Combining Quality Measures of Association Rules with Applications in Distributed Web Log Mining

Ion IANCU[1], Mihai GABROVEANU[1], Adrian GIURCĂ[2]

[1] Faculty of Mathematics and Computer Science
University of Craiova, Romania
i_iancu@yahoo.com , mihaiug@central.ucv.ro
[2] Department of Internet Technology
Brandenburg Technical University, Cottbus, Germany
giurca@tu-cottbus.de

Abstract. In this paper, we propose a technique for to combine two or more pairs of quality measures for association rules in order to obtain a better pair of quality measures. This combining method can be used to mining association rules from distributed databases. Also, the paper presents an use case for mining association rules from distributed WWW servers access logs, using proposed measures. Reasoning on web logs on the base of association rules might be an attempt to produce personalized content on commercial web sites and portals without storing explicitly of the user-profile information.
Keywords: association rule, support, confidence, web mining
Math. Subject Classification 2000: 64H20, 03E72

1 Introduction

Association rules provide a means for representing dependencies between attribute value objects (data records) stored in a (distributed) database. Association rules show attribute value conditions that occur frequently together in a given dataset.

Let $\mathcal{D} = \{x_1, \ldots, x_n\}$ be a transactional database contains records described by a set of binary attributes (items) $\mathcal{I}$. For a specified transaction $x \in \mathcal{D}$ we will retrieve the value of attribute $A \in \mathcal{I}$ using $A(x)$. This value is either 1 or 0 indicating whether or not item A bought in transaction x.

Typically, an association rule involve two sets, A and B, of attributes and has the following form $A \rightharpoonup B$. The meaning of a rule $A \rightharpoonup B$ is that when A is bought in a transaction, B is likely to be bought as well.

An example of an association rule is: "*70% of clients who bought printers also bought paper*".

The quality of an association rule is expressed by several measures, among which *support* and *confidence* are two essential ones [2]. The *support* is the

number of transactions in which both A and B occurs and the *confidence* represents the percentage of transactions containing A that contain B as well.

In this paper, starting from two pairs of quality measures, we propose a technique for their combining in order to obtain a new pair. The associativity of combining operation allow us to mine association rules from distributed databases. We give an application in mining association rules from distributed WWW servers access logs.

2 Overview of the support and confidence measures

Let $\mathcal{D}$ be a non-empty database containing records described by theirs values binary attributes. We will denote by D_A the set of all transactions x in which A was purchased, i.e.

$$D_A = \{x \in \mathcal{D} | A(x) = 1\};$$

in this way D_A is a subset of $\mathcal{D}$. We denote by coD_A the set of transactions that not contain the item A, i.e

$$coD_A = \{x \in \mathcal{D} | A(x) = 0\}.$$

Definition 1 (Support). *The* support *of an association rule $A \rightharpoonup B$ is usually defined as:*

$$supp(A \rightharpoonup B) = \frac{|D_A \cap D_B|}{|\mathcal{D}|}$$

where $|D_A|$ is the number of transactions that contain A.

The classical confidence is usually defined as below:

Definition 2 (Confidence). *The* confidence *of an association rule $A \rightharpoonup B$ is usually defined as:*

$$conf(A \rightharpoonup B) = \frac{|D_A \cap D_B|}{|D_A|} \tag{1}$$

The need for more refined confidence measures appears frequently in practical applications. Therefore others confidence measures were proposed by Hullermeier [8]. He suggested a formal framework for the systematic derivation of quality measures which is based on the classification of stored data into examples of a rule, counterexamples of that rule, and irrelevant cases. The result is the definition of *classical n-confidence*:

$$conf_n(A \rightharpoonup B) = \frac{|D_A \cap D_B|}{|D_A \cap coD_B|)} \tag{2}$$

Later De Cook et al. in [3] defined *pessimistic (p) confidence* and *optimistic (o) confidence*:

$$conf_p(A \rightharpoonup B) = \frac{|D_A \cap D_B|)}{|coD_A \cup coD_B|} \tag{3}$$

$$conf_o(A \rightharpoonup B) = \frac{|coD_A \cup D_B|}{|D_A \cap coD_B|} \tag{4}$$

They introduce these measures in order to capture incompletion cases. For example in the case of the rule "if X buy bread then X buy also butter", when determining the pessimistic confidence of a rule we have the following assumption in mind: if those people who did not buy bread, would have bought bread, they would not have bought butter as well. For the optimistic confidence measure on the other hand we assume that if those people who did not buy bread, would have bought bread, they would have bought butter as well.

