Random simulations of a datatable for efficiently mining reliable and non-redundant itemsets

Martine Cadot, Pascal Cuxac, Alain Lelu

Random simulations of a datatable for efficiently mining reliable and non-redundant itemsets

Martine Cadot[1], Pascal Cuxac[2], and Alain Lelu[3,4]

[1] UHP/Loria, Nancy
(e-mail: Martine.Cadot@loria.fr)
[2] INIST, Nancy
(e-mail: Pascal.Cuxac@inist.fr)
[3] INRA, Crebi, Jouy en Josas
[4] UFC/Laseldi, Besanon
(e-mail: Alain.Lelu@jouy.inra.fr)

Abstract. Our goal is twofold: 1) we want to mine the only statistically valid 2-itemsets out of a boolean datatable, 2) on this basis, we want to build the only higher-order non-redundant itemsets compared to their sub-itemsets. For the first task we have designed a randomization test (Tournebool) respectful of the structure of the data variables and independant from the specific distributions of the data. In our test set (959 texts and 8477 terms), this leads to a reduction from $126,000$ 2-itemsets to $13,000$ significant ones, at the 99% confidence interval. For the second task, we have devised a hierarchical stepwise procedure (MIDOVA) for evaluating the residual amount of variation devoted to higher-order itemsets, yielding new possible positive or negative high-order relations. On our example, this leads to counts of 7,712 for 2-itemsets to 3 for 6-itemsets, and no higher-order ones, in a computationally efficient way.
Keywords: Text Mining, Randomization Tests, Significant Itemsets, Statistical Interaction, Multiple Comparison.

1 Introduction : setting the problem

In a knowledge discovery task, knowledge nuggets are brought to light in the shape of itemsets, i.e. "interesting" conjunctions of boolean variables from which association rules may be derived. Presently, more than fifty different measures are used for trying to establish the quality of association rules [10]. The difficulty to use them comes from the diversity of their own (or common) semantics [12], but also from the uncertainty about their statistical robustness [14]. Two reasons behind these difficulties: 1) the concept of independance is simple to understand for 2 variables, but is not so obvious beyond [3]; the concept of dependance is still worse, as many definitions are available... 2) one catches easily the concept of interaction for 2 variables faced to a third one, but the extension to three and more is nothing but evident [15]. A k-itemset expresses an imprecise notion of relation between k variables, and it may refer implicitly to the first viewpoint or the second.

This is why we have developped two distinct inspiration lines for this unique reality. Moreover, combinatorial explosion is a recurrent issue for the item-sets extraction algorithms in this domain. We have developed two ideas for tackling the overall difficulty:

Idea 1: Given a text collection, two frequent words are much more likely to occur altogether in a text than two infrequent ones. Simulating random variants of the presence/absence data, it is thus possible to evaluate whether or not the support of any itemset is too small (or too large, in the case of rare itemsets) to be fortuitous. Our *TourneBool* method determines a context-dependant confidence interval for the support of each 2-itemset, thus selecting the only statistically validated ones.

Idea 2: A, B, C being three words in a text collection, if A and B often occur simultaneously in the same texts, the number of texts with A and C together does not differ much from the one with B and C together. For each itemset, our *Midova* method yields 1) the variation share left to the support of its super-itemsets (Midova *residue index*), 2) the deviation between its support and the expected median support, given its sub-itemsets (Midova *gain index*).

Hence our operating sequence: Given a set of objects characterized by a set of boolean features, our goal is to mine the only informative k-associations between features (k-itemsets):

- selecting the only statistically valid 2-itemsets, those "too (in)frequent" to be hazardous (TourneBool method),
- for $k = 2, 3, 4, ...$ selecting the only k-itemsets highlighted by an appreciable support variation left by their sub-itemsets (Midova method).

2   Statistical validation of 2-itemsets connections

TourneBool method:

- belongs to the class of randomization tests [13],
- validates the significant 2-itemsets with a "cascade-permutation" method [6,8].

