Relationships between Distance Measures Adopted for Transactional Data Analysis

Mihaela Vranić, Damir Pintar and Zoran Skočir
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
E-mail: {mihaela.vranic, damir.pintar, zoran.skocir}@fer.hr

Abstract— **Transactional data is today's great resource of information. Unsummarized form of this type of data can reveal interesting relationships between elements of transactions. Hierarchical clustering coupled with usage of appropriate measures can reveal various aspects of these relationships. The choice of measure is a key component for getting useful analysis results. In this paper we present the continuation of our research dealing with these measures by analyzing relationships between them, better understanding of which is a great asset for analysts using them in their transactional data analysis.**

Keywords— transactional data; hierarchical clustering; distance measures; relationships between distance measures; artificial data

I. INTRODUCTION

Today transactional data occupies great storage resources in commercial sector or science databases. It is often hard to uncover the most interesting relationships between observed objects. Descriptive data mining is commonly used as a first step in data mining analysis and should offer the analyst concise and clear representation of relationships between analyzed objects [1], [2]. Transactional data, which is in focus of our research, possesses some special characteristics. Usual methods for descriptive data mining need to be adopted and customized to these characteristics.

Most commonly used method in transactional data analysis is association rules generation, firstly presented in [3]. This method however has certain disadvantages. Its main problems are:

- the final set of rules is greatly dependent on parameters ([4]),
- method execution time is hard to predict (see p. 1962. of [4] for graphs representing exponential nature of the relationship between execution time and chosen minimal support),
- relationship between two elements is presented in different, often scattered rules (this problem is addressed in many publications like [5], [6], [7], [8]).

There are a number of researches focused on finding most interesting rules. They introduce new measures for determining interestingness and relevance of rules themselves or itemsets they are based on ([6], [8], [9]). Work presented in [10] gives an overview and presents additional ways of using these measures in various scenarios.

Another descriptive data mining method is hierarchical clustering. It provides succinct, easily understandable results which allow fast and easy insight into relationships between analyzed objects. However, specific properties of transactional data often make applying this particular method somewhat difficult. Achieving good results when doing hierarchical clustering on transactional data highly depends on the choice of the appropriate distance measure, one which must be specifically adapted for transactional data itself.

Research dealing specifically with development of distance measures for hierarchical clustering of transactional elements is presented in [12]. In that paper each of developed measures are elaborated and semantics behind each measure is presented. However, even with given semantics and use case scenarios it was proven hard for analysts to pick measures most suited for their analysis. Therefore in this paper we further investigate relationships between proposed distance measures and offer additional insight into their behavior.

The paper is structured as follows: Section II gives definition of transactional data, explains the used terminology gives insight in its presentation. The related work on the subject is also more closely presented in this section. In Section III, the solution framework for analysis is introduced. Section IV provides results of the analysis and compares the distance measures. Finally, in Section V a conclusion is given.

II. DEFINITIONS AND RELATED WORK

A. Transactional Data and Known Approaches

The most frequently used example of transactional data is market basket data. It refers to data depicting transactions in the specific store – it captures data about store, time of the purchase, sometimes data about customer etc. The most important data, which is in focus of our analysis, is data about presence of specific products/services in a mutual transaction. We actually focus on relationships between this transactional elements and succinctly presentation of those relationships.

Typical transactional dataset is presented in Fig. 1. Each row here presents one transaction while columns present specific products/services being purchased – i.e. transactional elements. Here presence of specific item in a transaction is noted by "1" appearing in appropriate cell. There are some alternative presentations suitable for better storage usage or convenient for specific data mining tools. In this paper, when referring to transactional data presentation, we will refer to here displayed table presentation.

transaction\elements	A	B	C	D	…
1	1	1	0	0	
2	1	1	0	1	
3	1	1	1	0	
4	1	1	1	1	
5	0	1	1	0	
6	0	0	0	1	
...					

Fig. 1. Transactional data example.

After this example, further definitions introducing basic concepts on which the whole after going research is based should be clear.

Let I be a set of possible elements found in transaction: $I = \{I_1, I_2, ...I_n\}$. **Transaction** $t \subseteq I$ is a record of a business event modeled as a set of elements and stored in the input dataset D. **Itemset** $A \subseteq I$ is a set of elements from I used by the data mining process.