An interesting relation between these measures is the following:

Proposition 1. *(De Cock et al. [3])*

$$conf_p(A \rightharpoonup B) \leq conf_n(A \rightharpoonup B) \leq conf_o(A \rightharpoonup B)$$

3 Combining quality measures

In this section we will consider pairs of quality measures (M_*, M^*) that verify the following property:

$$M_* \leq M^*$$

i.e. quality measures verifying relationships such as one from Proposition 1.

According with the relations (2),(3), (4) and the proposition 1 we introduced in [9] and finalized in [10] a collection of such measures. For instance:

1. $(conf\_po_*, conf\_po^*)$

$$conf\_po_*(A \rightharpoonup B) = \frac{|D_A \cap D_B|^2}{|D_A \cap D_B|^2 + |coD_A \cup coD_B|^2}$$

$$conf\_po^*(A \rightharpoonup B) = \frac{|coD_A \cup D_B|^2}{|coD_A \cup D_B|^2 + |D_A \cap coD_B|^2}$$

2. $(conf\_pn_*, conf\_pn^*)$

$$conf\_pn_*(A \rightharpoonup B) = \frac{|D_A \cap D_B|^2}{|D_A \cap D_B|^2 + |D_A \cap coD_B| \cdot |coD_A \cup coD_B|}$$

$$conf\_pn^*(A \rightharpoonup B) = \frac{|D_A \cap D_B| \cdot |coD_A \cup D_B|}{|D_A \cap D_B| \cdot |coD_A \cup D_B| + |D_A \cap coD_B|^2}$$

are such measures.

By analogy with Dempster's rule for combining the degrees of belief and plausibility from evidence theory [14], for to combine two pairs of confidence

measures $(M_{1*}, {M_1}^*)$ and $(M_{2*}, {M_2}^*)$, with $M_{1*}, {M_1}^*, M_{2*}, {M_2}^* \in [0,1]$, we propose the following formula

$$\widetilde{M_*} = \frac{M_{1*}{M_2}^* + {M_1}^* M_{2*} - M_{1*}M_{2*}}{1 - M_{1*}(1 - {M_2}^*) - M_{2*}(1 - M_1^*)} \tag{5}$$

$$\widetilde{M^*} = \frac{M_1^* M_2^*}{1 - M_{1*}(1 - {M_2}^*) - M_{2*}(1 - M_1^*)} \tag{6}$$

The measures introduced by (5) and (6) define an associative[1] composition operation (i.e. a combination rule) $\circ$:

$$(M_{1*}, {M_1}^*) \circ (M_{2*}, {M_2}^*) = (\widetilde{M_*}, \widetilde{M^*})$$

In addition the following relations hold:

$$\widetilde{M_*} \leq \widetilde{M^*} \text{ and } \widetilde{M^*} - \widetilde{M_*} < \min\{{M_1}^* - M_{1*}, {M_2}^* - M_{2*}\}.$$

Therefore, from a pair (M_1, M_2), $M1 \leq M_2$, we can define a single measure:

$$M = \lambda_1 \cdot M_1 + \lambda_2 \cdot M_2, \text{ with } \lambda_1 + \lambda_2 = 1$$

where λ_1 and λ_2 represent the belief degree in M_1 respectively in M_2.

