UDC 004.89
Original Scientific Paper

# Adapting hierarchical clustering distance measures for improved presentation of relationships between transaction elements

**Mihaela Vranić** *mihaela.vranic@fer.hr*
*University of Zagreb*
*Faculty of Electrical Engineering and Computing*

**Damir Pintar** *damir.pintar@fer.hr*
*University of Zagreb*
*Faculty of Electrical Engineering and Computing*

**Dragan Gamberger** *dragan.gamberger@irb.hr*
*Ruđer Bošković Institute*
*Department of electronics*

## Abstract

Common goal of descriptive data mining techniques is presenting new information in concise, easily interpretable and understandable ways. Hierarchical clustering technique for example enables simple visualization of distances between analyzed objects or attributes. However, common distance measures used by existing data mining tools are usually not well suited for analyzing transactional data using this particular technique. Including new types of measures specifically aimed at transactional data can make hierarchical clustering a much more feasible choice for transactional data analysis. This paper presents and analyzes convenient measure types, providing methods of transforming them to represent distances between transaction elements more appropriately. Developed measures are implemented, verified and compared in hierarchical clustering analysis on both artificial data as well as referent transactional datasets.

**Keywords:** Transactional Data, Distance Measures, Linkage Criteria, Hierarchical Clustering

## 1. Introduction

Descriptive data mining is commonly used as a first step in data mining analysis. Although frequently followed by certain predictive data mining methods, descriptive methods alone can also provide useful information based on which certain decisions and real world actions can be made. Results of descriptive methods are most useful when they offer concise and clear representation of relationships between analyzed objects [1], [2].

This paper will focus on transactional data analysis - where transactional data is defined as a set of digital records each describing a specific event (business or otherwise). Single transaction usually consists of a timestamp and a set of references to objects participating in described event. Today's information systems store huge amounts of such transactional data, which often after their initial usage in transactional systems remain unused. Subsequent analysis will most commonly focus on aggregated or summarized form of this data, while transactions on lowest level are often ignored. However, focusing solely on aggregated information will ultimately result in lack insights in interesting fine-grained relationships between objects referenced by transactions, even though these relationships may turn out to be extremely beneficial in the process of creating important real-world decisions. For this reason discovery and efficient representation of such relationships is chosen as the main focus of this paper.

When speaking of transactional data analysis, most commonly used method is association rule generation - introduced in [3]. This method has certain drawbacks, some being: large quantity of resulting rules, inherent redundancy between rules and difficulties in discerning helpful rules apart from those that do not provide useful information. Additionally, association rules generation method is usually very resource-intensive and highly sensitive to chosen threshold parameters. One of the other existing and widely used methods for discovering and visual representation of relationships between analyzed objects is the hierarchical clustering method. This method works very well in various data analysis scenarios. However, it is not very well suited for transactional data specifically, the reason being that hierarchical clustering relies on measuring distances between objects, and the usual ways of defining distance aren't suitable for depicting relationships between transactional data objects.

To deal with this issue, this paper suggests and discusses new distance measures specifically aimed toward transactional data. New measures are based on existing interestingness measures developed for transactional data, description and analysis of which can be found in [4], [5], [6], [7]. These measures of interestingness were primarily designed for assessment and ordering of association rules [8], [9]. Their potential usage in hierarchical clustering has not yet been investigated, mostly because by design they aren't intended to measure distance. This paper investigates the semantics behind these measures and describes the transformations required to make their new form suit the concept of distance more appropriately. In addition to transforming existing measures, an entirely new measure was also developed. The final result allows us to efficiently bridge the gap between transactional data objects and the method of hierarchical clustering.

To estimate the efficiency of all developed measures, a reference implementation was made. This implementation was based on open-source Orange data mining tool [10]. Two new modules were developed and subsequently used for both real-life and artificial dataset analysis.

This paper is organized as follows: Section 2 presents related work on the subject. Section 3 discusses the method of hierarchical clustering and properties of transactional data in more detail. After that, in Section 4 chosen measures of interestingness are presented and methods of transforming them to appropriately represent distance are described. Section 5 shows the application on these measures on an artificially constructed dataset as well as on a real-life input dataset. Thorough examination and comparison of measures is also provided. Finally, Section 6 offers final insights and conclusions on the issues presented in this paper.

## 2. Related Work

Hierarchical clustering technique is highly sensitive to the choice of distance measure. Choosing the right distance measure which suits best for given input dataset is one of the most delicate and important decisions the analyst using this technique must make, as discussed in [2]. Additionally, research done on this issue on behalf of the authors of this paper is given in [11], which describes development of a new measure with the intent of measuring distance between objects represented by attributes belonging to heterogeneous data types. A new challenge - providing a measure of distance suited for transactional data - has been introduced and analyzed in [12]. Following chapters describe the continuation of this research and present our proposed solutions.

Certain visualizations of relationships between transaction elements have already been developed. Those visualizations are most commonly tied with association rules discovery, which is considered to be a native method for transactional data analysis. In [13] certain ways of visualizing generated association rules are provided. One solution presents rule visualization through a three-dimensional graph created by placing the left side of the rule on the X axis, and right side of the rule on the Y axis. Rule confidence is represented by a column determined by the intersection of x and y coordinates. In this case, by looking at the graph the analyst can determine interesting groupings of rules, even though the ease of discerning the groupings will highly depend on the

order of transaction elements on the axes. Described visualization is available in the MineSet data mining tool provided by Silicon Graphics. Some of the example visualizations obtained by this tool can be found in [13]. One of the drawbacks of these visualizations is their focus on a narrow set of measures closely tied with association rules (mostly confidence or lift). In certain cases other types of measures founded on different semantics might be more suitable for representing relationships analyst might be interested in. A significant amount of research dealing with evaluation and discussion of possible measures which could be used with mined association rules ordering is given in [4], [5], [6], [7], [14].

One additional way of visualizing association rules is through mosaic plots, as shown in [15]. This visualization represents a single rule which can have two members on the left side and one member on the right side. While this visualization technique is intuitive and rather interesting, the fact that it is restricted to only one single rule at the time greatly limits its potential and usefulness in analysis with a broader scope.

Visualizations based on hierarchical clustering offer certain advantages over visualizations based on generated association rules. First and foremost, using hierarchical clustering ensures that each transaction element will be shown only once, greatly enhancing the clarity of presentation. Furthermore, resulting structure has hierarchical properties which directly reflect chosen measure and linkage criteria. Finally, hierarchical clustering is a much faster method since the distances are calculated only once for each pair of elements. Tree structures in particular are a natural choice when it comes to representing hierarchical structures. Some previous work by the authors revolves around using tree structures for representing relationships between transaction elements [16]. In that particular approach tree is being constructed sequentially, each step including a separate association rule generation process. Based on generated rules and other input parameter thresholds certain transaction elements are linked to form each tree level. Newly formed groups are treated as single elements in subsequent steps of the process which demands additional transformations of input data. This method results in useful structures but was shown to be extremely resource-intensive, mostly because of the repeated need to generate new association rules throughout the entire process.

This paper offers a solution for presenting relationships between transaction elements through hierarchical structures - dendrograms. This solution will avoid some of the pitfalls of previously mentioned visualizations, such as being restricted to representing only a small subset of transaction element relationships, showing same elements multiple times or creating structures in a very resource-intensive way. We feel that hierarchical clustering provides a good foundation for such a solution as long as appropriate distance measures are devised and the method is adjusted to better suit the peculiarities of transactional data. Provided solution is not resource-intensive, and resulting visualizations are clear, concise and easily interpretable by the analyst.