If we observe items or transactional elements as objects of our interest, than transactions play role of variables describing them. In this sense objects and variables will be referenced in the paper.

Most known and used method for transactional data analysis is association rules generation (presented in [3]). Its main disadvantages are indicated in the *Introduction section*. It is based on finding frequent itemsets (here *frequent* is determined by parameter *support*) and those itemsets are further examined to sort out the rules that comply with other parameters set by the user.

B. Hierarchical Clustering and Distance Measure Adaptation

Hierarchical clustering is descriptive data mining technique for uncovering natural groupings of objects in hierarchical manner. Obtained models are called dendrograms. Objects belonging to the same cluster (i.e. connected sooner in the dendrogram structure – near the leaves) must share certain properties between them – must be similar. Objects belonging to different clusters (or are connected afterwards in a structure – nearer to the structure root) need to exhibit some **dissimilarity**. This method is usually not applied to transactional data, however there is feasibility using this approach and initial research showcasing this can be found in [11].

Let the each object be described by p variables. For determining **dissimilarity** between objects A and B, particular formula must in a certain way take into account p values describing object A and p values describing object B. Obtained **dissimilarity** measure depicts from the certain point of view relationship between objects A and B. If we do this for every pair of observed objects (n of them) we would gain n x n matrix describing whole dataset. This matrix is called **dissimilarity matrix**. Hierarchical clustering algorithm uses this matrix to come up with the dendrogram structure. Generally speaking, dissimilarity matrix contains values $d(i, j)$ which reveal dissimilarity or 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. The choice of measure is crucial, as shown in [13].

Formula for determining dissimilarity or distance (as further referred to in the paper) is key component that must be further analyzed and adopted to transactional data. As presented in [13], distance calculation plays a major role in final organization of dendrogram. It should be picked with care and with understanding the consequences of its selection.

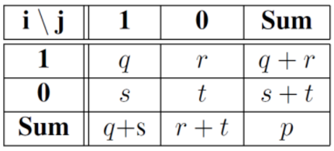
Variables using interval or relational scale (such as *weight*, *height*, *length*, *age*, *volume* etc.) most commonly use either Euclidian measure or Manhattan measure - specializations of Minkowski distance measure (formula 1).

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

However, when applying hierarchical clustering to transactional data, it must be emphasized that this measure is not appropriate since typical transactional data set contain values that comply neither to interval or relational scale. Data here is described by binary variables.

Source [2] provides a few formulas potentially useful for measuring distance between objects described through binary variables. Basis for determining distance between these objects is contingency table – presented in Fig. 2.

Here, 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: $q + r + s + t = p$.



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

Fig. 2. Contingency table for objects i and j described by p binary variables.

When it comes to application of these concepts for transactional data analysis – transactional elements present objects of our interest. If we observe typical transactional data example (Fig. 1), than specific transactions present specific properties or variables describing this objects. We can now use previously defined contingency table to calculate different distance measures for each pair of transactional objects.

Since presence and absence of elements in transactions is not equally valued, measure *Asymmetric Binary Dissimilarity* presented in [2] is used as one of measures to determine distance between transactional objects.

Work presented in [10] and [12] introduces specifically designed measures appropriate for hierarchical clustering of transactional data. Those measures are developed on basis of

TABLE I. MEASURES THAT ARE TRANSFORMED TO THE FORM SUITABLE FOR TRANSACTIONAL DATA ANALYSIS

Measure Code	Code Name	Formula	Value Range
s	support	$s(A \Rightarrow B) = s(A,B) = P(A,B)$	0···1
c	confidence	$c(A \Rightarrow B) = \frac{P(A,B)}{P(A)}$	0···1
L	lift	$L(A \Rightarrow B) = \frac{P(A,B)}{P(A)} * \frac{1}{P(B)}$	0···1···N
AV	added value	$AV(A \Rightarrow B) = P(B \mid A) - P(B)$	-0,5···0···1
PS	Piatetsky-Shapiro	$PS(A,B) = P(AB) - P(A)P(B)$	-0,25···0···0,25
IS	cosine	$IS = \sqrt{I * s(A,B)}$, $I = \frac{P(AB)}{P(A) * P(B)}$	0···$\sqrt{P(A,B)}$ ···1
α	odds ratio	$\alpha = \frac{P(AB)P(\bar{A}\bar{B})}{P(\bar{A}B)P(A\bar{B})}$	0···1···∞
Q	Yule's Q	$Q = \frac{\alpha - 1}{\alpha + 1}$	-1···0···1
Qxs	new balancing measure	$Qxs = Q * s$	-1···0···1
ABD	asymmetric binary dissimilarity	$ABR(A,B) = \\ = \frac{p(A,\bar{B}) + p(\bar{A},B)}{p(A,B) + p(\bar{A},B) + p(A,\bar{B})}$	0···1

