ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set

Marie Plasse[a, b, *], Ndeye Niang[a], Gilbert Saporta[a], Alexandre Villeminot[b], Laurent Leblond[b]

[a] *CNAM, Laboratoire CEDRIC, 292 Rue St Martin Case 441, 75141 Paris Cedex 03, France*
[b] *PSA Peugeot Citroën, Zone Aéronautique Louis Bréguet, Route Militaire Nord, 78943 Vélizy Villacoublay, France*

Abstract

A method to analyse links between binary attributes in a large sparse data set is proposed. Initially the variables are clustered to obtain homogeneous clusters of attributes. Association rules are then mined in each cluster. A graphical comparison of some rule relevancy indexes is presented. It is used to extract best rules depending on the application concerned. The proposed methodology is illustrated by an industrial application from the automotive industry with more than 80 000 vehicles each described by more than 3000 rare attributes.

Keywords: Association rules mining; Variable clustering; Large sparse matrix; Binary attributes; Rule relevancy index

0. Introduction

We consider the problem of discovering links between binary attributes in the case of large sparse matrices. Our sample data from the automotive industry consists of more than 80 000 vehicles each described by more than 3000 attributes. Each attribute is a binary variable equal to 1 if the vehicle has the attribute, 0 otherwise.

Our data can be considered as basket data and then a first idea is to mine association rules to find frequent co-occurrences of attributes. In our case, threshold configuration for support and confidence is particularly tricky. Minimum support has to be very low because vehicle attributes are extremely rare contrary to basket data. In addition, by a slight threshold variation, the number of rules increases rapidly.

To solve this problem we propose to cluster variables in order to build homogeneous groups of attributes and then mine association rules inside each of these groups. We have used several clustering methods and compared resulting partitions. The study shows that the combined use of association rules and classification methods is more relevant. Actually this approach brings about an important decrease in the number of rules produced. Furthermore, it appears that complex rules are always generated by the same grouped attributes identified through variable clustering.

* Corresponding author. CNAM, Laboratoire CEDRIC, 292 Rue St Martin Case 441, 75141 Paris Cedex 03, France. Tel.: +33 1 57 59 08 15; fax: +33 1 41 36 30 46.
E-mail address: marie.plasse@mpsa.com (M. Plasse).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.02.020

Even if we reduce number of rules, we still need to sort them from the most relevant to the less interesting. There are many indexes that measure statistical interest of association rules and the choice of one depends on the application. We have performed an empirical and graphical comparison to help to select the most appropriate.

After reviewing the basics of association rules mining in the first section, in the second we present an overview of variable clustering methods. Then in the third section we describe the combined use of these two methods. Finally, in the last section we compare some rule relevancy indexes. To illustrate our approach, each section contains a detailed example using industrial data.

1. Association rules mining

1.1. Algorithms to mine association rules

Association rules mining has been developed to analyse basket data in a marketing environment. Input data are composed of transactions: each transaction consists of items purchased by a consumer during a single visit. Output data are composed of rules. For example, a rule can be "90% of transactions that involve the purchase of bread and butter also include milk" (Agrawal et al., 1993). Even if this method has been introduced in the context of Market Business Analysis, it has many applications in other fields, like webmining or textmining. It can actually be used to search for frequent co-occurrences in every large data set.

A rule is an implication $A \rightarrow C$. The left part of the rule is called the antecedent and the right, the consequent. The sets A and C are disjointed as we cannot find the same item in both the antecedent and consequent. A rule makes sense thanks to its support $s = sup(A \rightarrow C) = P(A \cap C)$ and its confidence $c = conf(A \rightarrow C) = P(C/A)$.

The first efficient algorithm to mine association rules is *APriori* (Agrawal and Srikant, 1994). The first step of this algorithm is the research of frequent itemsets. The user gives a minimum threshold for the support and the algorithm searches all itemsets that appear with a support greater than this threshold. The second step is to build rules from itemsets found in the first step. The algorithm computes confidence of each rule and keeps only those where confidence is greater than a threshold defined by the user.

As we will see in the application, one of the main problems is to define support and confidence thresholds. *Apriori* is based on the property that every subset of a frequent itemset is also frequent. Candidate k-itemsets are generate in the kth read of the data set and their supports are computed in the $k+1$th read. If K is the largest size of frequent itemsets, the total number of reads is $K+1$. Other algorithms have been proposed to decrease the count of reads of the database and to improve computational efficiency. Among them, we can quote *Eclat* (Zaki, 2000), *Partition* (Savasere et al., 1995), *Sampling* (Toivonen, 1996), *DIC* (Brin et al., 1997a) or *FP-Growth* (Han et al., 2000). All of these algorithms furnish the same results as rules searching is deterministic. We have used *Apriori* and *Eclat* because they perform fastest on our sparse data.

1.2. Application of association rules mining to industrial data

To apply association rules mining, we consider vehicles as transactions and their binary attributes, as items.

1.2.1. A large sparse data set of industrial data

The sample data from the automotive industry consists of more than 80 000 vehicles each described by more than 3000 binary attributes. Simple graphics illustrate that we are dealing with extremely sparse data. Our binary matrix contains about 0.13% of "1". The most frequent attribute appears on 12% of vehicles but 97% of attributes appear on less than 1% of vehicles as shown in Fig. 1. In addition, a vehicle has an average of four attributes. A few vehicles have more than 10 attributes but most have between one and five (Fig. 2).

1.2.2. Too many association rules

With a minimum support of 500 vehicles and a minimum confidence of 50%, *the algorithms* produce 18 rules with a maximum size of three items (Table 1). As items are rare events, the minimum support threshold has to be reduced to identify less common links. We want to find rules with a minimum support of 100 vehicles. As Table 1 shows, the number and complexity of rules increase enormously.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

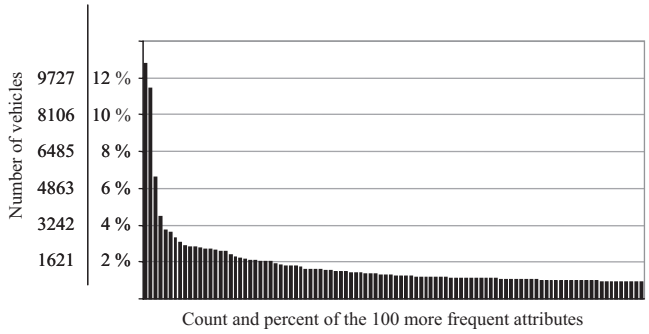

Fig. 1. Distribution of the 100 most frequent attributes.

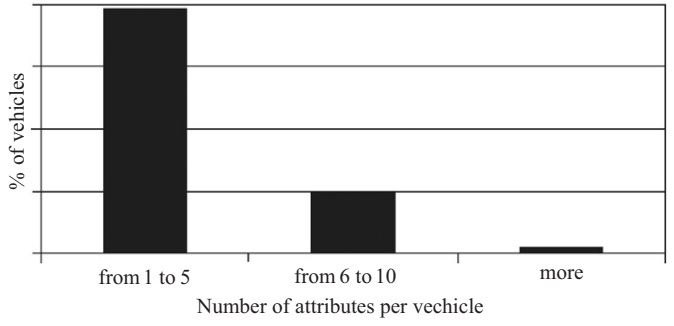

Fig. 2. Attributes per vehicle.