## 3. Solution Framework

### 3.1 Hierarchical clustering

Clustering is a well-known descriptive data mining technique which involves grouping physical or abstract objects into clusters based on their mutually shared characteristics. Objects belonging to the same cluster must share certain properties between them, while objects belonging to different clusters need to exhibit differences. Clustering is most commonly used when the analyst wishes to perceive natural groups of analyzed objects. One of the more popular clustering methods is hierarchical clustering, which enables formation of dendrograms that reveal object groupings in hierarchical manner. The usual input for clustering methods is a table where each row describes one of the objects involved in the analysis while columns represent variables used to specify properties of each object. This table is commonly called "data matrix".

Many clustering algorithms use so-called "dissimilarity matrices" which store information about distance (as opposite to closeness) between pairs of objects. If the input dataset contains $n$ objects, dissimilarity matrix will be a symmetrical $n \times n$ matrix. Dissimilarity matrix contains values $d(i,j)$ which reveal distance $d$ between objects $i$ and $j$ where $d(i,j) = d(j,i)$ and $d(i,i) = 0$. Measure $d$ is a positive number with the value being closer to zero as objects are more similar to each other.

Exact calculation of distance measure will depend on the way objects are described (or more precisely, it will depend on types of variables used to describe those objects). Variables of different types require different approaches to measuring distances between objects (as discussed in [2] and further investigated in [11]).

Variables using interval or relational scale (such as $weight$, $height$, $length$, $age$, $volume$ etc.) most commonly use either Euclidian measure or Manhattan measure. Variables may optionally be normalized before distance measure is calculated. Said measures are specializations of Minkowski distance measure shown by formula (1). This formula calculates distance between objects $i$ and $j$ represented through vectors in $p$-dimensional space. Euclidian measure assumes $q = 2$, while Manhattan uses $q = 1$.

$$d(i,j) = \sqrt[q]{\sum_{m=1}^{p} |x_{im} - x_{jm}|^q} \tag{1}$$

Another important part of forming dendrograms (when the agglomerative approach of forming progressively larger groups is used) is the choice of *linkage criterion*. In the resulting structure the nearest objects are the first to be connected. Then, in forming subsequent connections, a decision must first be made how to measure both distance between a formed group and a single element as well as distance between two formed groups. There are three most commonly used linkage criteria defined for establishing distance between already formed groups of elements (in this case $C_i$ and $C_j$ containing $n_i$ and $n_j$ elements respectively):

- minimal distance: $d_{min}(C_i, C_j) = min_{p \in C_i, p' \in C_j} d(p, p')$
- maximal distance: $d_{max}(C_i, C_j) = max_{p \in C_i, p' \in C_j} d(p, p')$
- average distance: $d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} d(p, p')$

Objects (be it single elements or formed groups) are in each following step connected with objects based on minimal value of distance (calculated using the chosen distance measure and linkage criterion), finally forming the resulting hierarchical structure. Example of the final result in dendrogram form is presented in Section 5, Figure 2.

### 3.2 Transactional data

Transactional data analysis is the chosen focus of this paper, or more specifically search and analysis of hidden relationships between transaction elements. Therefore expected input datasets will be data matrices in which rows represent transactions and columns are variables which describe transaction elements. For the purposes of analysis described in this paper the presence of transaction timestamp attribute is not required in the data matrix itself - it will be used in the data preparation stage for the analyst to select only the transactions from the time period he is interested in. Variables which are of interest represent transaction elements - references to objects that may be included in the transaction. These variables are in binary form, meaning that they denote either presence or absence of certain element in a particular transaction. While this particular format is useful for data mining, it is not the most efficient way of storing transactional data. Most commonly input datasets will originally come in different, more compact formats better suited for

saving storage space. However, transformations from such format to the one used in data mining analysis should be fairly trivial and straightforward.

Regarding relationships between transaction elements, this paper will focus on representing them through the concept of distance - in other words, the analysis partaken in provided solution will provide insight in how similar/dissimilar transaction elements are by evaluating a chosen measure of distance between them.

### 3.3 Adapting the distance measures

Depending on the goals the analyst has set up, relationships between transaction elements can be defined in various ways. The analyst might be interested in the probability of two elements appearing together in a transaction, or he may want to know how the presence of some elements affects the probability of other elements appearing in the same transaction. He can also value not just mutual presence, but also mutual absence of elements from transactions. There are various interpretations of "closeness" between elements. For this reason discussion of specific measures further in the paper will include information on how each measure affects different interpretations of relationships between transaction elements. It is important to emphasize the fact that Minkowski distance measure, being best suited for measuring interval data, isn't really the best choice for measuring distance between binary variables discussed in previous subsections. Choosing or devising new, more appropriate distance measures, might lead to results which will conform much better to the expected goals of the analyst.

Source [2] provides a few formulas potentially useful for measuring distance between objects described through binary variables. Furthermore, [4], [5], [6], [7] give an overview of measures for evaluating relevance or interestingness of a certain association rule or just its elements. These formulas and measures will provide foundations for devising new measures oriented specifically toward hierarchical clustering.

One of the main constructs used in said formulas and measures is a contingency table, an example of which is shown in Table 1.

| **i \ j** | **1** | **0** | **Sum** |
|---|---|---|---|
| **1** | $q$ | $r$ | $q+r$ |
| **0** | $s$ | $t$ | $s+t$ |
| **Sum** | $q$+s | $r+t$ | $p$ |

Table 1: Contingency table for objects $i$ and $j$ described by $p$ binary variables

Objects $i$ and $j$ are described by exactly $p$ binary variables, which indicate presence or absence of a specific property. The number of mutually present properties in both objects is given by number $q$, while the number of mutually absent properties is given by $t$. Number $r$ reflects how many properties are present in $i$ but not in $j$, and vice versa for number $s$. The sum of all these numbers must equal the total number of variables: $q + r + s + t = p$.

For the purpose of the analysis described in this paper, it is not distances between transactions we are interested in, but rather distances between transaction elements (e.g. objects the transactions refer to). Therefore contingency table will be constructed by examining the columns rather then rows of the input data matrix[1].

## 4. Adaptation of distance measures

This paper introduces new distance measures whose foundations come from the following sources:

1. Input data matrix can optionally be transposed to accentuate the fact that objects of interest are actually transaction *elements*, although in practice this step is not necessary.

- association rules measures of interestingness,
- measure for asymmetric binary variables,
- new distance measure constructed by adapting (weighing) the Yule's Q measure.

Following subsection describes chosen measures in detail, after which a process of adapting these measures so they can be used as distance measures is addressed.

### 4.1 Examined distance measures

**Support ($s$)** is one of the most important measures for discovering relationships between transaction elements. This measure denotes the probability of both elements appearing in a transaction:

$$s(A \Rightarrow B) = \frac{number\ of\ transactions\ containing\ A\ and\ B}{total\ number\ of\ transactions} \tag{2}$$

This is a symmetrical measure which means it is not sensitive to ordering of elements forming the rule. This allows us to use $s(A \Rightarrow B)$ and $s(A, B)$ interchangeably further in the text to accentuate the symmetrical nature of a measure. Support values will always be situated in the interval $[0, 1]$, with elements with larger support evaluating to values closer to 1.