parameters drown from association rules generation method (presented in [3]) and special measures used to evaluate relevance and interestingness of specific itemsets ([6], [8], [9]) – they mainly present different easily computed quantitative measures that should sort out most interesting rules or most interesting itemsets. TABLE I. presents all basic measures that are transformed to the form suitable for determining distances between transactional objects.

Transformation is handled in a way to keep their main semantics but it is put in perspective of distance capacity. Nine new measures are presented and some insight is given on relationship between them can also be found in [12]. Detailed elaboration on their logic and semantics accompanied with transformation to the form suitable for hierarchical clustering can be found there, and in [14] these measures are developed further by integrating them with quantitative attributes.

Among measures in TABLE I, there is also $Q\ xs$ measure. This measure is introduced in [10] and is trying to find balance between Q and s. Q treats appearing/not appearing in transaction in the same way. Still, for the analyst appearance is much more significant. This is the reason for introducing *Qxs* measure, and research presented in this document will shed more light on relationship between measures *Q*, *s*, *Qxs* among others.

C. Choosing the Right Measure for Specific Analysis

Even though the background and semantics for every developed measure is well known and easily understandable, it is sometimes hard for analysts to make a choice among them. The choice of the measure is usually paramount, and can result in dendrograms which may be similar but also very different. In that sense it is very useful for analyst to have concise comparison of measures – more precisely comparison of results they produce.

In [12], a comparison between measures has already been made, using the help of an artificially created dataset and a set of dendrograms resulting from applying each measure in turn for that particular dataset. Fixed number of transactions was agreed on in a sense that they depict every possible relationship between elements appearing in them. Final dataset (Fig. 3) included 8 transactions and 20 elements. Dendrograms for some measures were very similar – due to the algorithm forming a dendrogram and usage of minimal or average linkage criterion. It must be said that hierarchical clustering cannot show - nor is intended to show - all relationships between elements. Depending on the chosen linkage criterion it groups most similar objects in a hierarchical manner to form the final dendrogram. Its purpose is to summarize calculated distances between already formed groups, and with mentioned linkage criterions here smaller distances have greater influence on final structure. That summarization is the cause of loss of information, and though dendrograms presented in mentioned paper are very informative, they still do not offer insight in statistical relationships between offered distance measures.

To overcome this problem, authors of [12] chose the approach of choosing representative pairs of objects and presenting different calculated distance measures. In this paper,

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

Fig. 3. Artificially created transactional dataset covering all possible contingency tables involving 8 transactional elements.

however, we go a step further and take into account all pairs of elements. We again use the same dataset, however we made further transformations and filter out every reoccurring relationship to achieve appropriate input for the intended analysis.

III. Solution Framework

Ideally, to properly evaluate relationships between distance measures based on contingency tables, we would require a dataset which covers all possible contingency table variants (for a given number of variables) with each instance of each variant being taken into account only once. Artificial dataset described in Section II.C is therefore more practical then a real-life dataset, although it still requires further filtering of reoccurring contingency tables to achieve the desired pre-requisites. Subsequent analysis with larger datasets has shown that resulting relationships between measures do not change significantly and has proved that this reduced dataset is sufficient for the purposes of our inquiries.

Complete solution framework is shown on Fig. 4. For each distance measure a distance matrix was produced (using the artificial dataset as input), resulting in 9 separate distance matrices (the odds ratio measure was only used for the calculation of the Yule's Q measure and was not to be included in further research). Then, using the contingency tables for each pair of elements as a reference, we identified which pairs of elements produced a combination of contingency table values already covered by another pair of elements; these specific pairs were then flagged for removal from further analysis (for example pairs *1-2* and *1-16* result with the same contingency tables, so pair *1-16* was eliminated – which does not mean that element *1* or *16* themselves were excluded, they could still appear paired with other elements). After considering all the element combinations we ended up with a binary matrix which was used as filtering criteria for aforementioned distance matrices. This left 126 surviving pairs out of 190 possible combinations $\binom{20}{2}$.