Table 1
Number of frequent itemsets and rules with different support thresholds

Minimum support threshold	Minimum confidence threshold (%)	Number of frequent itemsets	Number of association rules	Maximum complexity (size) of rules
500 vehicles	50	188	18	3
400 vehicles	50	256	31	3
300 vehicles	50	389	213	5
200 vehicles	50	2398	86 836	9
100 vehicles	50	7704	600 632	11

Table 2
Number of rules with different confidence thresholds

Minimum confidence (%)	Number of rules
50	600 632
80	416 312
90	240 362
99	60 841

As each rule has to be analysed and eventually confirmed by an expert, we clearly have too many rules to be practical. A first attempt to reduce the number of rules by dropping the minimum confidence threshold (Table 2) shows that even if we are more selective the reduction is insufficient.

In our second attempt to reduce the number of rules, we keep the more interesting ones by sorting them by descending order of relevancy. The relevancy of a rule is given by a measure of its statistical interest. As we will see in Section 4,

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

there are many such measures. One of the most well-known is the lift (Brin et al., 1997b):

$$lift(A \Rightarrow C) = \frac{P(A \cap C)}{P(A).P(C)}.$$

Generated rules present great values of lift so it is impossible to make a significant selection of rules in this way.

As these two attempts to reduce the number of rules have failed, we have therefore proposed a new way by combining association rules mining and variable clustering. Before explaining this new approach, we present variable clustering methods in the next section.

2. Variable clustering methods

2.1. Overview

Variable clustering may be useful for finding multicolinearities between variables, for reducing a too high quantity of variables, and for converting variables into independent dimensions. In our case, we aim to build homogenous groups of correlated variables.

As with individual clustering, there are two families of variable clustering methods: hierarchical and non-hierarchical. Hierarchical methods are divided in two groups:

- Ascendant methods based on a agglomerative algorithm.
- Descendant methods performed by a divisive algorithm.

Ascendant hierarchical clustering leads to a hierarchy of nested clusters and is based on the choice of a similarity coefficient and an aggregation strategy. The similarity coefficient allows us to appraise the degree of proximity between two variables. Many coefficients have been proposed. A good overview has been made by Nakache and Confais (2005). The choice of a coefficient depends on the nature of both the variables and application concerned. There are many similarity coefficients for binary data (Fichet and Le Calvé, 1984), all computed from the contingency table of variables X_j and $X_{j'}$ (Fig. 3).

The Pearson's Φ^2 is obtained by the Khi^2 contingency measure:

$$\Phi^2_{jj'} = \frac{\chi^2_{jj'}}{n} = \frac{n_{11}n_{00} - n_{01}n_{10}}{n_{1.}n_{0.}n_{.1}n_{.0}},$$

where $\Phi^2_{jj'}$ is equal to $r^2_{jj'}$, the squared linear correlation coefficient between the indicator variables of the first category. We can also use the linear correlation coefficient. Other common coefficients are summarised in Table 3.

We can notice that the Russel–Rao coefficient is identical to the support of the itemset $\{X_j X_{j'}\}$:

$$s_{RR}(jj') = \frac{n_{11}}{n} = \frac{Number\ of\ co\text{-}occurrences}{Number\ of\ cases} = support\{X_j; X_{j'}\}.$$

Other coefficients may be used, such as the Qannari and Vigneau coefficient (Qannari et al., 1998) which is based on Escoufier's operator. In the case of two variables X_j and $X_{j'}$ with m and q categories, it is obtained by the

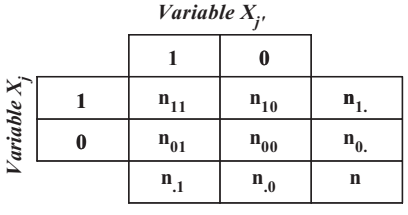

Fig. 3. Contingency table of variables X_j and $X_{j'}$.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Table 3
Similarity coefficients used in our case

Russel and Rao	$s_{RR}(jj') = \frac{n_{11}}{n_{11}+n_{10}+n_{01}+n_{00}} = \frac{n_{11}}{n}$
Jaccard	$s_J(jj') = \frac{n_{11}}{n_{11}+n_{10}+n_{01}} = \frac{n_{11}}{n-n_{10}} = \frac{n_{11}}{n_{1.}+n_{01}}$
Ochiai	$s_O(jj') = \frac{n_{11}}{\sqrt{(n_{11}+n_{10})(n_{11}+n_{01})}} = \frac{n_{11}}{\sqrt{n_{1.}n_{1.}}}$
Dice	$s_D(jj') = \frac{2n_{11}}{2n_{11}+n_{10}+n_{01}}$

Tschuprow T:

$$D^2(X_j, X_{j'}) = 2(1 - T^2(X_j, X_{j'})),$$

with $T^2(X_j, X_{j'}) = \chi^2_{jj'}/n\{(m-1)(q-1)\}^{1/2} = \Phi^2_{jj'}$ in the binary case.

The Matusita affinity (Nicolau and Bacelar-Nicolau, 2003) can also be used:

$$a_{jj'} = \sum_{i=1}^{n} \frac{n_{ij}}{n_{.j}} \frac{n_{ij'}}{n_{.j'}},$$

where n_{ij} is the number of occurrences of the couple (*individual* i, *variable* j) and $n_{.j}$ the margin of column j. In the binary case, the Matusita coefficient is the same as Ochiai's coefficient.

Hierarchical classification requires changing these similarity coefficients $s(j, j')$ into dissimilarity coefficients $d(j, j') = \max_{j,j'}(s(j, j')) - s(j, j')$.

Aggregation strategy allows us to build the hierarchy of nested clusters. We use the same strategies as those used for individual clustering: Ward criterion, minimum linkage, maximum linkage or average linkage. The Ward criterion is employed only with euclidian distances. Fichet and Le Calvé (1984), then Gower and Legendre (1986) have shown that coefficients like Jaccard are euclidian distances.

Descendant methods provide hierarchical trees where terminal segments represent a partition of variables. Variables in the same cluster have the maximum possible correlation and variables in different clusters have the minimum. The algorithm searches one-dimensional clusters, described by only one principal component. At the beginning, all variables are in the same cluster. A principal component analysis is performed on the entire set of variables. If the second greater eigenvalue is greater than one, the set of variables is divided into two clusters. Each variable is assigned to the component with which it has the higher squared correlation. From then on, the procedure is iterative. A principal component analysis is performed on each cluster, if the second greater eigenvalue is greater than one, the cluster is divided and if not, the division stops. The *Varclus* procedure (SAS, 2003) allows this kind of clustering.