The principle is to generate a randomized version of the initial datatable, under the constraint of keeping the same margins (row and column sums) as the original datatable (see figure 1). We have shown in [5] that this constraint was mandatory to distinguish structural effects (associations due to the sole effect of margins of the datatable) from meaningful ones.

This process is repeated at least a hundred times. For each 2-itemset in the initial data, the distribution of its support in the randomized tables is computed. The initial value is compared to the 2.5% head and 2.5% tail values of the distribution (in case of 95% confidence choice): if it falls outside this interval, the relation is declared significant, positively or negatively.

Texts	K1	K2	K3		K631	K632	Total
			Keywords				
T1	1	1	0		0	0	20
T2	0	0	1		0	0	12
...	0	0	0	C	1	1	8
T1358	1	1	1		0	0	5
T1359	1	1	0	S	1	0	30
T1360	0	0	1		0	0	2
T1361	1	1	1	.	0	0	14
Total	255	139	55	.	2	1	

Fig. 1. Example of a cascade exchange (C, squares), and a rectangular exchange (S, circles).

TourneBool algorithm: Let M a boolean matrix (n, p), with n objects in rows and p variables in columns.

Main : TourneBool

- 1. generate q randomized versions of M
- 2. for each column pair (i, j)
  - determine the lower and upper bound of the support's confidence interval after the list of the support values in the randomized matrices.
  - 3 cases:
    * if the original itemset support in M stays in between this interval, it is declared unsignificant and is thus eliminated,
    * if it is lesser than the lower bound, it is declared significantly rare, and is thus kept on,
    * if it is greater than the upper bound, it is declared significantly frequent, and is thus kept on.

Building q randomized versions of M:

- choose a number r of rectangular exchanges to execute
- 1. copy M to Mc
- 2. repeat q times :
  - 2.1 repeat r times :
    * choose at random with replacement a row pair and a column pair
    * if the zeros and ones alternate at the corners of this rectangle in Mc, then modify Mc moving each value into its complement to 1, else do nothing.
  - 2.2 store Mc

Building the confidence interval, at risk alpha, of an itemset (i, j) of M:

- 1. for each randomized version of M compute the support of the itemset (i, j) (dot product of the two columns) Store all the supports in a list.
- 2. sort the list in ascending order. The lower bound is the list element with rank $q*alpha/2$, and the lower bound the one with $q * (1 - alpha)/2$ rank.

3   MIDOVA algorithm and indices

Midova : Given three yet determined itemsets A, B, C, the MIDOVA method [7] looks for the variation interval of the ABC itemset support, in the framework of fixed support constraints for its sub-itemsets AB, AC, BC. Counts are established considering elementary unitary swaps ($A \rightarrow nonA$, $nonA \rightarrow A$, $B \rightarrow nonB$, ...)

Midova indices for a k-itemset M: Combinatorial considerations developed in [8] show that:

- the gain index g is a function of the support s of the k-itemset M, of its length k, and of the center c of the support variation interval, whose lower an upper bounds are respectively sl and su :
  $g = 2^{k-1}(s - c)$ where
  - $c = (sl + su)/2$
  - $sl = s - min("odd" \ zones)$; see figure 2
  - $su = s + min("even" \ zones)$
- the residue index r is a function of k, s, sl, su:
  $r = 2^{k-1}min(|s - sl|, |s - su|)$
  where g and r are expressed with the same units as the support s, i.e as a number of objects. A reasonable choice for the significance threshold of r seems $r \geq 2$. It is worth to notice that Midova parametered with the $s \leq t$, t being a given positive threshold, and $r \geq 2$, amounts to the A-priori algorithm [1,2], intrinsically devoted to frequent itemsets.

