B. Sologub, (2011) On measuring of similarity between tree nodes, Young Scientists Conference in Information Retrieval, 63-71.

[8] Gras and Kuntz, (2008) An overview of the Statistical Implicative Analysis (SIA) development, Statistical Implicative Analysis - Studies in Computational Intelligence (Volume 127), Springer-Verlag, 11-40.

[9] H. X. Huynh et al., (2007) A graph-based clustering approach to evaluate interestingness measures: a tool and a comparative study (Chapter 2), Quality Measures in Data Mining, Springer-Verlag, 25-50.

[10] H. X. Huynh et al., (2012) Classification of objective interestingness measures, Journal of Can Tho University (2011:20a), 147 – 158.

[11] Hiep Xuan Huynh, Lan Phuong Phan, Nghia Quoc Phan, Bac Hoai Le, Fabrice Guillet, (2015) Classification of objective interestingness measures based on interestingness criteria, International of Expert System with Applications, submitted.

[12] Michel Marie Deza and Elena Deza, (2014) Encyclopedia of Distances, Springer.

[13] Mohammed Dabboor, John Yackel, Mosharraf Hossain and Alexander Braun, (2013) Comparing matrix distance measures for unsupervised POLSAR data classification of sea ice based on agglomerative clustering, International Journal of Remote Sensing, 1492-1505.

[14] Murtagh and Legendre, (2013) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification* (in press).

[15] Nghia Quoc Phan, Hiep Xuan Huynh, Fabrice Guillet and Régis Gras, (2015) Classifying objective interestingness measures based on the tendency of value variation, VIII Colloque International –VIII International Conference, A.S.I. Analyse Statistique Implicative - Statistical Implicative Analysis Radès (Tunisie) - Novembre 2015, 143-172.

[16] Régis GRAS, Pascale KUNTZ et Nicolas GREFFARD, (2015) NOTION DE CHAMP IMPLICATIF EN ANALYSE STATISTIQUE IMPLICATIVE, VIII Colloque International –VIII International Conference, A.S.I. Analyse Statistique Implicative - Statistical Implicative Analysis Radès (Tunisie) - Novembre 2015, 29-46.

[17] Sadaaki Miyamoto, (2012) An overview of hierarchical and non-hierarchical algorithms of clustering for semi-supervised classification, Springer-Verlag Berlin, Heidelberg ©2012, 1-10.

[18] Satoshi Takumi and Sadaaki Miyamoto, (2011) Agglomerative hierarchical clustering using asymmetric similarity based on a bag model and application to information on the web, Springer-Verlag Berlin, Heidelberg ©2011, 187-196.

[19] Satoshi Takumi and Sadaaki Miyamoto, (2012) Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering, Granular Computing (GrC), IEEE International Conference, 459 – 464.

[20] Sinnwell and Schaid, (2015) Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous.

[21] Tew et al., (2013) Behavior-based clustering and analysis of interestingness measures for association rule mining, Journal of Data Mining and Knowledge Discovery 28, Springer-Verlag, 1004-1045.

[22] Yu Tang, Yilun Cai, Nikos Mamoulis, (2015) Scaling Similarity Joins over Tree-Structured Data, Proceedings of the VLDB Endowment , Volume 8 Issue 11, 1130-1141.