The second mentioned family of methods is non-hierarchical or partitional. These methods search directly for a specific partition of variables. The number of clusters must be known. Vigneau and Qannari (2003) propose a method which searches K clusters $G_1, G_2, \ldots, G_K$ and K corresponding latent variables $c_1, c_2, \ldots, c_K$. The aim is to maximise the correlation between a variable in a cluster and the corresponding latent variable. This method is sometimes presented as a variant of the *Varclus* procedure due to its maximising criterion, but it is in fact a non-hierarchical method and its iterative procedure appears to be the algorithm k-means (Forgy, 1965).

2.2. Applying clustering methods to industrial data

Whatever clustering method is used, we need to decide how many clusters to use. As a rule, hierarchical trees can help suggest the appropriate number of clusters. Here, the large number of variables makes the hierarchical trees unreadable.

Within all tests performed with different numbers of clusters, there is always a large cluster that contains a high proportion of variables. This is illustrated in Tables 4 and 5 which show the number of variables in the five largest clusters, from 10 to 100. These results have been obtained through ascendant hierarchical clustering using the Ward strategy together with the squared correlation coefficient (Table 4) and the Russel–Rao coefficient (Table 5). Table 6

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Table 4
Composition of biggest clusters with ward strategy, R^2

Ward/R^2	Number of variables in the five largest clusters				
Total number of clusters	●	●	●	●	●
10	3058	10	6	5	4
20	3013	14	10	6	6
30	2983	14	10	6	6
50	2938	14	10	6	6
100	2798	10	9	7	7

Table 5
Composition of biggest clusters with ward strategy, Russel–Rao index

Ward/Russel–Rao	Number of variables in the five largest clusters				
Total number of clusters	●	●	●	●	●
10	2928	117	16	12	10
20	2895	55	38	28	16
30	2886	46	28	22	16
50	2675	167	26	26	13
100	2542	90	35	22	21

Table 6
Composition of clusters obtained by the Varclus procedure

Cluster	1	2	3	4	5	6	7	8	9	10
Number of variables	1282	1001	349	156	111	61	41	12	60	28

points out the number of variables in a 10-cluster partition generated by the *Varclus* procedure. We see that the clusters are more balanced.

As far as the choice of an aggregation strategy in the case of ascendant hierarchical strategy goes, the most interesting partitions have been obtained by the use of the Ward strategy. Fig. 4 illustrates that average and minimum linkages provide exactly the same partition, i.e. one variable by cluster except for one cluster containing all of the others. This is not compatible with our objective of making groups of variables. Maximum linkage gives the same kind of results except when the R^2 is used. As a result we conclude that the Ward strategy offers the best results whatever the similarity coefficient.

The Rand index measures agreement between two partitions of the same data (Youness and Saporta, 2004). It gives the percentage of pairs in agreement. A pair in agreement is two variables that are clustered together in two partitions or two variables that are clustered in different clusters within the same two partitions. Table 7 shows that all partitions from agglomerative clustering procedure are very close to each other as the Rand coefficient is close to 0.8. On the other hand, partition from divisive clustering procedure is quite distant from the others but we could have expected that because this partition has more regular clusters.

3. Association rules mining with primary variable clustering

We have used several clustering methods to get groups of correlated variables. Once we have obtained a partition of the variable set, association rules mining is performed inside each cluster as illustrated in Fig. 5 (Plasse et al., 2005a,b).

As we have shown in the previous paragraph, the Ward strategy produces interesting clusters. In other respects, the choice of the Russel–Rao coefficient is relevant because of its link with the support used to extract association

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Cluster	R²	Jaccard	Russel Rao	Dice	Ochiai
1	3092	3092	3092	3092	3092
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1
10	1	1	1	1	1

Average linkage strategy

Cluster	R²	Jaccard	Russel Rao	Dice	Ochiai
1	1378	3086	3084	3086	3083
2	837	7	9	7	10
3	295	1	1	1	1
4	232	1	1	1	1
5	160	1	1	1	1
6	143	1	1	1	1
7	53	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1
10	1	1	1	1	1

Maximum linkage strategy

Cluster	R²	Jaccard	Russel Rao	Dice	Ochiai
1	3092	3092	3092	3092	3092
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1
10	1	1	1	1	1

Minimum linkage strategy

Cluster	R²	Jaccard	Russel Rao	Dice	Ochiai
1	4	2973	2	2690	2762
2	3	72	12	61	201
3	3058	5	2	5	5
4	3	6	5	7	6
5	4	4	117	298	11
6	6	10	4	11	84
7	4	5	10	5	5
8	10	8	5	6	8
9	5	12	16	12	13
10	4	6	2928	6	6

Ward criterion

Fig. 4. Comparison of aggregation strategies.

Table 7
Paired comparison of partitions provided by the Rand index

	Ascendant clustering (with Ward criterion)				
	Coefficients	R^2	Ochiai	Jaccard	Dice	Russel–Rao
Ascendant clustering (with Ward criterion)	Ochiai	0.82				
	Jaccard	0.94	0.87			
	Dice	0.78	0.79	0.82		
	Russel–Rao	0.87	0.8	0.84	0.86	
Descendant clustering	(Varclus)	0.31	0.39	0.34	0.41	0.35

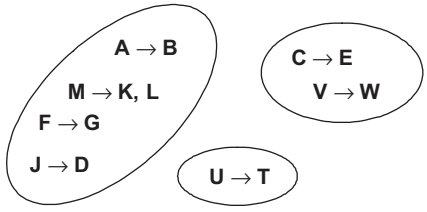

Fig. 5. Association rules mining inside each cluster.

rules. Here, we present rules found from the partition generated by the Ward strategy with the Russel–Rao coefficient. Table 8 shows the number of frequent itemsets with a minimum support threshold of 100 vehicles and the number of rules and their maximum size with different levels of the minimum confidence.

Cluster 2 brings together 12 attributes that are extremely correlated. Moreover, these attributes are gathered regardless of the clustering method used and correlations between them produce a high number of complicated association rules.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Table 8
Number of rules with different confidence thresholds after clustering

Clusters	Number of frequent itemsets	Minimum confidence = 99%		Minimum confidence =90%		Minimum confidence =80%		Minimum confidence = 50%	
		Number of rules	Maximum complexity	Number of rules	Maximum complexity	Number of rules	Maximum complexity	Number of rules	Maximum complexity
1	3	0	0	0	0	0	0	1	2
2	4095	33 233	11	171 229	11	315 134	11	481 169	11
3	3	0	0	0	0	0	0	1	2
4	24	3	4	14	4	15	4	24	4
5	173	4	4	12	4	24	4	53	4
6	15	1	4	9	4	14	4	22	4
7	57	0	0	3	4	7	4	33	4
8	17	0	0	4	4	11	4	22	4
9	59	0	0	1	2	1	2	1	2
10	421	8	4	17	4	22	4	60	4
All clusters	4867	33 249	11	171 289	11	315 228	11	481 386	11
All except cluster No. 2	772	16	4	60	4	94	4	217	4

Table 9
Reduction of number and size of rules

		Number of rules	Maximum size of rules	Reduction of number of rules
Without previous clustering		600 636	12	—
Ascendant clustering with	R^2 index	43	4	more of 99%
Ward criterion	Jaccard's index	479	5	
	Russel–Rao's index	218	4	
	Ochiai's index	459	5	
	Dice's index	478	5	
Descendant clustering	Varclus procedure	165	4	

So we consider this atypical group separately and isolate this cluster from the others to mine association rules. As shown in Fig. 7, the number and maximum size of rules found in others clusters is moderate.