To produce the final dataset for analysis, we put all distance matrices through the filtering criteria and then merged the filtered distances in a resulting 9 x 126 matrix (Resulting Dataset Matrix - further in text referred to as *RDM*), each column depicting each considered measure and each row belonging to a specific element pair. This matrix served us to evaluate relationships between distance measures.

For the final determination of relationships between measures we observed correlation of results in *RDM* for each pair of distance measures. This has resulted in a 9 x 9 symmetrical correlation matrix (which we will further refer to as CM). Significant values from this matrix are presented in TABLE II. As expected, some measures which were in previous experiments observed as behaving pretty similarly had demonstrated high correlation scores, while some exhibited diverging characteristics resulting in low correlation score values. This means that measures resulting in scores near zero actually favor different types of relationships between transactional elements.

One way to capture all calculated correlations and present them in a visually informative way is to use the heat map of *CM* (Fig. 5). Brighter squares show strongly correlated pairs while dark spots represent measures with a rather larger difference in behavior.

Another good way to demonstrate relations between distance measures is to use descriptive data mining. Hierarchical clustering itself is an interesting choice of method here, especially since it relies heavily on the choice of measure and it is measures themselves we are using as an object of

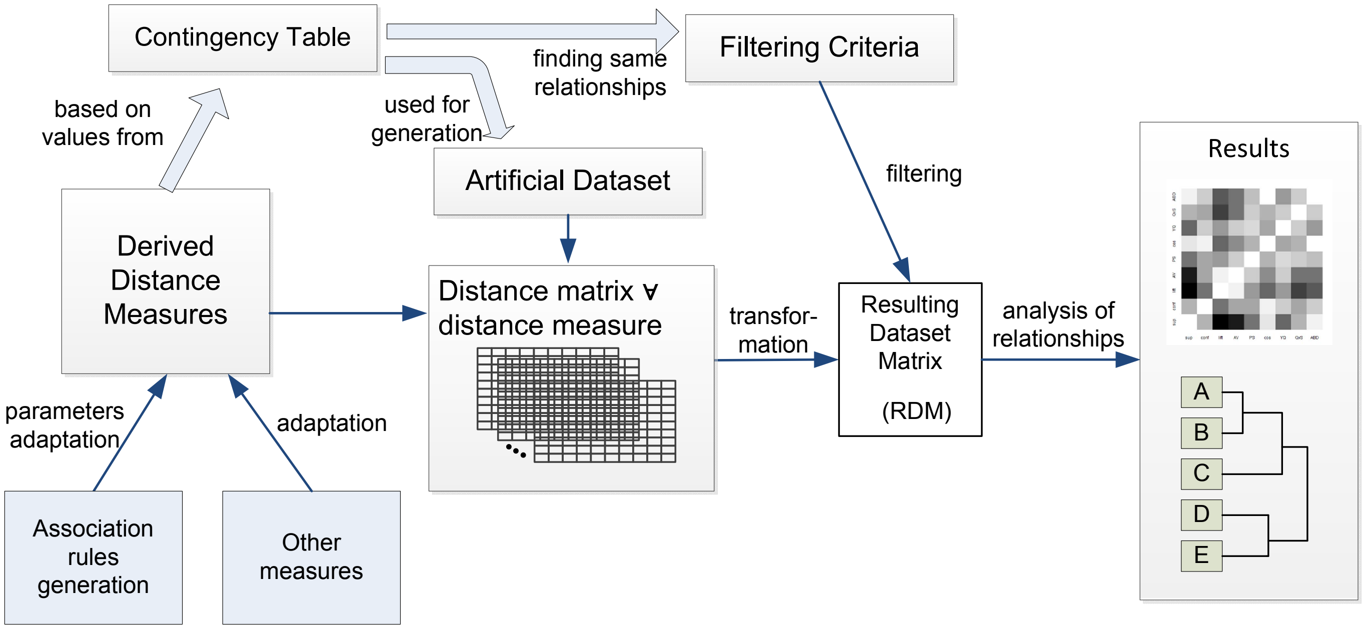


Fig. 4. Solution framework – concisely presented analysis flow.