Example 1. Consider the following sample transaction database presented in Table 1. The $\widetilde{M_*}$ and $\widetilde{M^*}$ are computed for the pairs $(conf\_po_*, conf\_po^*)$ and

TID	I_1, I_2, I_3, I_4, I_5	TID	I_1, I_2, I_3, I_4, I_5	TID	I_1, I_2, I_3, I_4, I_5
T_1	1, 0, 1, 0, 0	T_7	0, 0, 1, 1, 1	T_{13}	1, 1, 1, 0, 1
T_2	1, 0, 1, 0, 0	T_8	1, 0, 0, 1, 0	T_{14}	0, 1, 1, 0, 1
T_3	0, 0, 1, 0, 1	T_9	0, 1, 0, 1, 1	T_{15}	1, 0, 1, 0, 1
T_4	1, 1, 0, 1, 0	T_{10}	0, 1, 1, 1, 0	T_{16}	0, 1, 1, 0, 0
T_5	1, 1, 1, 0, 1	T_{11}	1, 0, 0, 1, 1	T_{17}	1, 1, 0, 1, 0
T_6	1, 0, 0, 0, 1	T_{12}	1, 1, 0, 0, 1	T_{18}	1, 0, 1, 1, 1

Table 1. A database sample

$(conf\_pn_*, conf\_pn^*)$ and The table 2 shows the values obtained for quality measures above presented.

4 Web Log mining use case

The Web Log mining use case was widely studied in the context of adaptive web sites and services. A number of papers such as [5], [13] and [12] address

[1] This permits to combine an arbitrary number of measures

$Rule$	$conf\_po_*$	$conf\_po^*$	$conf\_pn_*$	$conf\_pn^*$	$\widetilde{M}_*$	$\widetilde{M}^*$
$I_1 \rightharpoonup I_3$	0,200	0,800	0,077	0,667	0,196	0,581
$I_2 \rightharpoonup I_3$	0,129	0,925	0,088	0,814	0,180	0,776
$I_4 \rightharpoonup I_1$	0,129	0,962	0,114	0,893	0,214	0,874
$I_3, I_5 \rightharpoonup I_1$	0,075	0,962	0,087	0,870	0,145	0,847
$I_1, I_4 \rightharpoonup I_3, I_5$	0,003	0,925	0,014	0,467	0,015	0,433
$I_1 \rightharpoonup I_4, I_5$	0,015	0,390	0,012	0,138	0,007	0,055

Table 2. Values for quality measures $\widetilde{M}_*$, $\widetilde{M}^*$

this scenario with different techniques. In this section we use the same scenario by addressing the quality measures combination issues i.e. by using proposed measures (5) and (6).

One natural assumption is that the server access logs contain users requested files into a chronologically order. The Table 3 shows a fragment of Apache Web server.

```
220.226.35.198 - - [23/Feb/2007:15:54:10]
         "GET /products/product-A.html HTTP/1.1" 200 12482
220.226.35.198 - - [23/Feb/2007:15:54:12]
         "GET /products/images/A.jpg HTTP/1.1" 200 12482
216.122.48.128 - - [23/Feb/2007:18:43:51]
         "GET /library.html HTTP/1.1" 200 42227
220.226.35.198 - - [23/Feb/2007:15:54:10]
         "GET /products/product-B.html HTTP/1.1" 200 15482
172.30.110.140 - - [23/Feb/2007:21:28:01]
         "GET / HTTP/1.0" 200 23903
195.113.25.239 - - [23/Feb/2007:22:46:55]
         "GET /html/d711.html HTTP/1.1" 200 4513
```

Table 3. A fragment of Apache Web Log

Typically, these web server logs contain millions of records therefore they are appropriate to mine association rules and then use them in the context of web sites personalization. Another argument in the favor of using association rules is that the logs are never complete and exhaustive therefore the conclusions cannot be derived using standard first order logic. Technically, for each visit, web log records the remote users host name or IP address, the time when the request arrived, the HTTP method (GET, POST, etc.) the remote user used, the URL of the visited web document, the status code of the HTTP response (for example: 200 for successful access, 404 for file not found), and the number of bytes returned to the user (in most cases is the document size).

For a given web server log the mining process start with a preprocessing row data. The fist step is a clean process [11]. When a web page is requested by user then automatically all images, clips embedded in page are recorded in web access log. These records can be removed because are irrelevantly for mining process.

The next step is to extract user sessions from web logs. A user session is a relatively independent sequence of web requests from the same user. We consider web log data as a sequence of distinct web pages, where subsequences, such as user sessions can be observed within unusually long gaps between consecutive requests.