**Confidence ($c$)** indicates the probability of an element appearing on the right side of the rule if left side of the rule is present in the transaction:

$$c(A \Rightarrow B) = \frac{s(A, B)}{s(A)} \tag{3}$$

Confidence also evaluates to a number contained in the interval $[0, 1]$ and gains larger values as elements are "closer" together. As opposed to support however, this is an asymmetrical measure, meaning that the order of attributes is important and that different orderings can evaluate to different values.

**Lift ($L$)** is a measure used in association rules generation which implies the strength of the relationship between two elements as opposed to those elements being together in a transaction by random chance:

$$L(A \Rightarrow B) = \frac{c(A \Rightarrow B)}{s(B)} = \frac{s(A, B)}{s(A)s(B)} \tag{4}$$

Lift is a symmetrical measure with preferred values being larger then 1. Values below 1 are undesirable and will be eliminated from further consideration (i.e. when this measure is transformed to represent distance, the distance between these elements will be set to maximum). Lift measure is constrained by the following values [17]:

$$\frac{max\left\{s(A) + s(B) - 1, 1/n\right\}}{s(A)s(B)} \leqslant L(A \Rightarrow B) \leqslant \frac{1}{max\left\{s(A), s(B)\right\}} \tag{5}$$

If we want this measure's values to fall into the $[0, 1]$ interval we can use an adapted measure called **"normalized positive lift ($NPL$)"** (we will use the term *normalization* as shorthand for "constraining to the $[0, 1]$ interval"). This measure uses a minimal relative support value given by the analyst ($min\_supp$).

$$NPL(A \Rightarrow B) = \frac{L(A \Rightarrow B) - 1}{1/min\_supp - 1} \tag{6}$$

Values closer to 1 reflect stronger connections between elements while values approaching value of 0 mean that the appearance of one elements doesn't support the reasoning that the other

element will also appear in the same transaction. NPL is symmetrical, and uses the chosen minimal support value ($min\_supp$) to normalize the lift measure.

**Added value ($AV$)** is an asymmetrical measure which evaluates how much our conclusions about the possibility of certain elements being contained in a transaction (e.g. element $B$) can improve with our knowledge of existing elements in the transaction (element $A$), again as opposed to elements we are interested in being in the transaction by pure chance:

$$AV(A \Rightarrow B) = c(A \Rightarrow B) - s(B) \tag{7}$$

Added value will rise together with the rise of the lift measure, with the ultimate limit of 1 (when $B$ almost never appears randomly in a transaction but always appears with $A$). Values below 0 aren't desirable and will also be ultimately represented by maximal distance. Again, we can adapt this value to a measure called **"normalized positive added value (NPAV)"** which will constrain this measure to the $[0, 1]$ interval:

$$NPAV(A \Rightarrow B) = max\left\{0, \frac{AV(A \Rightarrow B)}{1 - min\_supp}\right\} \tag{8}$$

**Cosine measure ($IS$)** is a measure which evaluates how interesting a combination of two elements is by taking into account both how strongly connected those elements are and how often they appear together taking the entire dataset into account. This is done by calculating a geometric average between the interestingness factor (which has the same formula as lift [6]) and support:

$$IS(A \Rightarrow B) = \sqrt{L(A,B)s(A,B)} = \sqrt{\frac{s(A,B)}{s(A)s(B)}s(A,B)} = \frac{s(A,B)}{\sqrt{s(A)s(B)}} \tag{9}$$

The name "cosine" comes from the interpretation of a measure as a scalar product of two vectors defined by attributes in a $n$-dimensional transactional space. Maximal length of such vectors is 1, which means this measure will be already normalized, falling into the [0,1] interval (value zero will be reached when the vectors are orthogonal to each other, meaning that their sets of elements have no intersection).

**Piatetsky-Shapiro measure ($PS$)** gives insight into how different the probability of two elements appearing together is from the probability of their appearance if they are considered statistically non-correlated:

$$PS(A,B) = s(A,B) - s(A)s(B) \tag{10}$$

This measure is identical to the measure called "weighted relative accuracy" [14] with a different interpretation but with the same final formula. $PS$ is a symmetrical measure limited by the value 0.25 on the positive side and achieving negative values when $A$ and $B$ are negatively correlated. Therefore we can adapt it to a **normalized positive Piatetsky-Shapiro ($NPPS$)** measure by applying the following transformation:

$$NPPS(A,B) = max\left\{0, \frac{PS(A,B)}{0.25}\right\} \tag{11}$$

**Yule's Q measure ($Q$)** is based on calculating a ratio of various probabilities, mostly combinations between presence and absence of attributes in a transaction:

$$Q(A,B) = \frac{s(A,B)s(\bar{A},\bar{B}) - s(\bar{A},B)s(\bar{A},\bar{B})}{s(A,B)s(\bar{A},\bar{B}) + s(\bar{A},B)s(\bar{A},\bar{B})} \tag{12}$$

Since this measure is contained in the $[-1, 1]$ interval we can opt to use a **normalized positive Yule's Q ($NPQ$)** by setting negative values to zero:

$$NPQ(A,B) = max\{0, Q(A,B)\} \tag{13}$$

Normalized Yule's Q measure can further be used for devising additional measure which combines the Yules's Q measure and the support measure. We can call this measure simply $Qxs$:

$$Qxs(A,B) = NPQ(A,B) \cdot s(A,B) \tag{14}$$

**Asymmetric binary difference ($ABD$)** is a measure which takes into account the possible asymmetrical nature between the importance of two states binary variable can have:

$$ABD(A,B) = \frac{s(A,\bar{B}) + s(\bar{A},B)}{s(A,B) + s(A,\bar{B}) + s(\bar{A},B)} \tag{15}$$

This measure is symmetrical and is already normalized. Furthermore, as opposed to all the above measures, this is the only measure which gets smaller values as variables are more similar, which is precisely a feature a distance measure needs to have.

The following subsection delves in detail on what are the expected properties of distance measures and which steps need to be taken to adapt the measures described above to be used for measuring distance.

### 4.2 Distance measure properties

Before describing the development of new distance measures some basic distance measure properties of binary variables and distance measures in general need to be discussed. Binary variables used to describe an object can be symmetric or asymmetric. Variables are symmetric if both states are equally important and are weighed the same by the analyst (such as a variable describing person's gender). Asymmetric variable has one state that is considered more important and interesting then another - for example a variable which describes the presence or absence of an infectious disease in a hospital patient. With transactional data it is safe to assume that binary variables describing transactions are asymmetrical since presence of an item in a transaction is usually more important to note than its absence.

It is important to emphasize the difference between the concept of asymmetry when applied to the binary variables and the concept of asymmetry when applied to a measure itself. When talking about the symmetry of distance measures, this basically means that distance must be the same regardless of which object is chosen as the reference point. In other words, if $A$ and $B$ are objects and $d$ a function of distance, than the following must be true: $d(A,B) = d(B,A)$. With some measures designed for evaluating the interestingness of association rules this will not necessarily be true, which is problematic if they were to be used as measures of distance. The easiest way to deal with this is defaulting to the smaller value of the two when using such measure for representing distance. In other words, if $d_m$ is a distance measure based on an asymmetric measure of distance $m$, then $d_m(A,B) = min(m(A,B), m(B,A))$. Justification for this can be found in the fact that two objects can be deemed as similar if distance between them is small in at least one of the directions. Further discussion and elaboration on this can be found in [6].