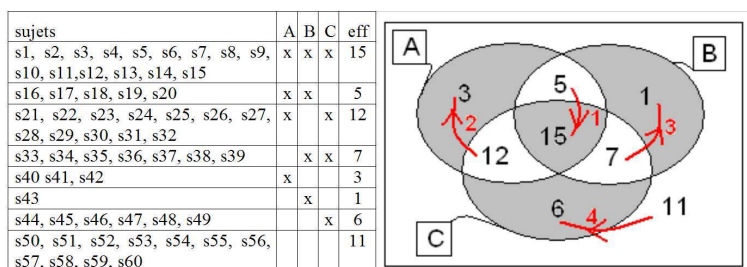| sujets	A	B	C	eff
s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11,s12, s13, s14, s15	x	x	x	15
s16, s17, s18, s19, s20	x	x		5
s21, s22, s23, s24, s25, s26, s27, s28, s29, s30, s31, s32	x		x	12
s33, s34, s35, s36, s37, s38, s39		x	x	7
s40 s41, s42	x			3
s43		x		1
s44, s45, s46, s47, s48, s49			x	6
s50, s51, s52, s53, s54, s55, s56, s57, s58, s59, s60				11

Fig. 2. Distribution of 60 subjects over 3 boolean features.

The choice of $s \leq 0$ yields infrequent itemsets, i.e. pairs, triplets, etc. of "opposite" variables, less encountered than expected at random. These interesting negative relations are generally not considered in the knowledge extraction literature.

Midova algorithm

- 0) Initializing: A threshold value is chosen as the measure of a negligible gap, expressed as a number of objects (0, 1, 2, ...)
- 1) Level 1 (1-itemsets, i.e. boolean variables): the support s of each itemset is somewhere between 0 and N. Compute:
  - the difference to the "neutral center" $N/2$ of this interval,
  - the variation share left to the supports of its super-itemsets (i.e. 2-itemsets), which writes s if $s > N/2$, else $N - s$,
  then the 1-itemsets for which this share is negligible ($s \leq e$) or almost equal to N ($s \geq N-e$) have exhausted their variation potential, and are therefore eliminated in the sequel of the process.
- 2) Level k;

- For each k-itemset Mk issued from the candidate itemsets of lower order $k-1$, the support s of Mk is stored, and its variation interval $[sl\ ;\ su]$ is computed, as well as the difference with the center of this interval, and the variation share left to the super-itemsets of higher order: this amount is the Midova residue (i.e. the difference with the closest interval limit sl or su). If this amount is negligible, the current itemset is considered to have exhausted its variability potential, and is thus eliminated from the list of combinable itemsets at the next $k+1$ stage.
- While k-itemsets stay uneliminated: increment k and return to 2).

4   Application to a real text database:

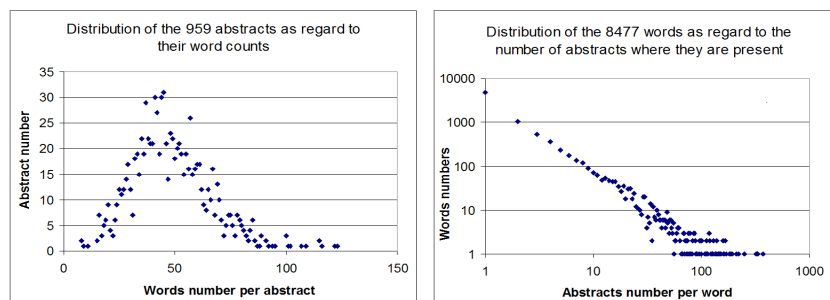## 4.1   Presentation of our "Geotechnics" database

Fig. 3. Distribution of the abstracts and words in the corpus

Our test set of 959 scientific paper abstracts is drawn from the CNRS/INIST Pascal bibliographic database: papers published in 2003, in english, in the field of Geotechnics. We have applied to this corpus a rough NLP procedure, giving rise to 8477 lemmatized words distributed over the sole noun, verb, adjective categories, and eliminated syntactic particles.

In the left part of the figure 3, the word distribution in the abstracts appears to be approximately binomial, with a central value of about 50 words per text. The right part shows the very unequal, Zipfian, distribution of word counts in the corpus, as usual in language statistics.

