# A Data Analysis Approach for Evaluating the Behavior of Interestingness Measures

Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand

LINA CNRS 2729 - Polytechnic School of Nantes University,
La Chantrerie BP 50609 44306 Nantes cedex 3, France
{xuan-hiep.huynh, fabrice.guillet, henri.briand}@polytech.univ-nantes.fr

**Abstract.** In recent years, the problem of finding the different aspects existing in a dataset has attracted many authors in the domain of knowledge quality in KDD. The discovery of knowledge in the form of association rules has become an important research. One of the most difficult issues is that an enormous number of association rules are discovered, so it is not easy to choose the best association rules or knowledge for a given dataset. Some methods are proposed for choosing the best rules with an interestingness measure or matching properties of interestingness measure for a given set of interestingness measures. In this paper, we propose a new approach to discover the clusters of interestingness measures existing in a dataset. Our approach is based on the evaluation of the distance computed between interestingness measures. We use two techniques: agglomerative hierarchical clustering (AHC) and partitioning around medoids (PAM) to help the user graphically evaluates the behavior of interestingness measures.

## 1 Introduction

Knowledge quality has become an important issue of recent researches in KDD. The problem of selecting the best knowledge for a given dataset has attracted many authors in the literature. Our approach is based on the knowledge representation in the form of association rules [2], one of the few models dedicated to unsupervised discovery of rules tendencies in data. With association rules, many authors have proposed a lot of interestingness measures to evaluate the best matched knowledge from a ruleset: to select the best measures or the best rules. According to Freitas [5], two kinds of interestingness measures existing can be differentiate: objective and subjective. Subjective measures are strongly influenced by the user's goals and his/her knowledge or beliefs, and are combined to specific supervised algorithms in order to compare the extracted rules with what the user knows or wants [13] [12], rule novelty and unexpectedness in point of view of the user are captured. Objective measures are statistical indexes that depend strictly on the data structures. The definitions and properties of many objective measures are proposed and surveyed [3] [8] [16] to study the behavior of the objective measures to design a suitable measure or to help the user to select the best ones with their preferences. We focus on objective measures (called measure for short) as a natural way to discover different hidden aspects in the data.

A. Hoffmann, H. Motoda, and T. Scheffer (Eds.): DS 2005, LNAI 3735, pp. 330–337, 2005.

Many interesting surveys on objective measures can be found in the literature. They mainly address two related research issues: the definition of the set of principles or properties that lead to the design of a good measure; their comparison from a data-analysis point of view to study measure behavior in order to help the user to select the best ones [8][16][17][10].

In this paper, we propose a new approach to evaluate the behavior of 35 interestingness measures discussed in the literature to discover the clusters of interestingness measures existing in the user's dataset. Our approach is based on the distance computed between interestingness measures by using the two clustering methods agglomerative hierarchical clustering (AHC) and partitioning around medoids (PAM) to help the user to discover the behavior of the interestingness measures studied in his/her dataset graphically.

The paper is organized as follows. In Section 2, we present the correlation and the distance between measures. In Section 3, we introduce two views for evaluating the behavior of a set of 35 measures on a dataset. Finally, we conclude and introduce some future researches.

## 2 Distance Between Measures

Based on the idea of measuring the the statistical surprisingness of implication theory [7] that we have mentioned in [10], we continue to extend the principles discussed from [10]. Let $R(D) = \{r_1, r_2, ..., r_p\}$ denote input data as a set of $p$ association rules derived from a dataset $D$. Each rule $a \Rightarrow b$ is described by its itemsets $(a, b)$ and its cardinalities $(n, n_a, n_b, n_{a\bar{b}})$. Let $M$ be the set of $q$ available measures for our analysis $M = \{m_1, m_2, ..., m_q\}$. Each measure is a numerical function on rule cardinalities: $m(a \Rightarrow b) = f(n, n_a, n_b, n_{a\bar{b}})$.

For each measure $m_i \in M$, we can construct a vector $m_i(R) = \{m_{i1}, m_{i2}, ..., m_{ip}\}$, $i = 1..q$, where $m_{ij}$ corresponds to the calculated value of the measure $m_i$ for a given rule $r_j$.