Finally, by examining the structure of contingency tables (which, as said, are commonly used in conjunction with binary variables), it is apparent that the values on the main diagonal are rising as two objects are more similar. However, when using distance to represent objects' similarity, smaller distance values mean more similarity between objects. In other words, distance measure needs to be inversely proportional with the objects' similarity. The measures of interestingness commonly used for association rules showcase a behavior that is exactly the opposite. To address this issue, certain measures need to be reversed so they can actually be used to measure distance. Additionally, measures of interestingness that emphasize dissimilarity rather than similarity of transactional objects should be defaulted to maximal distances.

Therefore, basic transformation steps are as follows:

1. If measure is not normalized to the [0,1] interval, use a normalized version of a measure;
2. If measure's value rises proportionally to objects' similarity, invert it (e.g. if $m$ is a measure, use $1 - m$ as its distance measure counterpart);
3. If measure is asymmetrical, calculate values for both orderings and pick the smaller.

Table 2 shows the notations and final transformation formulas for all of the derived distance measures, together with respective source measures and original ranges. Measures which are symmetrical use the "comma" notation in formulas, while asymmetrical measures use the association rule "right arrow" notation. Also, measures which do not fit to the $[0, 1]$ range use their normalized counterparts in the final formula.

| **Measure notation** | **Source measure** | **Original range** | **Final formula** |
|---|---|---|---|
| $d\_s$ | support | $0 \ldots 1$ | $1 - s(A, B)$ |
| $d\_c$ | confidence | $0 \ldots 1$ | $1 - max\{c(A \Rightarrow B), c(B \Rightarrow A)\}$ |
| $d\_L$ | lift | $0 \ldots N$ | $1 - NPL(A, B)$ |
| $d\_AV$ | added value | $-0.5 \ldots 1$ | $1 - max\{NPAV(A \Rightarrow B), NPAV(B \Rightarrow A)\}$ |
| $d\_PS$ | Piatetsky Shapiro | $-0.25 \ldots 0.25$ | $1 - NPPS(A, B)$ |
| $d\_IS$ | cosine | $0 \ldots 1$ | $1 - IS(A, B)$ |
| $d\_Q$ | Yule's Q | $-1 \ldots 1$ | $1 - NPQ(A, B)$ |
| $d\_Qxs$ | Qxs | $0 \ldots 1$ | $1 - NPQ(A, B) \cdot s(A, B)$ |
| $d\_ABD$ | Asymmetric binary difference | $0 \ldots 1$ | $ABD(A, B)$ |

Table 2: Derived distance measures

## 5. Distance measures implementation, verification and comparison

This section presents the preconstructed dataset on which the derived measures are applied. The reasonings for preconstructed dataset usage and design process are also given. After that, the results are analyzed and used for insight into measures' performance and mutual comparison.

For verification purposes we have implemented a "New Expanded Attribute Distance" widget which can be used in Orange data mining tool environment, and is based on the default "Attribute Distance Widget". "New Expanded Attribute Distance" widget, coupled with support classes which implement the algorithms for calculating derived measures and offer additional utilities for formatting input and output datasets, allowed us to test all the measures and get results in tabular form. Screenshot of scheme using this widget along with its interface is shown in fig. 1.

### 5.1 Designing the artificial dataset

Carefully preconstructed artificial data can be of immense benefit to the analysis and comparison between distance measures. Data needs to be constructed in such a way to emphasize the differences between measures as well as augment and clearly present the properties of each derived measure. In the process of choosing the most convenient measure for a particular analysis, the

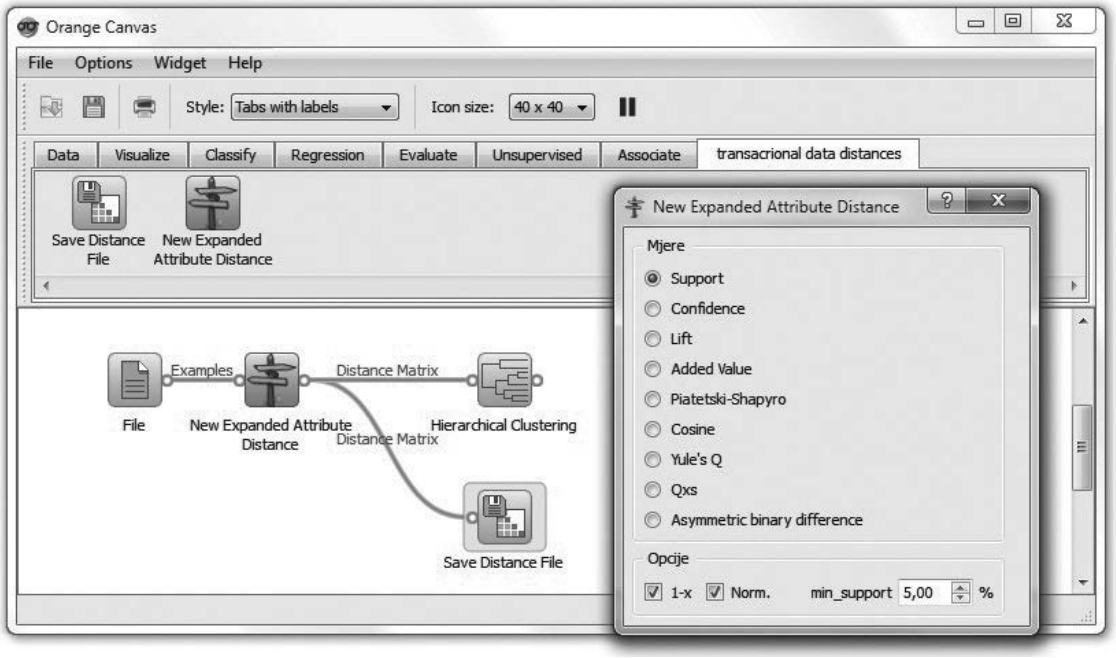



Figure 1: Orange scheme using "New Expanded Attribute Distance" widget

analyst can use the preconstructed dataset as a reference and useful guideline. Domain knowledge and specific analysis conditions will dictate which particular relationships between elements are deemed to be the most interesting, and artificial data can provide a template to help with estimating which measure conforms the best to given demands. This approach to measure selection is suggested in [6].

As said, the preconstructed dataset needs to showcase the measures' peculiarities, while at the same time it would help if it was being concise and succinct. With this in mind, the artificial data was restricted to only eight transactions. These transactions need to model as many possible relationship types between binary variables as possible. The next step involves forming attributes and their values for each transaction in such a way to cover all possible relationships.

Since relationships between pairs of elements is what interests us in the first place, it is helpful to remember that the nature of this relationships can be reflected in the contingency tables, an example of which was shown in Table 1. We need to see how many different contingency tables we can construct with only eight transactions, taking into account that each contingency table needs to showcase different view on the relationships between elements. Also, contingency tables which do not present any new or interesting information should be ignored.