Hidden atypical cluster 2 has been identified through the previous clustering regardless of the method used. So if we withdraw this cluster from every obtained partition, we end up with a significant reduction in the number and complexity of rules produced.

Table 9 shows that reduction is better than 99% in every case. In addition, the maximum complexity is not higher than five items, so rules will be easier to analyse.

4. Application to another example

4.1. Presentation of the data set and basic association rules mining

We have also applied our approach to the BMS-WebView-2 data set which contains several months worth of click-stream data from an e-commerce web site. This data set was used in KDD-cup 2000 competition and is available at KDD-cup 2000 home page (`url: http://www.ecn.purdue.edu/KDDCUP/`). Each transaction in this data set is a web session, i.e. the list of product-pages viewed in that session. The goal is to find associations between products viewed by visitors in a single visit to the web site.

This data set is comparable to our industrial data. Indeed, BMS-WebView-2 data set contains 77 512 transactions and 3340 items. Moreover, 4.62 is the average transaction size and the corresponding binary matrix contains about 0.14% of "1".

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Table 10
Results of a basic association rules mining (tr = transactions)

Minimum support threshold	Number of frequent itemsets	Number of association rules with minimum confidence = 50%	Maximum complexity (size) of rules	Number of association rules with minimum confidence = 90%	Maximum complexity (size) of rules
1000 tr	46	9	3	0	—
750 tr	16	90	3	0	—
500 tr	220	141	4	0	—
400 tr	380	378	5	2	5
300 tr	734	1230	5	28	5
200 tr	1954	7122	7	366	6
100 tr	11 927	179 381	9	17 506	9

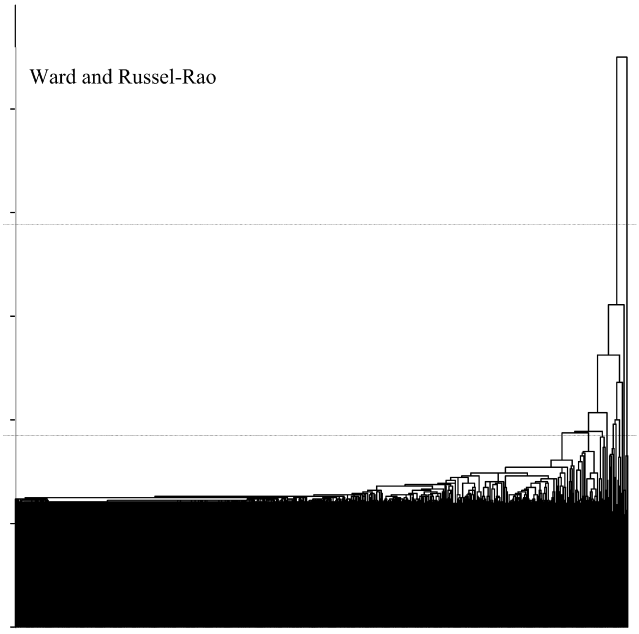

Fig. 6. Hierarchical tree obtained by ascendant hierarchical clustering with Ward method and Russel & Rao index.

We have proceeded to a basic association rules mining on this data set. The minimum threshold of the support has to be very low to find some association rules. With a minimum support fixed at 1000 transactions, we obtain only nine rules with a minimum confidence of 50%, and 0 rules with a minimum confidence of 90%. By reducing minimum support as shown in Table 10, the number of rules increases and reaches 179 381 with a minimum support of 100 transactions. In next section, we use variable clustering methods in order to reduce both the number and the complexity of these rules.

BMS-WebView-2 data set has been also used by Iodice D'Enza et al. (2005). In their paper, the authors propose a two interactive steps: graphical approach using factorial plans to identify a priori potentially interesting items. Like us, they propose to first perform a clustering method: incremental K-means with Jaccard index, but whereas we cluster items, they cluster transactions in order to produce homogeneous groups on which they apply a multidimensional factorial analysis.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

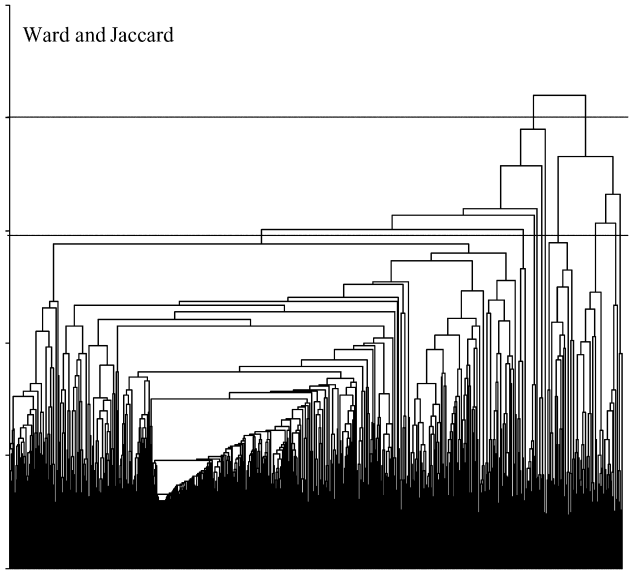

Fig. 7. Hierarchical tree obtained by ascendant hierarchical clustering with Ward method and Jaccard index.

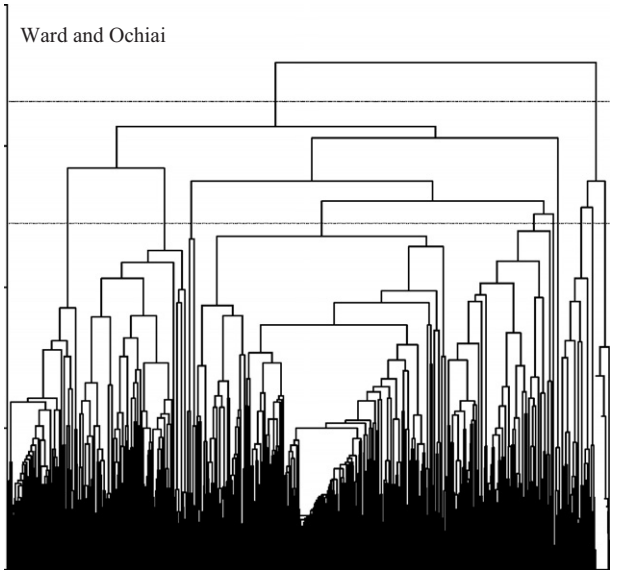

Fig. 8. Hierarchical tree obtained by ascendant hierarchical clustering with Ward method and Ochiai index.