The correlation value between any two measures $m_i, m_j \{i, j = 1..q\}$ on the set of rules $R$ will be calculated by using a Pearson's correlation coefficient $CC$ [15], where $\overline{m_i}, \overline{m_j}$ are the average calculated values of vector $m_i(R)$ and $m_j(R)$ respectively.

$$CC(m_i, m_j) = \frac{\sum_{k=1}^{p}[(m_{ik} - \overline{m_i})(m_{jk} - \overline{m_j})]}{\sqrt{[\sum_{k=1}^{p}(m_{ik} - \overline{m_i})^2][\sum_{k=1}^{p}(m_{jk} - \overline{m_j})^2]}}$$

In order to interpret the correlation value, we introduce the two following definitions:

**Definition 2.** *Correlated measures* ($\tau$-correlation). Two measures $m_i$ and $m_j$ are correlated with respect to the dataset $D$ if their absolute correlation value is greater than or equal to a threshold $\tau$: $|CC(m_i, m_j)| \geq \tau$.

**Definition 3.** *Distance.* The distance $d$ between two measures $m_i$,$m_j$ is defined by:

$$d(m_i, m_j) = 1 - |CC(m_i, m_j)|$$

As both correlation and distance are symmetrical, the $q(q-1)/2$ values can be stored in one half of a table $q \times q$. We then use the distances computed from this table for both the AHC and PAM methods.

## 3 Measure Behavior

### 3.1 Data Description and Used Measures

To study the measure behavior, we try to evaluate the effect of measures based on the distance calculations for the dataset $D_1$. We use the categorical dataset *mushroom* ($D_1$) from Irvine machine-learning database repository [4]. We then generate the set of association rules (ruleset) $R_1$ from the dataset $D_1$ using the the algorithm Apriori [1]. We use 35 interestingness measures to this study (34 measures are referenced in [10] and a measure $II = 1 - \sum_{k=max(0,n_a-n_b)}^{n_{a\bar{b}}} \frac{C_{n_b}^{n_a-k} C_{n_{\bar{b}}}^{k}}{C_n^{n_a}}$). A remark is that EII[$\alpha = 1$] and EII[$\alpha = 2$] are two entropic versions of the II measure). Hereafter, we use this ruleset as our knowledge data for analysis.

**Table 1.** Ruleset description

| Dataset | Items | Transactions | Average length of transactions | Number of rules (support threshold) | Ruleset |
|---|---|---|---|---|---|
| $D_1$ | 118 | 8416 | 22 | 123228 (12%) | $R_1$ |

Our aim is to discover the behavior of the measures via two views: the strong relation and the relative distance between measures occur when they are applied to the distance matrix (or distance table) calculated from $R_1$ (see Sec. 2). This result is useful because we can capture the different aspect or *the nature of the available knowledge* existing in the rulesets. We use the two techniques AHC and PAM for each of these views respectively.

### 3.2 With AHC

Fig. 1 illustrates the result computed from $R_1$. The horizontal line goes throughout the cluster dendrogram has the small distance 0.15 determining the clusters of measures having strong relation (strongly correlated). The assignment $\tau = 0.85 = 1 - 0.15$ of $\tau$-correlation is used because this value is widely acceptable in the literature. The clusters are represented in details in Tab. 2.

Intuitively, the user can choose the biggest cluster in Tab. 2 contains the measures Lift, Rule Interest, Phi-Coefficient, Kappa, Similarity Index, Putative Causal Dependency, Dependency, Klosgen, Pavillon for their first choice. In this cluster we can easily see two strong related clusters with four measures for each. This cluster gives the strongest effect on evaluation the similarity between two parts of an association rule. Another observation illustrates the existence of a