As was seen in Table 1, each contingency table is basically a set of four values - $q$, $r$, $s$ and $t$ - which describe how many "matches" are between attributes in two transactional objects. $q$ describes how many attributes are mutually present in both transactions, $t$ how many are mutually absent, while $s$ and $t$ describe how many discrepancies are respectively. Summation of all these values must be equal to the number of transactions (in this case - eight), and each of these numbers can potentially evaluate to zero. Since the existence of zeros eliminates some of the combinations, the dataset will be described first in relation on how many zeros are present in the contingency table. After grouping the cases based on number of zeros, all possible combinations of values which amount to eight will be considered. Table 3 presents how many combinations are expected.

It is important to emphasize why the "two zero" combinations are so few compared to the "one zero" and "no zeros" scenarios. The only feasible "two zero" scenario is when $q = 0$ and $t = 0$, other combinations aren't useful because:

| **Two zeros present** | | **One zero** | | **No zeros** | |
|---|---|---|---|---|---|
| numbers | orderings | numbers | orderings | numbers | orderings |
| $1, 7$ | $\binom{2}{2} \cdot 2 = 2$ | 1,1,6 | $\binom{4}{3} \cdot 3!/2 = 12$ | 1,1,1,5 | 4 |
| $2, 6$ | $\binom{2}{2} \cdot 2 = 2$ | 1,2,5 | $\binom{4}{3} \cdot 3! = 24$ | 1,1,2,4 | $\binom{4}{2} \cdot 2 = 12$ |
| $3, 5$ | $\binom{2}{2} \cdot 2 = 2$ | 1,3,4 | $\binom{4}{3} \cdot 3! = 24$ | 1,1,3,3 | $\binom{4}{2} = 6$ |
| $4, 4$ | $\binom{2}{2} \cdot 1 = 1$ | 2,2,4 | $\binom{4}{3} \cdot 3!/2 = 12$ | 1,2,2,3 | $\binom{4}{2} \cdot 2 = 12$ |
| – | – | 2,3,3 | $\binom{4}{3} \cdot 3!/2 = 12$ | 2,2,2,2 | 1 |
| SUBTOTAL: | 7 | SUBTOTAL: | 84 | SUBTOTAL: | 35 |
| | | | | **TOTAL:** | 126 |

Table 3: Contingency table cell value combinations

- if $q = 0$ and $r = 0$ then $B$ is always absent
- if $q = 0$ and $s = 0$ then $A$ is always absent
- if $r = 0$ and $t = 0$ then $A$ is always present
- if $s = 0$ and $t = 0$ then $B$ is always present
- if $r = 0$ and $s = 0$ then $A$ and $B$ are identical

Table 4 shows the final dataset with eight transactions and twenty transaction attributes. This dataset covers all of the 126 proposed contingency table scenarios which was proved algorithmically. The following step is to apply developed distance measures on this dataset and compare the results. To avoid presenting each of the 126 individual results per measure, a summarized approach will be used - minimal linkage dendrograms are provided as well as graphs which allow clear comparison between measures and showcase their differences on representative number of examples.

| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |

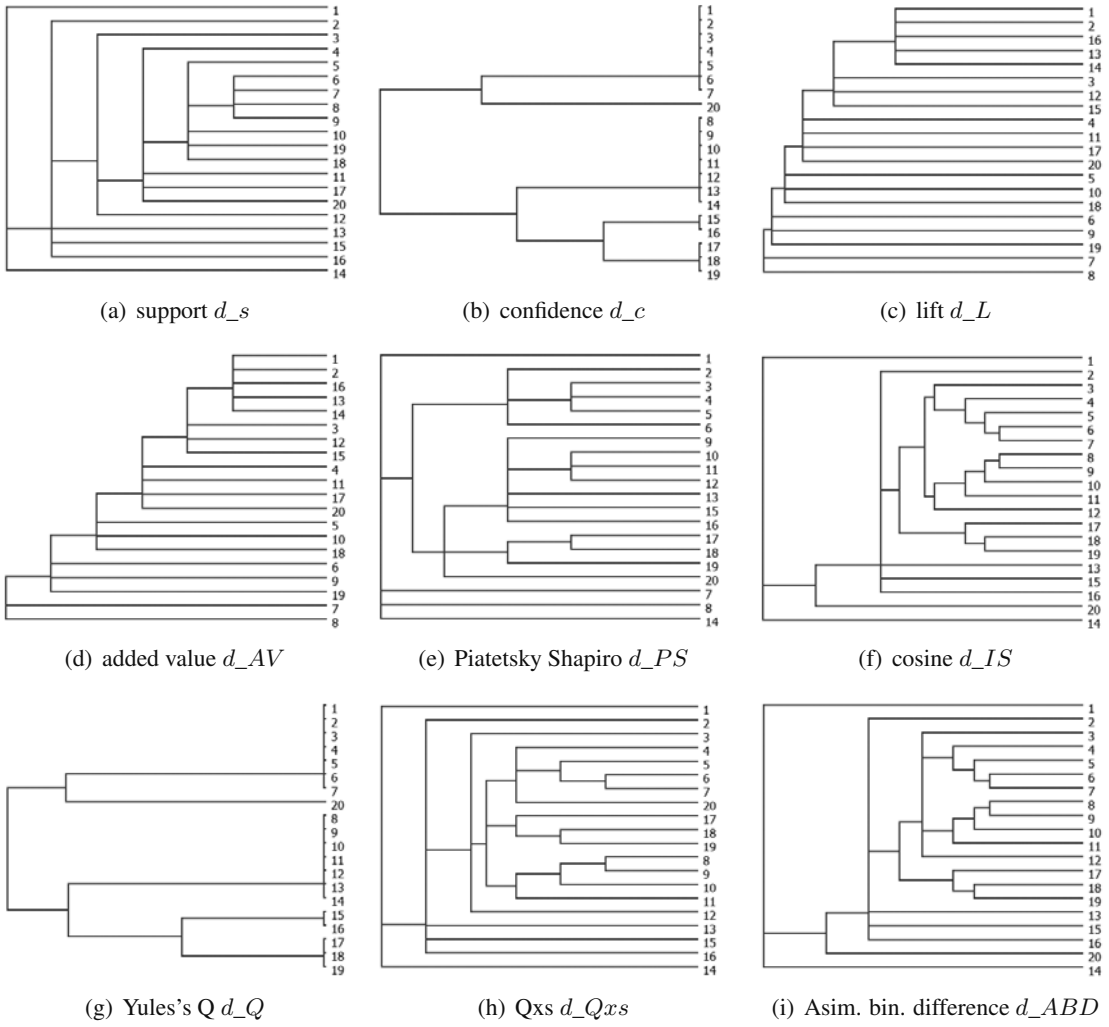
Table 4: Final preconstructed input dataset

### 5.2 Results analysis and measures comparison

All derived measures were applied on the described preconstructed dataset and results were summarized in dendrogram and graph form to allow analysis and comparison. In this subsection a representative excerpt of analysis results are shown and gained insight and conclusions about proposed measures are given. First set of results shows resulting dendrograms generated by sequentially applying hierarchical clustering method on given dataset across all of the proposed

measures. Second set presents comparison between exact measure values for chosen representative pairs of elements.

Figure 2 shows the resulting nine dendrograms. All generated dendrograms are based on minimal linkage criterion, except 2(b) and 2(g) which use average linkage criterion, since minimal linkage scenario for confidence and Yule's Q measure results in connecting all the elements on zero level.