TABLE II. CORRELATION VALUES BETWEEN ALL PAIRS OF MEASURES

	conf	lift	AV	PS	cos	YQ	QxS	ABD
sup	0.71	0.03	0.16	0.48	0.89	0.42	0.68	0.92
conf		0.48	0.65	0.67	0.92	0.80	0.66	0.81
lift			0.92	0.62	0.41	0.61	0.28	0.36
AV				0.81	0.54	0.81	0.49	0.49
PS					0.71	0.83	0.81	0.75
cos						0.68	0.72	0.97
YQ							0.81	0.62
QxS								0.77

interest for this particular analysis. It was fitting to discover that for the hierarchical clustering of measures specifically devised to cluster transactional elements, the default Euclidian measure proved as an adequate choice. Hierarchical clustering was applied using RDM input, and the resulting dendrogram is shown on Fig. 6.

IV. COMPARISON OF DISTANCE MEASURES

This section will explain and summarize gathered results obtained through processes described previously.

As expected certain measures behave pretty similarly, such as *lift* and *added value*. Correlation value between these measures was 0.92 (0 being the lowest and 1 being the highest possible score) and hierarchical clustering has group these measures very early on. This was an expected result since both measures are based on similar semantics and their formulas (shown previously in TABLE I) are a ratio of probabilities and a difference in probabilities, both decreasing with increased probability of the second pair element. If the choice between distance measures comes down to lift vs. added value, then the analyst should opt for the added value measures, since as shown they have similar semantics and behave in a similar fashion yet added value usually offers somewhat visually more acceptable dendrograms.

One unexpected result of our analysis was similar behavior of the *cosine* measure and *ABD* measure because their underlying semantics are rather different.

Another interesting observation is taking into account resulting correlations between the *support* measure, *Yule's Q* measure and the *QxS* measure, which was basically an amalgamation between the first two measures devised to more strongly accentuate a presence of an element in a transaction as opposed to its absence. As far as correlation goes, *support* and *Yule's Q* were shown to be very different, and *QxS* behaved exactly as expected – as a balance of sorts, being loosely correlated with both *support* and the *Yule's Q* yet holding its own as a separate measure with its own separate semantics.

Another interesting result of this analysis is demonstrating how using two different methods of analysis – hierarchical clustering and correlation measures – in many aspects offer similar insights (for example showing previously stated conclusions that *lift* and *added value* are similar, or *support* and *ABD*, or – which wasn't mentioned yet – *cosine* and *confidence* measure) yet in certain details the results vary. For example, from the correlation matrix one my conclude that *Piatetsky Shapiro* and *Yule's Q* are more similar to each other than *Piatetsky Shapiro* and *QxS* measure, yet the hierarchical clustering grouped *Piatetsky* and *QxS* together while leaving the Yule's Q measure ungrouped until very late in the hierarchy. This results obviously means that further research is required when it comes to relationships between these three measures.

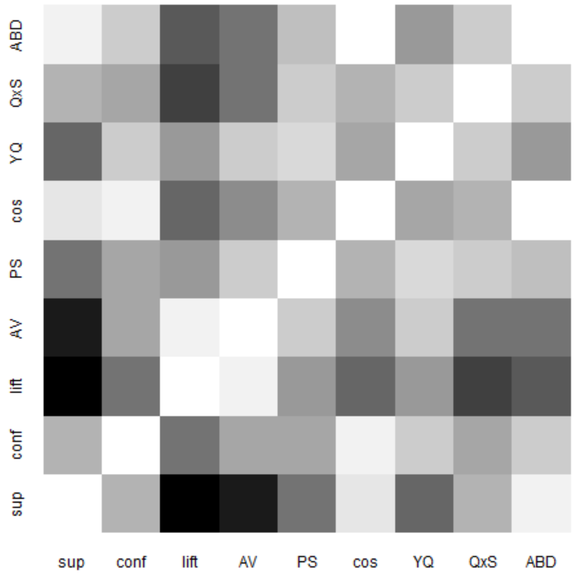


Fig. 5. Heat map of correlations between results produced by different distance measures