4.2. Using clustering methods

To cluster items, we employed an ascendant hierarchical clustering with the Ward method, and the same similarity indexes as above (Russel–Rao, Jaccard, Dice, Ochiai and Pearson). It leads to hierarchical trees presented from Figs. 6 to 9.

We have chosen two types of partitions: partitions in 2 clusters and partitions in 10 clusters. Table 11 shows that, whatever is the partition, there always is a big cluster containing most of variables and smaller clusters.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

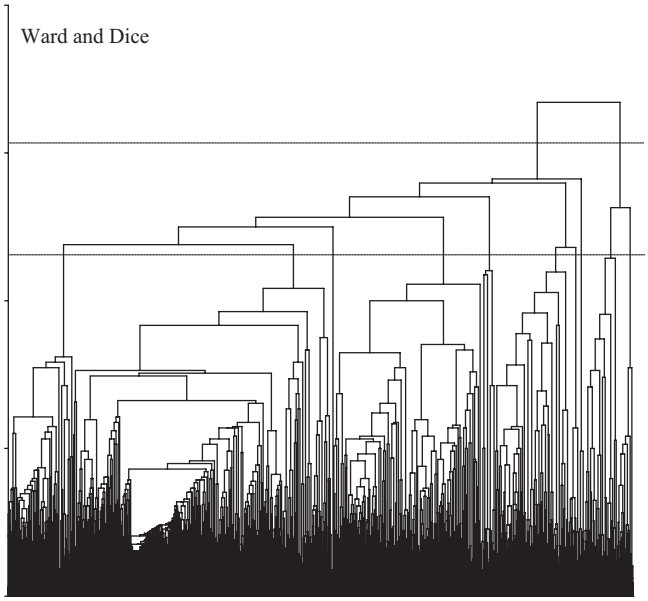

Fig. 9. Hierarchical tree obtained by ascendant hierarchical clustering with Ward method and Dice index.

Table 11
Number of variables per cluster in the two partitions

	Cluster	Russel–Rao	Jaccard	Dice	Ochiai	Pearson
2-clusters partition	1	3325	2927	3081	3065	3331
	2	15	413	259	275	9
10-clusters partition	1	3051	2767	1382	1390	3328
	2	131	156	793	602	3
	3	41	155	391	585	2
	4	38	81	345	390	1
	5	26	55	175	162	1
	6	15	51	84	92	1
	7	13	32	72	53	1
	8	11	21	55	34	1
	9	7	11	32	21	1
	10	7	11	11	11	1

Then, we have mined association rules in each cluster of each partition with a minimum support of 100 transactions and two minimum thresholds for the confidence: 50% and 90%. Results are presented in Table 12 for the 2-cluster partitions and in Table 13

for the 10-cluster partitions.

There is a reduction of the number of rules only when Russel–Rao and Pearson indexes are employed but there is no simplification of the rules. Best results are obtained with Pearson's index with a 95%-reduction of the number of rules in the case of the minimum threshold for confidence is 90%.

In the 10-cluster partitions, when there is not a significant reduction, we note that there is always a cluster in which the number of rules explodes: cluster 3 for Jaccard index, cluster 7 for Dice and Ochiai indexes. Those clusters gather almost the same group of variables. A small part of these variables that are very dependent causes the explosion of the number of rules. These variables are separate in different clusters in the other partitions obtained with Russel–Rao and Pearson indexes. That is why there is a reduction of the number of rules in those two partitions. It is possible to isolate these very dependent variables to analyse them separately but the interpretation of the contents of the clusters has to be made by an expert who knows the data well.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Table 12
Results of association rules mining in the 2-cluster partitions

		Minimum support = 100 tr		Minimum confidence = 50%			Minimum confidence = 90%		
		Number of frequent itemsets	Reduction of number of frequent itemsets (%)	Number of rules	Maximum size	Reduction of number of rules (%)	Number of rules	Maximum size	Reduction of number of rules (%)
Without classification		11 927	—	179 381	9	—	17 506	9	—
Russel–Rao	Cluster 1	2716		15 487	8		616	8	
	Cluster 2	4305		76 810	9		6231	9	
	Total	7021	41	92 297	9	49	6847	9	61
Jaccard	Cluster 1	926		316	5		8	5	
	Cluster 2	10 917		179 010	9		17 497	9	
	Total	11 843	1	179 326	9	0	17 505	9	0
Dice	Cluster 1	11 779		179 378	9		17 506	9	
	Cluster 2	148		3	2		0	0	
	Total	11 927	0	179 381	9	0	17 506	9	0
Ochiai	Cluster 1	11 783		179 378	9		17 506	9	
	Cluster 2	144		3	2		0	0	
	Total	11 927	0	179 381	9	0	17 506	9	0
Pearson	Cluster 1	2826		15 602	8		621	8	
	Cluster 2	511		6290	9		214	9	
	Total	3337	72	21 892	9	88	835	9	95

To conclude this section, even if the reduction of the number of rules is not systematic, the previous clustering of variables makes it possible to identify atypical clusters to analyse more closely. In this direction, previous clustering simplifies the search for association rules. Nevertheless it is necessary to take care to use several similarity indexes because they seem to bring complementary information.

Lastly, to extract the most interesting rules, we use some relevancy measures. This is the topic of the next paragraph.

5. Selecting relevant rules

To complete the support-confidence approach in the selection of interesting rules, there are many relevancy indexes. Association rule $A \rightarrow C$ is completely described by its corresponding contingency table. That is why every relevancy index (or measure of interest) is built from Table 14.

5.1. Choosing relevancy indexes

There are so many relevancy indexes that it is difficult to select just one. The choice of an index depends on both the data and its context. We can refer to previous works (Lenca et al., 2004; Vaillant et al., 2004) to help us classify and then select an appropriate index. From these works, we have selected four indexes: centred confidence, lift, odd multiplier (OM) and Loevinger. We have then studied their behaviour using our data in order to select the best one (Plasse et al., 2006). Our data present many uninteresting rules where consequent is relatively frequent with regard to antecedent.

The four indexes mentioned above present very high values for these kinds of rules. For that reason we have added the Jaccard coefficient and agreement–disagreement index (ADI) which better discriminate relevant rules in our application. ADI is an old coefficient (Kulczynski, 1927) and is equivalent to the Jaccard coefficient, i.e. $1/Jaccard = (1/MAD) + 1$. While they lead to the same classification of rules, the Jaccard coefficient has the advantage of varying between 0 and 1. Table 15 shows definitions of the six compared indexes and gives their particular values.