(a) support $d\_s$ (b) confidence $d\_c$ (c) lift $d\_L$

(d) added value $d\_AV$ (e) Piatetsky Shapiro $d\_PS$ (f) cosine $d\_IS$

(g) Yules's Q $d\_Q$ (h) Qxs $d\_Qxs$ (i) Asim. bin. difference $d\_ABD$

Figure 2: Resulting dendrograms across all the measures

Dendrogram 2(a) which uses the support measure reflects the way the initial dataset was constructed - shifting the groups of 1s as can be seen in Table 4. Furthermore, it is obvious that dendrogram 2(c) belonging to the lift distance measure ($d\_L$) is very similar to dendrogram 2(d) generated by using the added value measure ($d\_AV$) - which was to be expected since their formulas clearly show that they represent similar relative relationships between elements, only connecting them on different levels. The apparent visual difference is that $d\_AV$ measure has a higher range of values, which is considered favorable in hierarchical clustering analysis. Also, lift and added value dendrograms are pretty different from other ones since they connect the most frequent elements (7, 8 and 19) only on the very last level.

Dendrograms 2(f), 2(h) and 2(i) - belonging to cosine, Qxs and asymmetrical binary difference measures respectively - also share certain similarities. Differences in connected elements start showing up in higher levels of the structure but overall appearance is pretty close, which leads to

conclusion that those measures reflect similar types of relationships between elements. Finally, confidence and Yule's Q measures (dendrograms 2(b) and 2(g)) also exhibit similar features which may infer their inherent correlation.

Since dendrograms are basically summarized representations of elements relationships where one relationship can hide other ones, another approach which provides more detailed insight into relationships may be required. As mentioned before, artificial dataset contains 126 different types of relationships between elements which is impossible to present in a complete yet concise way on a single two-dimensional graph. Therefore, a representative subset of element pairs was chosen which will reflect and represent how each measure works and which relationships it models most closely. The analyst can use these results as a guideline to which measure suits his requirements the most.

Chosen element pairs are as follows: (6,7), (5,6), (5,7), (8,11), (9,11), (4,5), (3,4), (3,5), (6,17), (6,10), (5,20), (2,16), (3,16), (4,20) and (3,10). All chosen pairs have different contingency tables, and the ordering of pairs is done by sorting them by support and ties manually reordered to achieve better clarity of presentation. Figure 3 presents a graph of measure values for described set of element pairs (lift measure is absent for reasons of clarity since it behaves almost identical to the added value measure).

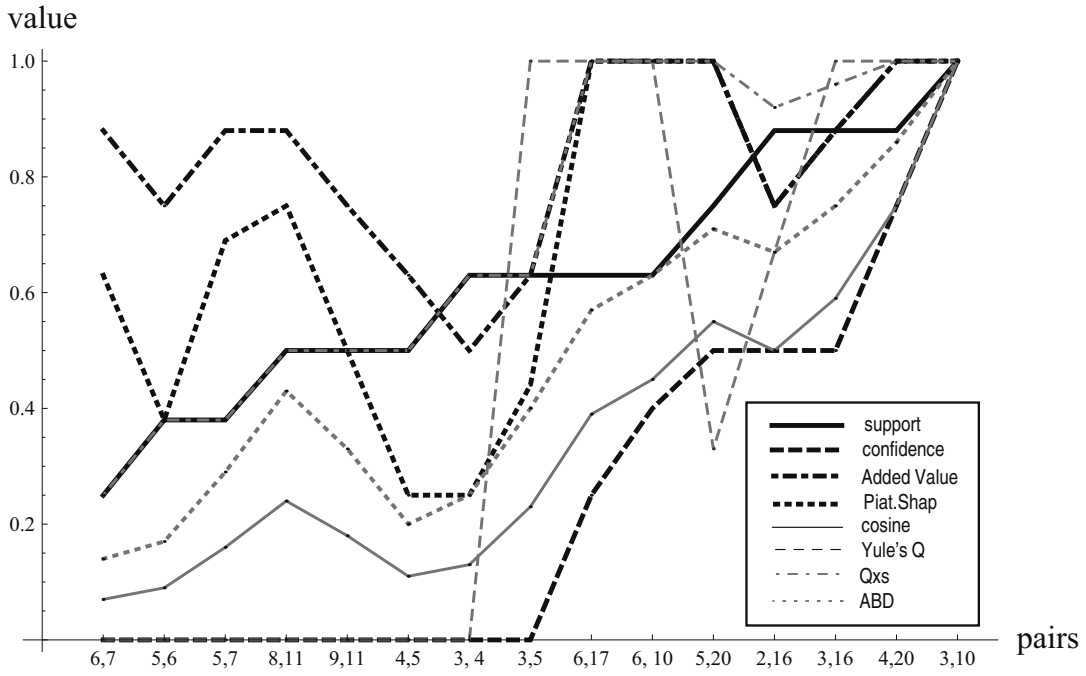



Figure 3: Measure comparison

The graph structure reveals pairs of elements on which measures act relatively similar and pairs where measures act significantly different. It is important to notice that element pairs on the left side of the graph have larger support and therefore are more interesting representations of measures' performance. Element pair (6,7) for example is the one where measure scattering is probably the most drastic, with added value reaching local maximum and cosine and asymmetrical binary difference achieving local minimums. Element pair (8,11) on the other hand shows big differences in actual measure values, but at the same time reflects similarities between measures in the sense that they almost all reach their local maximums on this "large support" portion of the graph. Elements (3,4) and (4,5) visually group the measures into two subsets, Piatetsky Shapiro and asymmetric binary difference belonging to one, and added value, support and Qxs belonging to other. These two groups evaluate to similar values which means that both model these particular

relationships in a similar way. The overall conclusion for the left portion of the graph may be that element pairs with larger support (and therefore smaller distances between them) demonstrate rather significant differences between measures.

Right side of the graphs belongs to element pairs with smaller support, and here is where all the measures act pretty similar. Sole exception is the element pair (2,16) where difference of absolute values is apparent, but derivation is consistent since almost all measures show a relative decline. It is also useful to notice that support and Qxs measures which showcased similar behavior with smaller support values now start to significantly diverge from each other.

Analysis requirements often include a minimal support value which element pairs need to surpass to be deemed interesting. If support-based-pruning is required, than the graph would have a cut-off point beyond all distances would default to "1". In this particular graph shown on fig. 3, with a given minimal support of 40% the point where all the values become "1" would be at (and including) the element pair (3,4).

If we take the remaining elements and re-order them based on added value (and resolve ties by using the Piatetsky Shapiro measure), the resulting graph would be the one shown on fig. 4. This graph lacks Yule's Q and confidence measures since they evaluate to zero for all the given pairs of elements. This graph demonstrates a concise representation of measure's behavior and offers a visually clear guideline for measure's comparison. For example, one obvious conclusion may be that cosine, Qxs and asymmetrical binary difference measures are more closely related than added value and Piatetsky Shapiro, since they consider element pair (6,7) to be much closer than what the remaining two measures evaluate to.