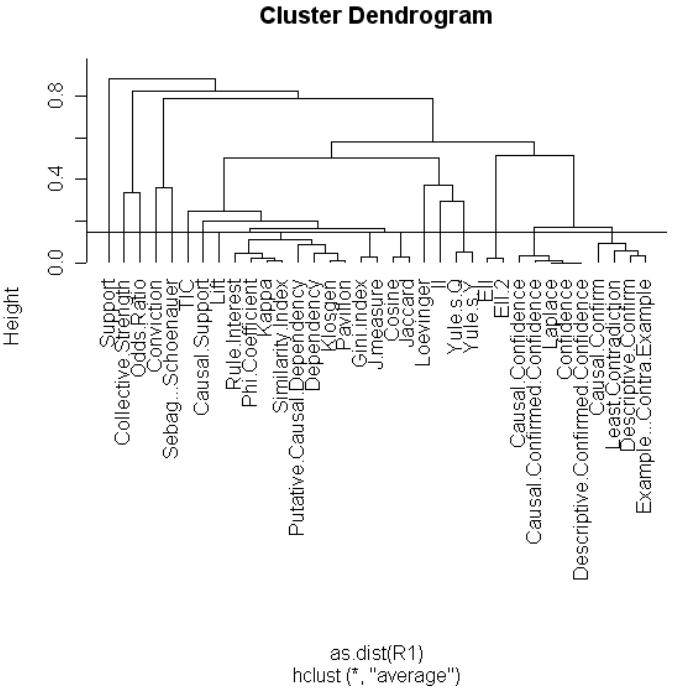


**Fig. 1.** View on the strong relation between measures

**Table 2.** Clusters of measures with AHC (distance = 0.15)

| Cluster | $R_1$ |
|---|---|
| 1 | Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace |
| 2 | Causal Confirm, Descriptive Confirm, Example & Contra-Example, Least Contradiction |
| 3 | Causal Support |
| 4 | Collective Strength |
| 5 | Conviction |
| 6 | Cosine, Jaccard |
| 7 | Dependency, Kappa, Klosgen, Lift, Pavillon, Phi-Coefficient, Putative Causal Dependency, Rule Interest, Similarity Index |
| 8 | EII, EII 2 |
| 9 | Gini-index, J-measure |
| 10 | II |
| 11 | Loevinger |
| 12 | Odds Ratio |
| 13 | Sebag & Schoenauer |
| 14 | Support |
| 15 | TIC |
| 16 | Yule's Q, Yule's Y |

confidence cluster (the first cluster in Tab. 2) with Causal Confidence, Causal Confirmed-Confidence, Laplace, Confidence, Descriptive Confirmed-Confidence. The user can then select this cluster to discover all the rules have the effect of high confidence.

This view is useful because the user can determine the strong relation between interestingness measures via the graphical representation. The hierarchical

structure allows the user clearly seeing the clusters of measures that are connected closely with the hierarchical level computed.

### 3.3 With PAM

We can see relatively the distance between clusters by applying the principal component analysis, the number of cluster is determined from the first view (Sec. 3.2). For example, Fig. 2 illustrates the result obtained from $R_1$. Each symbol from Fig. 2 represents every measure in the same cluster. PAM is very useful because it gives a graphical view of clusters intuitively.

The user can now choose the aspects in the ruleset by viewing the clusters with their distances calculated (Fig. 2) based on the projection on the two principal components. The measures that have the smallest distances between them will be grouped in one cluster. In Tab. 3 the two clusters 1 (Causal Confirmed-Confidence, Descriptive Confirmed-Confidence, Confidence, Laplace, Causal Confidence) and 2 (Least Contradiction, Example & Contra-Example, Causal Confirm, Descriptive Confirm) as two different aspects the most close with the very small between-distance or separation. Then, the user can obtain automatically the representative measures for each of these two clusters are Causal Confirmed-Confidence and Example & Contra-Example. Another useful