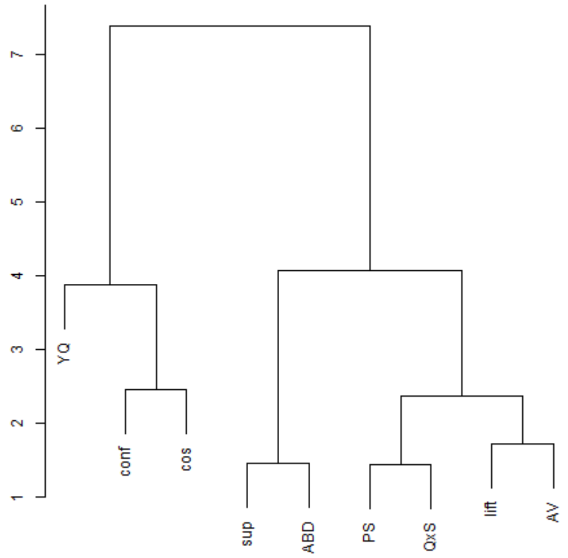


Fig. 6. Hierarchical clustering of analyzed distance measures on the basis of results produced by them

V. CONLUSION

In this paper we present our further investigation of distance measures devised for the usage of hierarchical clustering method on transactional data. Direct motivation for this was observed difficulties in choosing which measure (out of nine) would be the most appropriate for a particular analysis. Even though the formulas and semantics behind measures are the expected default guideline for the analyst, additional insight in their behavior and their relationships is an often asked feature.

In this paper we used an artificially created dataset which ensured all possible combinations of relationships between a certain number of transactional elements are covered. We then identified uniquely presented relationships between transactional elements and used them to find our object of interest for this paper - relationships between derived measures themselves. For this we used two methods – calculation of a correlation matrix for each pair of measure results which evaluated the similarity between two measures through a numeric value between 0 and 1, and a hierarchical clustering method used to group the measures themselves into clusters.

Results of the research in certain aspects solidified our already formed hypotheses about how some measures behave and which exhibit similar characteristics. Some of the more interesting conclusions included the measure *QxS*, a measure specifically developed to combine two different semantic approaches to emphasize an aspect deemed important in transactional analysis – specifically the larger importance of an element being present in a transaction as opposed to it being absent from it.

Regardless of what the results of the research in this paper show, we feel that a very important factor in deciding which distance measure is the most appropriate for a particular dataset needs to be the actual semantics behind chosen measure. Research shown here represents a useful guideline and should serve to offer additional insight into expected behavior of the measures, yet it is still the semantics of the measure which will affect which exact relationships are emphasized in a particular analysis. In another words, if a measure provides results which are attractive to the analyst, he/she should still check the semantics behind the measure to see whether regardless of the satisfying results the measure actually fits the analysis in question.

REFERENCES

[1] Goodman, A; Kamath, C; Kumar, V. Data analysis in the 21st century. Statistical Analysis and Data Mining, 1(1), pp.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] R. Agrawal, T. Imielinski, and A. Swami, “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, pp. 207-216, New York, NY, USA, 1993

[4] Vranić, Mihaela; Pintar, Damir; Banek, Marko. The impact of threshold parameters in transactional data analysis // Proceedings of the the 35th International ICT Convention – MIPRO 2012 / Biljanović, Petar, editor(s). Rijeka : Grafik, 2012. 1959-1964 (lecture,international peer-review,published,scientific).

[5] K. Gouda and M. J. Zaki. Genmax, “An efficient algorithm for mining maximal frequent itemsets”, Data Min. Knowl. Discov., pages 223-242, 2005

[6] P.-N. Tan, V. Kumar, and J. Srivastava, “Selecting the right objective measure for association analysis”, Inf. Syst., 29:293-313, June 2004

[7] M. J. Zaki, “Generating non-redundant association rules”, In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 2000

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

[9] 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.

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

[11] M. Vranić, D. Pintar, and Z. Skočir, “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, pp. 48-53, IEEE Computer Society, Washington, DC, USA, 2010

[12] M. Vranić, D. Pintar, and D. Gamberger, “Adapting hierarchical clustering distance measures for improved presentation of relationships between transaction elements”, Journal of Information and Organizational Sciences, vol. 36, No. 1, pp. 69-86, Varaždin, Croatia, 2012.,

[13] 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.

[14] Vranić, Mihaela; Pintar, Damir; Skočir, Zoran. Integrating quantitative attributes in hierarchical clustering of transactional data. Lecture Notes in Artificial Intelligence. 7327 (2012) ; 94-103