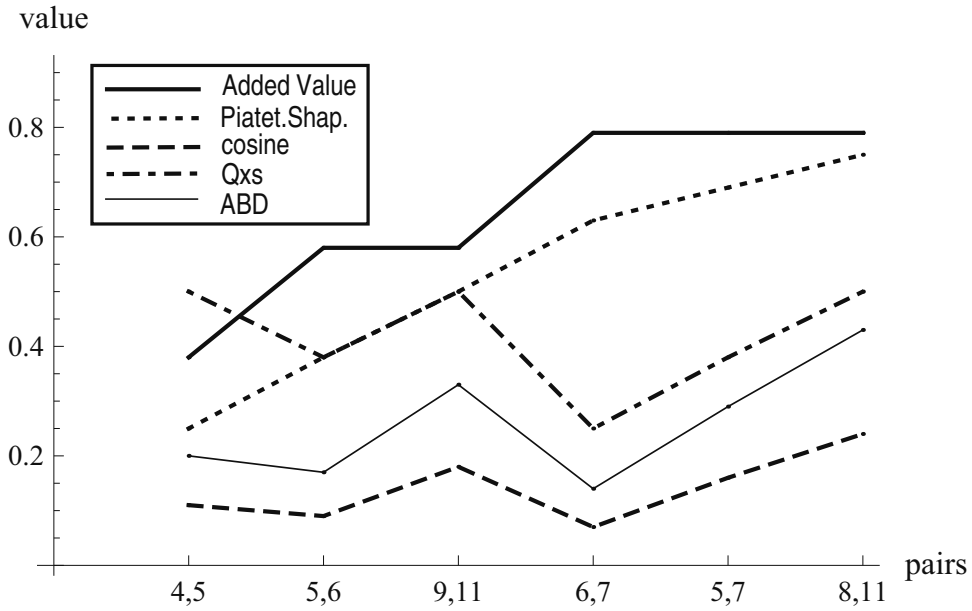



Figure 4: Measure comparison - higher support

### 5.3 Referent dataset analysis

To adequately estimate the performance and effectiveness of developed measures, in addition to applying them to the artificially constructed dataset, they should also be tested on a realistic set of data. For this purpose we used the *Computer Shop* transactional dataset (available as one of the test datasets for the *Oracle 11g Database Management System*). Some earlier research regarding tree structures formations over this dataset is presented in [16]. That research dealt with grouping of items forming frequent itemsets which result in association rules satisfying additional parameters.

Although results were very informative, developed process proved to be very resource intensive and with less control over group formation than the approach used in this paper.

*Computer Shop* dataset contains 940 transactions and 14 transaction elements. Each transaction is a record of a purchase in a shop dealing with electronic equipment. Average number of elements per transaction is 2.98, and the most frequent element appears in 32% of transactions. In this subsection we will emulate the process of analyzing dataset using developed measures, providing possible rationalizations made by the analyst for choosing particular measures and briefly discussing gained results.

The simplest and most logical measure is the *support* measure, $d\_s$, which will emphasize relationships between elements that appear most frequently together (and as such might be considered most interesting and important). Figure 5 shows a dendrogram which results from using *support* as the chosen measure of distance and minimal linkage criterion.

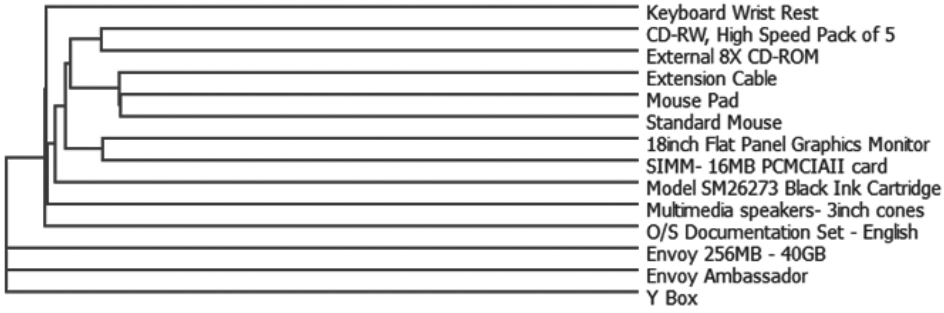



Figure 5: *Computer Shop* dataset - dendrogram gained by using the $d\_s$ measure

From this dendrogram the analyst can gain certain logical insights about this particular dataset. For example, one obvious fact is that mouses and mouse pads are very frequently bought together, with the extension cable being bundled a little bit less frequently. Graphics card is also closely tied with the monitor, as is a pack of CDs with external CD-ROMs. Elements *"Envoy Ambassador"* , *"Envoy 256MB"* and *"Y Box"* are most rarely appearing with other items.

Next choice of measure might be *confidence*, with reasoning that this measure will emphasize conditional probability. This means that by using this measure the probability of one element actually appearing in a single transaction will not be considered as important as the chance of an element appearing if it is already known some other element has appeared in the transaction.This distance measure could reveal relationships between elements that show up relatively rarely, but are closely tied together, which is commonly regarded in literature as the "*vodka and caviar problem*". Confidence measure coupled with minimal linkage criterion results in a dendrogram shown on fig. 6.

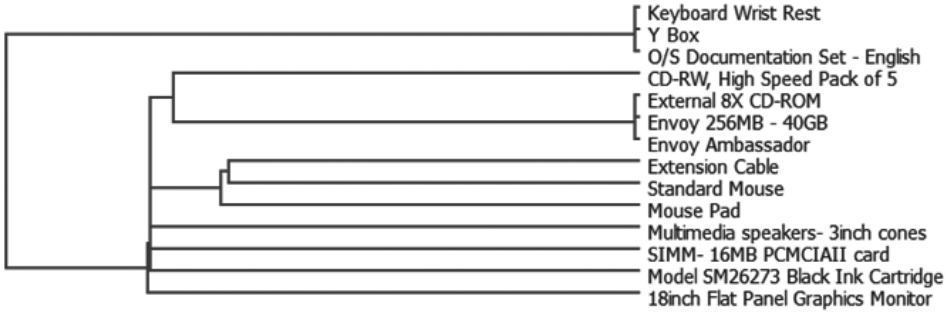



Figure 6: *Computer Shop* dataset - dendrogram gained by using the $d\_c$ measure

This dendrogram reveals additional information that elements *"Y Box" , "Keyboard Wrist Set"* and *"O/S Documentation Set"* are very closely tied together when considering confidence as a key measure of closeness. Even though these elements are relatively rarely bought, in the cases when one of them is actually chosen by the buyer there is great likelihood that one or two other elements coupled with it will also be purchased with it. Similar reasonings can be made by examining other grouped elements in the dendrogram.

For further examination analyst can reference artificial dataset described in previous chapter. If his notion of closeness between elements most closely resembles the Piatetsky Shapiro measure depicted in fig. 4, and he wants that presented relationships are supported at leased by 6% of transactions, then further analysis will lead him to the dendrogram presented in fig 7. As on previous dendrograms minimal linkage criterion was used.