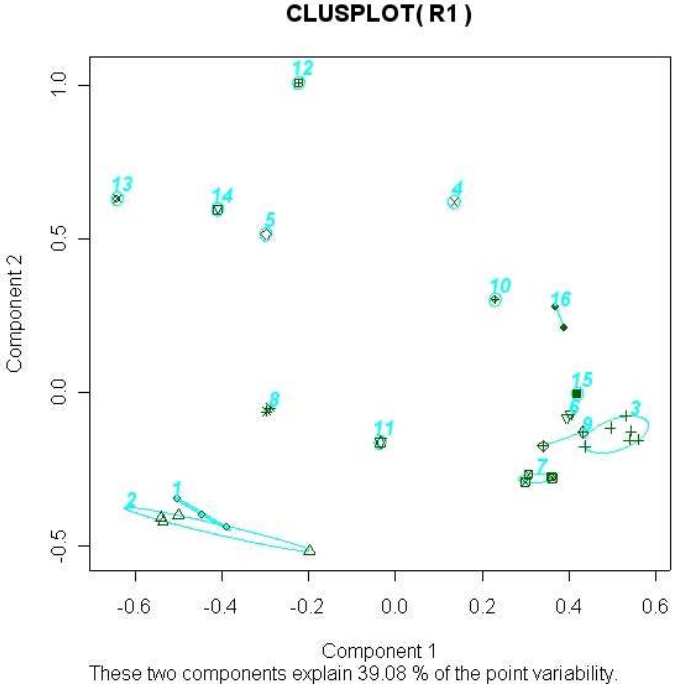


**Fig. 2.** Views on the relative distance between clusters of measures

**Table 3.** Clusters of measures with PAM

| Cluster | $R_1$ |
|---|---|
| 1 | Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace |
| 2 | Causal Confirm, Descriptive Confirm, Example & Contra-Example, Least Contradiction |
| 3 | Causal Support, Kappa, Lift, Phi-Coefficient, Rule Interest, Similarity Index |
| 4 | Collective Strength |
| 5 | Conviction |
| 6 | Cosine, Jaccard |
| 7 | Dependency, Klosgen, Pavillon,Putative Causal Dependency |
| 8 | EII, EII 2 |
| 9 | Gini-index, J-measure |
| 10 | II |
| 11 | Loevinger |
| 12 | Odds Ratio |
| 13 | Sebag & Schoenauer |
| 14 | Support |
| 15 | TIC |
| 16 | Yule's Q, Yule's Y |

information is that the diameter of the cluster 1 is smaller than the cluster 2 so this observation illustrates the strongly coherent interestingness values computed from the measures in cluster 1, representing the high value of the confidence aspect. Another choice is that the user can select in Tab. 3 one aspect formed by the cluster 3 (Causal Support, Kappa, Lift, Phi-Coefficient, Rule Interest, Similarity Index) that is very far from the two clusters 1&2 but the nearest cluster with the others such as 9,7,15 (Fig. 3) and having Kappa as the representative measure for this cluster. The user can also interest in the cluster 10 (II) in Tab. 3 standing isolated with other clusters (Fig. 3).

This view based on relative distance has an important role because it allows the user to choose the aspects that he/she takes interested by regarding the scale between them. The distance between clusters will help the user to evaluate more precisely the near or far between these aspects.

### 3.4 Comparing with AHC and PAM

With two different evaluations based on the two views of AHC and PAM we can obtain some interesting results: cluster that seems independent from the nature of data and the selection of rules. Comparing from Tab. 2 and Tab. 3 we can easily see sixteen clusters agreed perfectly (see Tab. 4).

To understand the behavior of the measures we will examine some important clusters in Tab. 4. For example, the first cluster (Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace) has most of the measures issued from the Confidence measure. The fifth cluster (Cosine, Jaccard) has a strong relation with the fifth property proposed by Tan et al. [16]. The sixth cluster (Dependency, Klosgen, Pavillon, Putative Causal Dependency) is necessary to distinguish between the strength of the rule $a \Rightarrow b$ from $b \Rightarrow a$. The seventh cluster (EII, EII 2) are two measures obtained with different parameters of the same original formula and very useful in evaluating the entropy of implication intensity. The ninth cluster (II) has only one measure provides the strong evaluation on the intensity of implication. The tenth cluster (Kappa, Lift, Phi-Coefficient, Rule Interest, Similarity Index) mainly gathers the