4.2   Overall results on our "Geotechnics" texts set

With TourneBool (alpha-risk=0.01) and Midova (residue $r \geq 5$), all the item-sets are extracted, whatever their support ($s \geq 0$). Starting from the 8477 variables (words), it yields :

- step 1, Midova: given 8477 itemsets, 1707 are kept (with $r \geq 5$) for composing 2-itemsets.
- step 2, TourneBool: given 1,456,071 2-itemsets, 7712 significant ones are kept with 0.01. alpha risk.
- step 3, Midova: 5046 2-itemsets are kept on (with $r \geq 5$) for setting up 3-itemsets
- step 4, Midova: starting from 4160 3-itemsets, 2214 are kept on (with $r \geq 5$) for setting up 4-itemsets
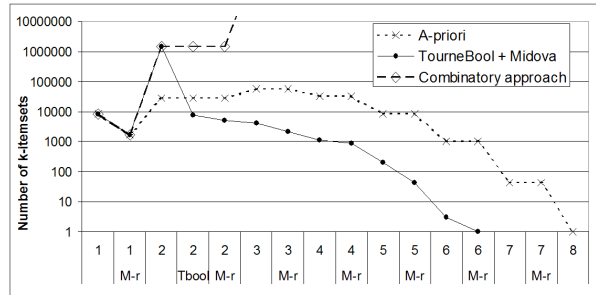
Fig. 4. Comparison TourneBool+Midova vs. A-priori (1)

- step 5, Midova: starting from 1123 4-itemsets, 881 are kept on (with $r \geq 5$) for setting up 5-itemsets
- step 6, Midova: starting from 207 5-itemsets,43 are kept on (with $r \geq 5$) for setting up 6-itemsets
- step 7, Midova: starting from 3 6-itemsets, 1 is kept on (with $r \geq 5$) for setting up 7-itemsets
- step 8 : no 7-itemset is set up and the algorithm stops.

The process is represented in the figure 4 in full lines, and in figure 5 by black bars, which show that our process clearly outperforms A-priori method by the number and length of itemsets criteria.
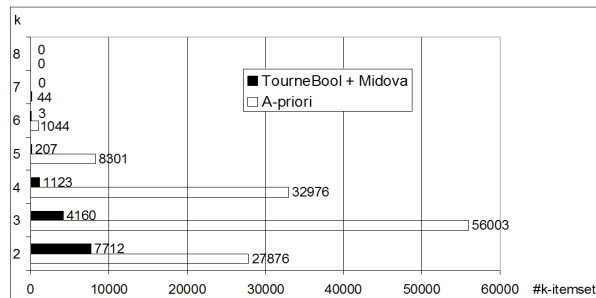

Fig. 5. Comparison TourneBool+Midova vs. A-priori (2)

4.3   Interpretation of our "Geotechnics" k-itemsets

We will limit our interpretation trial to the significant 2-itemsets issued from TourneBool, with extreme Midova-g values. we have selected a few interesting examples to our domain expert's eyes:

- 2-itemsets with support lesser than expected, and with strong negative gain:

- 'rock', 'soil', Mr=-163, support=25
- 'site', 'theory', Mr=-81, support=3
- 2-itemsets with support greater than expected, and with strong positive gain:
  - 'pore', 'pressure', Mr=51, support=68
  - 'conductivity', 'hydraulic', Mr=17, support=21
  - 'ash', 'fly', Mr=9, support=9
- 4-itemsets with strong positive gain:
  - 'model', 'modelling', 'finite', 'numerical', Mr=28, support=10
- Words contributing to significantly frequent 2-itemsets with 'site' (with $Mr \geq 4$): 'amplification', 'amplify', 'ground-motion', 'near-surface', 'recording', 'SEC', 'Shear-wave', 'spectral', 'spectrum', 'S-wave', 'unconsolidated'. The expert has admitted that these associated terms evoke the concept of *seism*.