To illustrate that the Jaccard index and ADI make a better selection of rules, Fig. 10 presents an example of an irrelevant rule in our case because of the high value of the consequent compared to that of the antecedent. The lift for example is high, whereas the Jaccard index and ADI are low.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Table 13
Results of association rules mining in the 10-cluster partitions

		Minimum support = 100 tr		Minimum confidence = 50%			Minimum confidence = 90%		
		Number of frequent itemsets	Reduction of number of frequent itemsets (%)	Number of rules	Maximum size	Reduction of number of rules (%)	Number of rules	Maximum size	Reduction of number of rules (%)
Without classification		11 927	—	179 381	9	—	17 506	9	—
Russel–Rao	Cluster 1	717		38	3		0	0	
	Cluster 2	192		19	3		0	0	
	Cluster 3	139		193	5		8	5	
	Cluster 4	145		116	4		3	4	
	Cluster 5	99		28	3		0	0	
	Cluster 6	4305		76 810	9		6231	9	
	Cluster 7	1042		14 232	8		566	8	
	Cluster 8	71		66	4		0	0	
	Cluster 9	113		573	6		38	5	
	Cluster 10	67		116	4		0	0	
	Total	6890	42	92 191	9	49	6846	9	61
Jaccard	Cluster 1	859		311	5		8	5	
	Cluster 2	61		0	0		0	0	
	Cluster 3	10 769		179 007	9		17 497	9	
	Cluster 4	81		3	2		0	0	
	Cluster 5	15		0	0		0	0	
	Cluster 6	50		5	2		0	0	
	Cluster 7	1		0	0		0	0	
	Cluster 8	6		0	0		0	0	
	Cluster 9	1		0	0		0	0	
	Cluster 10	0		0	0		0	0	
	Total	11 843	1	179 326	9	0	17 505	9	0
Dice	Cluster 1	701		869	6		42	6	
	Cluster 2	154		10	3		0	0	
	Cluster 3	390		216	5		8	5	
	Cluster 4	67		2	2		0	0	
	Cluster 5	67		0	0		0	0	
	Cluster 6	81		3	2		0	0	
	Cluster 7	9325		172 821	9		17 233	9	
	Cluster 8	15		0	0		0	0	
	Cluster 9	1		0	0		0	0	
	Cluster 10	1		0	0		0	0	
	Total	10 802	9	173 921	9	3	17 283	9	1
Ochiai	Cluster 1	769		883	6		42	6	
	Cluster 2	158		55	4		0	0	
	Cluster 3	335		222	5		8	5	
	Cluster 4	112		11	3		0	0	
	Cluster 5	57		0	0		0	0	
	Cluster 6	81		3	2		0	0	
	Cluster 7	9291		172 813	9		17 233	9	
	Cluster 8	1		0	0		0	0	
	Cluster 9	6		0	0		0	0	
	Cluster 10	1		0	0		0	0	
	Total	10 811	9	173 987	9	3	17 283	9	1
Pearson	Cluster 1	2742		15 529	8		618	8	
	Cluster 2	7		2	2		0	0	
	Cluster 3	3		2	3		0	0	
	Cluster 4	1		0	0		0	0	
	Cluster 5	1		0	0		0	0	
	Cluster 6	1		0	0		0	0	

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Table 13 (*Continued*)

	Minimum support = 100 tr		Minimum confidence = 50%			Minimum confidence = 90%		
	Number of frequent itemsets	Reduction of number of frequent itemsets (%)	Number of rules	Maximum size	Reduction of number of rules (%)	Number of rules	Maximum size	Reduction of number of rules (%)
Cluster 7	1		0	0		0	0	
Cluster 8	1		0	0		0	0	
Cluster 9	1		0	0		0	0	
Cluster 10	1		0	0		0	0	
Total	2759	77	15 533	8	91	618	8	96

Table 14
Contingency table of a rule

	C	$\bar{C}$	Lines profiles
A	$P(AC)$	$P(A\bar{C})$	$P(A)$
$\bar{A}$	$P(\bar{A}C)$	$P(\bar{A}\bar{C})$	$P(\bar{A})$
Columns profiles	$P(C)$	$P(\bar{C})$	1

Table 15
Selected indexes of rules relevancy

Relevancy indexes	Definition	Incompatibility $P(AC)=0$	Independency $P(AC)=P(A)P(C)$	Logical rule $P(C/A)=1$
Centred confidence	$P(C/A)-P(C)$	$-P(C)$	0	$P(\bar{C})$
Lift	$\frac{P(AC)}{P(A).P(C)}$	0	1	$\frac{1}{P(C)}$
OM	$\frac{P(AC)P(\bar{C})}{P(A\bar{C})P(C)}$	0	1	$+\infty$
Loevinger	$\frac{P(C/A)-P(C)}{P(\bar{C})}$	$\frac{-P(C)}{P(\bar{C})}$	0	1
ADI	$\frac{P(A\cap C)}{P(A\Delta C)}$	0	$\frac{P(A)P(C)}{P(A)+P(C)-2P(A)P(C)}$	$\frac{P(A)}{P(C)-P(A)}$
Jaccard	$\frac{P(A\cap C)}{P(A\cup C)}$	0	$\frac{P(A)P(C)}{P(A)+P(C)-P(A)P(C)}$	$\frac{P(A)}{P(C)}$

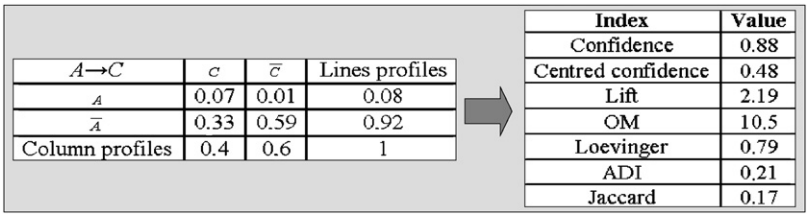

Fig. 10. Values of rules relevancy indexes for the same rule.

5.2. Comparing indexes using graphics

We propose a graphical tool to make it easier to compare the six indexes. This innovative representation using level lines or curves allows us to compare behaviour of selected indexes of rule relevancy. This work requires the rewriting of indexes according to the conditional probabilities $\lambda_A = P(C/A)$ and $\lambda_C = P(A/C)$. As we can see in Table 16, only the Jaccard index and ADI can be written according to the two conditional probabilities simultaneously; for this reason they improve rules discrimination. Figs. 11–16 show level lines for centred confidence, lift and Loevinger and level curves for OM, Jaccard index and ADI.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

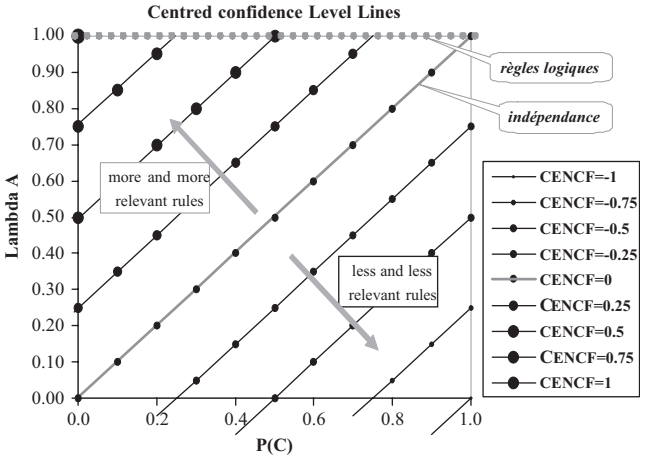

Fig. 11. Level lines or curves according to $P(C)$ and λ_A.