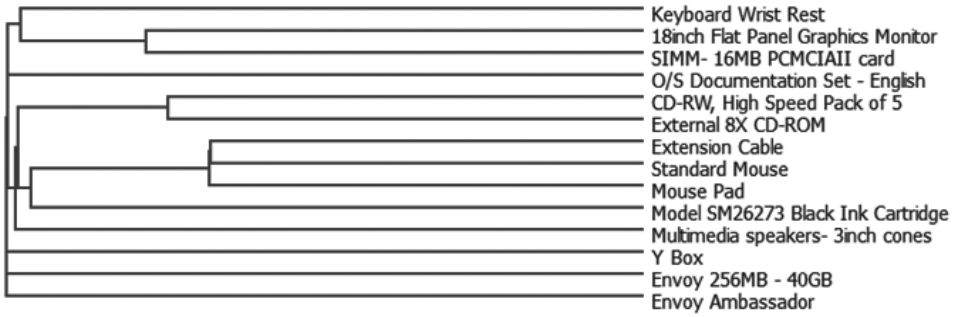



Figure 7: *Computer Shop* dataset - dendrogram gained by using the $d\_PS$ measure

This final dendrogram, in addition to presenting already discovered relationships, demonstrates a revelation of even more previously hidden relationships. For example, *"Model SM26273 Black Ink Cartridge"* and *"Keyboard Wrist Rest"* are now much closer to the dendrogram leaves than they were positioned in previous dendrograms.

By further experimentation with measure choice, linkage criteria and support based pruning, additional dendrograms will be gained which the analyst can use to get even further insight into the dataset. The ultimate effectiveness of the analysis will mostly depend on the actual needs of the analyst, his objective and subjective estimations of which relationships and which elements are deemed the most interesting, and finally his ability to interpret the results and come up with actionable decisions.

## 6. Conclusion

The focus of this paper was our research concerning concise presentation of relationships between transaction elements. Hierarchical clustering is a well known descriptive data mining method which provides results in the form of dendrograms - easily interpretable visual representations of mentioned relationships which are acceptable to the analysts with different background knowledge. However, hierarchical clustering in itself is not well explored when it comes to mining transactional data, a type of data often underutilized in decision making. The reason for the incompatibility between transactional data and hierarchical clustering is that traditional distance measures are unadapted to inherent features of transactional data. Distance measures are key components of hierarchical clustering and the choice of a distance measure is crucial for achieving a desirable and useful outcome. Therefore the focus of this paper was providing appropriate measures which would suit transactional data specifically. The goal was to come up with a set of measures which accurately model distance and as such make hierarchical clustering a good com-

plement and/or alternative to association rule generation method in the process of transactional data analysis.

Since the association rule generation method is complemented by a set of formal measures which objectively estimate the interestingness and usefulness of generated rules, we decided to use these measures as a foundation for modeling distance measures between pairs of elements in hierarchical clustering. We considered those interestingness measures, examined the general properties of the concept of "distance", and proposed transformations of those measures so their adapted form models distance more accurately. We also considered one genuine distance measure developed to model distance between objects described by asymmetrical binary attributes - Asymmetric binary difference. Transformation of considered interestingness measures led us also to devise a new distance measure called $Qxs$. All transformed measures (support, confidence, lift, added value, cosine measure, Piatetsky Shapiro measure, Yule's Q), one genuine distance measure and one newly proposed measure form the final set of nine new measures available to the analyst.

Constructing the artificial dataset on which the new measures were applied through hierarchical clustering algorithms allowed us to verify the effectiveness of measures and compare their features. The results proved that different measures act differently in various situations, which leads to conclusion that they model different aspects of relationships between elements. If the analyst chooses hierarchical clustering as his analysis method for available transactional data, he can decide on using a particular developed distance measure based on its semantics. As it can often be difficult to evaluate available measures, provided artificial dataset results can also be used as a guide to choose a measure that is the most suitable for specific analyst's needs. In this paper we also present one possible usage scenario of proposed measures on the real-world dataset and present interpretation of obtained patterns.

Future work includes extensive appliance of developed measures to various input datasets, and further research of their feasibility and effectiveness. Additionally, consulting with domain experts is planned so proper feedback is acquired about which measures are most suited to which scenarios when it comes to real-life transactional data analysis.

Finally, in regards to the comparison between hierarchical clustering and association rule generation, we have come to the following conclusions. Hierarchical clustering is a much less resource-intensive method, and the results it provides are presented in a significantly different manner. Also, the results of hierarchical clustering inevitably include a certain loss of information in the final presentation due to their conciseness, which is compensated by their clarity and ease of interpretation. With different distance measures at analyst's disposal, as well as the added option of support-based pruning and tweaking the linkage criteria, hierarchical clustering represents a powerful tool for rapid and efficient transactional data exploration and a good alternative to association rule generation.

## References

[1] Goodman, A; Kamath, C; Kumar, V. Data analysis in the 21st century. *Statistical Analysis and Data Mining*, 1(1):1–3, 2008.

[2] Han, J; Kamber, M. *Data mining: concepts and techniques*. The Morgan Kaufmann series in data management systems. Elsevier, San Francisco, 2006.

[3] Agrawal, R; Imieliński, T; Swami, A. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, SIGMOD '93, pages 207–216, New York, NY, USA, ACM, 1993.

[4] Geng, L; Hamilton, H.J. Interestingness measures for data mining: A survey. *ACM Computer Survey*, 38(3), New York, NY, USA, ACM, 2006.

[5] Piatetsky-Shapiro, G. Discovery, Analysis, and Presentation of Strong Rules. *Knowledge Discovery in Databases,* pages 229–248, AAAI/MIT Press, Cambridge, MA, 1991.

[6] Tan, P.-N; Kumar, V; Srivastava, J. Selecting the right interestingness measure for association patterns. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41, New York, NY, USA, ACM, 2002.

[7] Tan, P.-N; Kumar, V; Srivastava, J. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[8] Webb, G.I. Discovering significant rules. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 434–443, New York, NY, USA,ACM, 2006.

[9] Webb, G.I. Self-sufficient itemsets: An approach to screening potentially interesting associations between items. *ACM Transactions on Knowledge Discovery From Data*, 4:1–20, 2010.

[10] Demšar, J; Zupan, B; Leban, G; Curk, T. Orange: from experimental machine learning to interactive data mining. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD '04, pages 537–539, New York, NY, USA, Springer-Verlag New York, Inc., 2004.

[11] Pinjušić, S; Vranić, M; Pintar, D. Improvement of hierarchical clustering results by refinement of variable types and distance measures. Automatika: Journal for Control, Measurement, Electronics, Computing and Communications, 52(4):353-364, 2011.

[12] Vranić, M. *Designing concise representation of correlations among elements in transactional data*. PhD thesis, FER, Zagreb, Croatia, 2011.

[13] Berthold, M.R; Hand, D.J, editors. *Intelligent Data Analysis: An Introduction*. Springer Verlag, 2nd edition, 2003.

[14] Lavrac, N; Flach, P.A; Zupan, B. Rule evaluation measures: A unifying view. In *Proceedings of the 9th International Workshop on Inductive Logic Programming*, ILP '99, pages 174–185, London, UK, Springer-Verlag, 1999.

[15] Hofmann, H; Siebes, A.P.J.M; Wilhelm, A.F.X. Visualizing association rules with interactive mosaic plots. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '00, pages 227–235, New York, NY, USA, ACM, 2000.

[16] Vranić, M; Pintar, D; Skočir, Z. Generation and analysis of tree structures based on association rules and hierarchical clustering. In *Proceedings of the 2010 Fifth International Multi-conference on Computing in the Global Information Technology*, ICCGI '10, pages 48–53, Washington, DC, USA, IEEE Computer Society, 2010.

[17] McNicholas, P.D; Murphy, T.B; O'Regan, M. Standardising the lift of an association rule. *Comput. Stat. Data Anal.*, 52:4712–4721, June 2008.