**Table 4.** Clusters agreed with both AHC and PAM

| Cluster | $R_1$ |
|---|---|
| 1 | Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace |
| 2 | Causal Confirm, Descriptive Confirm, Example & Contra-Example, Least Contradiction |
| 3 | Collective Strength |
| 4 | Conviction |
| 5 | Cosine, Jaccard |
| 6 | Dependency, Klosgen, Pavillon,Putative Causal Dependency |
| 7 | EII, EII 2 |
| 8 | Gini-index, J-measure |
| 9 | II |
| 10 | Kappa, Lift, Phi-Coefficient, Rule Interest, Similarity Index |
| 11 | Loevinger |
| 12 | Odds Ratio |
| 13 | Sebag & Schoenauer |
| 14 | Support |
| 15 | TIC |
| 16 | Yule's Q, Yule's Y |

measures from different properties such as symmetry, anti-symmetry [16]. The last cluster (Yule's Y, Yule's Q) gives a trivial observation because the measures are all derived from Odds Ratio measure, that is similar to the second property proposed by Tan et al. [16].

## 4 Conclusion

To understand the behavior of the interestingness measures on a specific dataset, we have studied and compared the various interestingness measures described in the literature to find the different aspects existing in a dataset. Our approach is the first step towards the process of evaluating the knowledge issued in the form of association rules in the domain of knowledge quality research. We use a data analysis approach based on the distance computed between interestingness measures (with two clustering methods AHC and PAM) in order to evaluate the behavior of 35 interestingness measures. These two graphically clustering methods can be used to help a user in selecting the best measures. We also determine sixteen clusters with some interesting results: cluster that seems independent from the nature of data and the selection of rules. We also evaluate the behavior of the measures on some important clusters agreed with both AHC and PAM. With this result, the decision-maker will decide what measures are interesting to capture the best knowledge.

Our future research will be investigated in introducing a new approach to facilitate the the user's decision making from the best interestingness measures to select the best association rules (the best knowledge discovered).

## References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. Proc. of the 20th VLDB. Santiago, Chile (1994) 487–499
2. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Proc. of the ACM-SIGMOD Int. Conf. on Management of Data. Washington DC, USA (1993) 207–216

3. Bayardo, Jr.R.J., Agrawal, R.: Mining the most interestingness rules. KDD'1999. San Diego, CA, USA (1999) 145–154
4. Blake, C.L., Merz, C.J.: {UCI} Repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html. University of California, Irvine, Dept. of Information and Computer Sciences, (1998).
5. Freitas, A.A.: On rule interestingness measures. Knowledge-Based Systems, 12(5-6). (1999) 309–315
6. Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., Peter, P.: Quelques critères pour une mesure de qualité de règles d'association. Mesures de Qualité pour la Fouille de Données, RNTI-E-1. Cépaduès Editions (2004) 3–31
7. Gras, R.: L'implication statistique - Nouvelle méthode exploratoire de données. La pensée sauvage édition (1996)
8. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interestingness. Kluwer Academic Publishers (2001)
9. Huynh, X.H., Guillet, F., Briand, H.: ARQAT: an exploratory analysis tool for interestingness measures. ASMDA'05. (2005) 334–344
10. Huynh, X.H., Guillet, F., Briand, H.: Clustering interestingness measures with positive correlation. ACM ICEIS'05. (2005) 248–253
11. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, (1990)
12. Liu, B., Hsu, W., Mun, L., Lee, H.: Finding interestingness patterns using user expectations. IEEE Trans. on Knowledge and Data Mining (11). (1999) 817–832
13. Padmanabhan, B., Tuzhilin, A. : A belief-driven method for discovering unexpected patterns. KDD'1998. (1998) 94–100
14. Piatetsky-Shapiro, G.: Discovery, analysis and presentation of strong rules. Knowledge Discovery in Databases, G. Piatetsky-Shapiro and W. Frawley editors. MIT Press, Cambridge, MA (1991) 229–248
15. Saporta, G.: Probabilité, analyse des données et statistique. Edition Technip, (1990)
16. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems 29(4). (2004) 293–313
17. Vaillant, B., Lenca, P., Lallich, S.: A clustering of interestingness measures. DS'04. (2004) 290-297