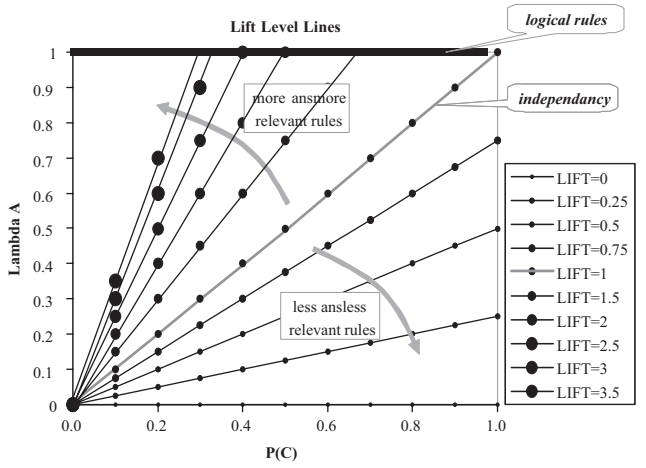

Fig. 12. Level lines or curves according to $P(C)$ and λ_A.

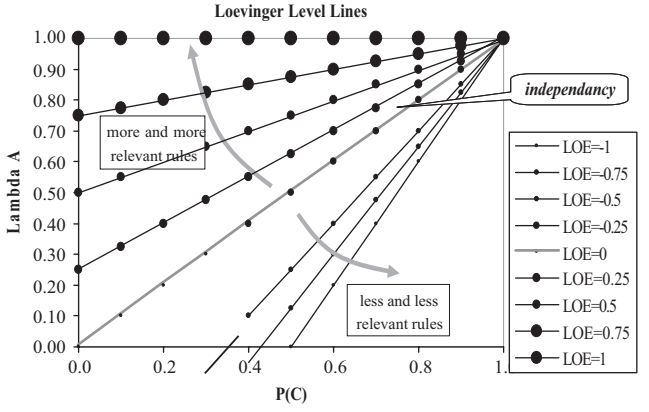

Fig. 13. Level lines or curves according to $P(C)$ and λ_A.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

Table 16
Equations of level lines and curves

Indexes	Definition	Indexes	Definition
Centred confidence	$\lambda_A = CONFCEN - P(C)$	Loevinger	$\lambda_A = LOE(1 - P(C)) + P(C)$
Lift	$\lambda_A = LIFT.P(C)$	ADI	$\lambda_A = \frac{1}{\frac{1}{IAD} - \frac{1}{\lambda_C} + 2}$
OM	$\lambda_A = \frac{MC.P(C)}{1-P(C)+MC.P(C)}$	Jaccard	$\lambda_A = \frac{1}{\frac{1}{JAC} - \frac{1}{\lambda_C} + 1}$

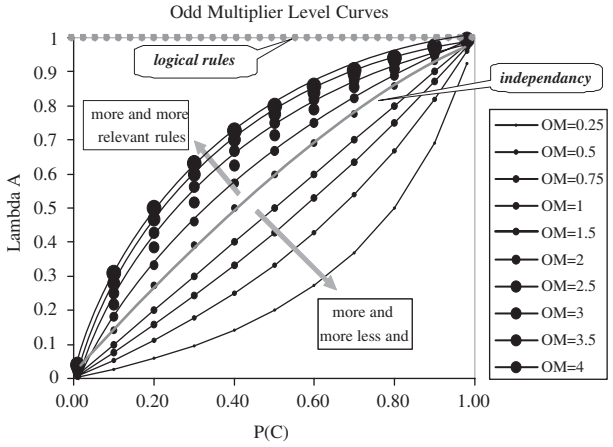

Fig. 14. Level lines or curves according to $P(C)$ and λ_A.

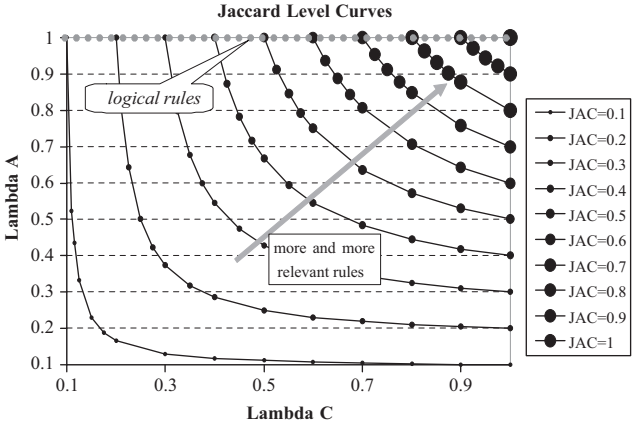

Fig. 15. Level curve according to λ_A.

If λ_A is high, lift and centred confidence give priority to rules where the consequent is not very frequent whereas the Loevinger index and OM depend above all on λ_A. A rule will be relevant according to Jaccard and ADI if the two conditional probabilities are high. These two indexes are appropriate in our case but do not suit every application. Note that they are symmetric, contrary to the others. This is not a drawback in our application where the direction of the rule is of little importance.

In addition, the Jaccard index and support can be represented on the same graph as shown in Fig. 17. This figure can be read in two ways. Support curves are read with $P(A)$ and $P(C)$ whereas Jaccard curves are read with λ_A and λ_C. Knowing only three of them, it is possible to deduce the others using this graphical representation. This visualisation tool can be used to quickly evaluate the interest of a rule without computing the Jaccard index.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

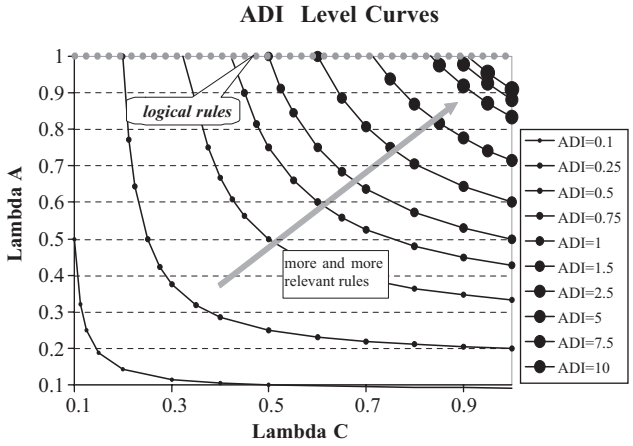

Fig. 16. Level curve according to λ_C.