5   Conclusion

Computer efficiency: Implemented as sparse matrices processing, out of any particular optimization effort, the CPU efficiency seems promising: less than 15 minutes on a standard 2.2 GHz PC for the 200 simulations part of a TourneBool run, on our 1000 X 8500 Geotechnics data; a few minutes for Midova.

- Compared to A-priori, the TourneBool-Midova sequence has reduced the number of "interesting" k-itemsets from $126,200$ to $13,200$, i.e. about two orders of magnitude.
- TourneBool has divided the overall width of the k-itemsets pyramid by 3.6, at the type-1 error risk of 0.01 (see figure 5).
- Midova has reduced the overall height of the k-itemsets pyramid from $k = 8$ to $k = 6$ (with $r \geq 5$).

Advantages and limits: Our TourneBool algorithm allows one to build a statistical test well-adapted to data mining (many variables, data-flows...):

- it works for small amounts of variables as well as big ones; doing so, it tackles the "multiple comparison" problem [11],
- it fits to any probability distribution, with no need to explicitly specify this distribution [9] ("distribution-free" property).

Our choice for the central value of the gain is not yet ascertained, thus strong conclusions can be drawn for the only extreme positive or negative values of the gain. Negative interactions are tricky to interpret and merit further investigations.

References

1. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, Fast Discovery of Association Rules, in Advances in Knowledge Discovery and Data Mining, *U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy editors,* AAAI Press / MIT Press, Menlo Park, California, 1996, p. 307-328.
2. Bastide Y., *Data mining : algorithmes par niveau. techniques d'implantation et applications,* Thèse d'informatique, Université Blaise Pascal, Clermont-Ferrand, 2000.
3. Bavaud Franois, *Modèles et données : une introduction la Statistique uni-, bi- et trivariée.* Paris ; Montréal (Qc) : L'Harmattan, 1998
4. Brin S., Motwani R., Silverstein C. (1997). Beyond Market Baskets: Generalizing Association Rules to Correlations. *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, 1997, Tucson, Arizona, USA
5. Cadot M., Napoli A., 2003, Une optimisation de l'extraction d'un jeu de règles s'appuyant sur les caractéristiques statistiques des données, *RSTI-RIA-ECA-16/2003* pp. 631-656
6. Cadot M. , (2005) A Simulation Technique for extracting Robust Association Rules, *CSDA 2005* (Chypre)
7. Cadot M., Cuxac P., Franois C., (2006) Règles d'association avec une prémisse composée : mesure du gain d'information. EGC 2006: p. 599-600
8. Cadot M., (2006) *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association* Thèse d'informatique, Université de Franche-Comté, 2006
9. Droesbeke J.-J., Fine J., éditeurs. Inférence non paramétrique, les statistiques de rangs. Journées d'Etude en Statistiques de l'Association pour la Statistique et ses Utilisations, Edition de l'Université de Bruxelles, Ellipses,.1996
10. Guillet F. (2004), Mesure de qualité des connaissances en ECD, *Cours donné lors des journées EGC 2004*, Clermont-Ferrand, 20 janvier 2004.
11. Jensen D., Multiples comparisons in induction algorithms. *Kluwer Academic Publishers*, 1998 Boston p1-33
12. Lenca P., Meyer P., Picouet P., Vaillant B., Lallich S., Critères d'évaluation des mesures de qualité en ECD, *JS 2003*, Proceedings pp. 647-650, Lyon, 2003.
13. Manly B.F.J., *Randomization, Bootstrap and Monte Carlo methods in Biology.* Chapman & Hall/CRC, Boca Raton, Florida, USA. Texts in Statistical Science, 1997
14. James S. Press, The role of Bayesian and frequentist multivariate modeling in statistical Data Mining, dans *Statistical Data Mining and Knowledge Discovery*, H. Bozdogan, Chapman & Hall/CRC, Boca Raton, US, 2004
15. Winer B.J., Brown D.R., Michels K.M.() *Statistical principles in experimental design* (third edition) 1991