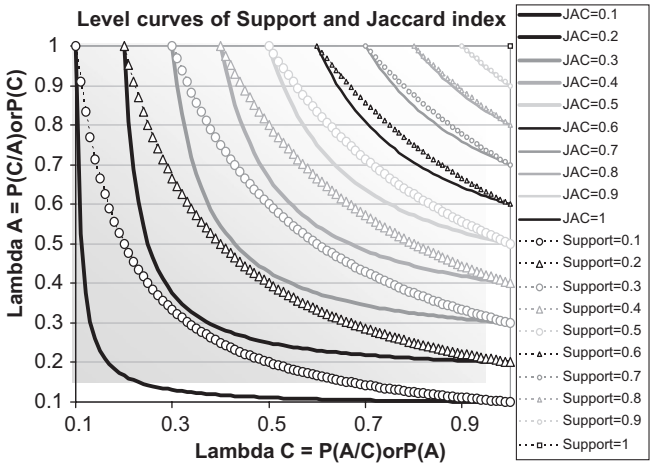

Fig. 17. Simultaneous representation of support and Jaccard index.

5.3. Application examples

We can use the Jaccard index to complete the support–confidence approach. After mining association rules inside each cluster, we select the most relevant rules using a minimum threshold for the Jaccard index. For instance, referring to Table 8, we have mined rules with a minimum support of 100 vehicles and a minimum confidence of 90%. Within 17 rules found in cluster 10, 8 have a Jaccard index above 0.6. So we analyse these 8 as a matter of priority.

In other respects, some clusters can lead to a large number of interesting rules depending on the Jaccard index. For instance, atypical cluster 2 produces 171 229 rules. Amongst them, 74% have a Jaccard index of over 0.6 and 17% above 0.8. So, minimum thresholds have to be high and severe in order to select the best rules first.

6. Conclusion

In this paper, we have proposed a way of discovering hidden links between binary attributes in a large sparse data set. Performed in isolation, association rules mining produces an enormous and impractical number of rules. Variable clustering can be used to reduce this number significantly. In our application it provides homogenous clusters of variables and highlights an atypical cluster of highly correlated attributes that produce numerous and complex set of

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020

rules. Once this group has been isolated, rules that are normally invisible when mining association rules on the whole data set are now revealed.

Our approach has been developed in order to analyse real data—a large sparse collection of industrial data from the car manufacturing industry. Our aim was to investigate links between attributes that describe a set of vehicles. The reduction in the number of rules allowed us to analyse and validate some interesting links.

Finally, we have identified more relevant rules using the Jaccard index, proving our hypothesis by comparing graphically rule relevancy indexes. This innovative representation is based on the use of level curves or lines and reveals to us how indexes behave according to different rules.

Following this study, we are currently investigating the benefits of applying simultaneous clustering of rows and columns to our approach rather than variable clustering. This simultaneous clustering will provide us with homogeneous blocks of zero or one representing vehicles with identical attributes. Once obtained, we will mine association rules inside each of these blocks.

References

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases (VLDB), Santiago, Chile.

Agrawal, R., Imielinski, T., Swani, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of the ACM-SIGMOD Conference on Management of Data, Washington DC, USA.

Brin, S., Motwani, R., Ullman, J.D., Tsur, S., 1997a. Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the ACM-SIGMOD Conference on Management of Data, Tucson, Arizona, USA.

Brin, S., Motwani, R., Silversein, C., 1997b. Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the ACM-SIGMOD Conference on Management of Data, Tucson, Arizona, USA.

Fichet, B., Le Calvé, G., 1984. Structure géométrique des principaux indices de dissimilarité sur signes de présence-absence. Statist. Anal. Données 9 (3), 11–44.

Forgy, E.W., 1965. Cluster analysis of multivariate data: efficiency versus interpretability models. Biometrics 61 (3), 768–769.

Gower, J.C., Legendre, P., 1986. Metric and Euclidian properties of dissimilarity coefficients. J. Classification 3, 5–48.

Han, J., Pei, J., Yin, Y., 2000. Mining frequent patterns without candidate generation. In: Proceedings of the ACM-SIGMOD Conference on Management of Data, Dallas, Texas, USA.

Iodice D'Enza, A., Palumbo, F., Greenacre, M., 2005. Exploratory data analysis leading towards the most interesting binary association rules. In: XIth International Symposium on Applied Stochastic Models and Data Analysis Brest, France.

Kulczynski, S., 1927. Classe des Sciences mathématiques et Naturelles. Bull. Internat. Acad. Polonaise Sci. Lett. Sér. B (Sci. Natur.) (Suppl. II) 57–203.

Lenca, P., Meyer, P., Vaillant, B., Picouet, P., Lallich, S., 2004. Evaluation et analyse multicritère des mesures de qualité des règles d'association. Mesures de qualité pour la fouille de données, no spécial RNTI Revue des Nouvelles Technologies de l'Information, Cepaduès.

Nakache, J.P., Confais, J., 2005. Approche pragmatique de la classification. Ed. Technip.

Nicolau, F.C., Bacelar-Nicolau, H., 2003. Teaching and learning hierarchical clustering probalistic models for categorical data. ISI 54th session, Berlin.

Plasse, M., Niang, N., Saporta, G., 2005a. Utilisation conjointe des règles d'association et de la classification de variables. In: 37emes Journées de Statistique de la SFdS, Pau, France.

Plasse, M., Niang, N., Saporta, G., Gauthier, D., 2005b. Combined use of association rules mining and clustering methods. In: Third World Conference on Computational Statistics & Data Analysis, IASC, Limassol, Cyprus.

Plasse, M., Niang, N., Saporta, G., Leblond, L., 2006. Une comparison de certains indices de pertinence des règles d'association. Revue des Nouvelles Technologies de l'Information, In: Actes 6e Conférence Extraction et Gestion des Connaissances, EGC'06, Série E, no. 6, vol. II, Lille, pp. 561–568.

Qannari, E.M., Vigneau, E., Courcoux, P., 1998. Une nouvelle distance entre variable—application en classification. Rev. Statist. Appl. XLVI (2), 21–32.

SAS Institute Inc, 2003. SAS/STAT User's Guide. SAS Institute Inc, Cary, NC.

Savasere, A., Omiecinski, E., Navathe, S., 1995. An efficient algorithm for mining association rules in large databases. In: Proceedings of the 21st VLDB Conference, Zurich, Swizerland.

Toivonen, H., 1996. Sampling large databases for association rules. In: Proceedings of the 22nd VLDB Conference, Mumbai (Bombay), India.

Vaillant, B., Lenca, P., Lallich, S., 2004. Etude expérimentale de mesures de qualité de règles d'association. Revue des Nouvelles Technologies de l'Information, Actes 4e Conférence Extraction et Gestion des Connaissances, EGC'04, Série E, no. 2, vol. 2, Clermont-Ferrand, pp. 341–352.

Vigneau, E., Qannari, E.M., 2003. Clustering of variables around latent component; application to sensory analysis. Commun. Statist. Simulation Comput. 32 (4), 1131–1150.

Youness, G., Saporta, G., 2004. Some measures of agreement between close partitions. Student 5 (1), 1–12.

Zaki, M.J., 2000. Scalable algorithms for association mining. IEEE Trans. Knowledge Data Eng. 12 (3), 372–390.

Please cite this article as: Plasse, M., et al., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.02.020