For example, assume that the web log consists of the following user visit sequence:

User IP	Date	Requested Page	User IP	Date	Requested Page
IP_1	00:00:00	A	IP_1	00:00:19	E
IP_2	00:00:05	B	IP_2	00:00:25	A
IP_3	00:00:10	D	IP_1	01:00:19	E
IP_1	00:00:15	C	IP_1	01:02:10	A

Table 4. Visited page sequence

User IP	Session	User IP	Session
IP_1	A, C, E	IP_3	D
IP_2	B, A	IP_1	E, A

Table 5. User sessions

The Table 5 shows users sessions according to user IP's.

Using association rule discovery techniques we can find correlations such as:

50% of clients who accessed the Web page with URL `/products/product-A.html` also accessed `/products/product-B.html`

or

35% of clients who accessed `/special-offers.html` placed an online order in `/products/product-C.html`

Discovery of such rules for online shops can help in the development of effective marketing strategies. Also, association rules discovered from WWW access logs can give an indication of how to better organize the organization's Web space.

For example, if one discovers that

80% of the clients accessing `/products` and `/products/product-A.html` also accessed `/products/product-B.html`,

but

only 30% of those who accessed `/products` also accessed `/products/product-B.html`, then it is likely that some information in `product-A.html` leads clients to access `product-B.html`.

This correlation might suggest that this information should be moved to a higher level (e.g., `/company/products`) to increase access to `/products/product-B.html`.

At this date the majority of WWW serves have mirrors for reducing the pages access time and avoid their overloading. In this situation the concern is for determining association rules for the web pages. A simple and coarse solution would be to keep all the log data from these servers in a single place and to apply a standard association rules determination algorithm. This approach is inefficient due to the large size that these logs can have.

Our proposal is: Let suppose we have two servers $S1$ and $S2$. By applying the standard association rules determination algorithm we can find, for every server, a set of association rules between pages together with their attached confidence pairs $(P_{1*}, {P_1}^*)$ and $(P_{2*}, {P_2}^*)$, respectively. On this basis we can compute a common set of rules by combining the $S1$ and $S2$ rules according to (5) and (6) formulas.

5 Conclusions

In this paper we propose a combining technique of two pairs of measures. This combining technique is useful for evaluate quality of association rules that rise from different sources. The obtained results can be extended to fuzzy association rules, similarly with papers [3], [4], [6]. Potential applications can be used by encoding association rules into a web rule language such as RuleML [1], [17] or R2ML [15], [16] which will permit a web distributed architecture of association rules repositories covering different subjects. Builders of portals as well as commercial web sites may use these rules to create user-specific content. Some efforts on that are already done but specifically in the Semantic Web area [7] and their main criticism is that they adopt a non-distributed solution of user profiles.

References

[1] **H. Boley, S. Tabet, G. Wagner**:- Design Rationale of RuleML: A Markup Language for Semantic Web Rules, Proc. SWWS'01, Stanford, July/August 2001

[2] **R. Agrawal, R. Srikant**:- Fast Algorithms for Mining Association Rules, In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, Santiago, Chile, 1994, 487-499

[3] **M. De Cock, C. Cornelis, E. E. Kerre**:- Fuzzy Association Rules: a Two-Sided Approach, In: Y. Liu, G. Chen, K. Y. Cai (Eds.): Proceedings of FIP2003 (International Conference on Fuzzy Information processing: Theories and Applications), Tsinghua Univ. Press, 2003, 385-390

[4] **M. De Cock, C. Cornelis, E.E. Kerre**:- A clear view on quality measures for fuzzy association rules, In: Proceedings of Int. Conf. on Fuzzy Sets and Soft Computing in Economics and Finance, St. Petersburg, 2004, 54-61

[5] **A. Demiriz**:- webSPADE: A Parallel Sequence Mining Algorithm to Analyze Web Log Data, In: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), Maebashi City, Japan, December 2002, 755-758

[6] **D. Dubois, E. Hüllermeier, H. Prade**:- A Note on Quality Measures for Fuzzy Association Rules. In: Proceedings IFSA-03, 10th International Fuzzy Systems Association World Congress, Lecture Notes in Artificial Intelligence, number 2715, Springer-Verlag, 2003, 346-353

[7] **N. Henze**:- Personal Readers: Personalized Learning Object Readers for the Semantic Web, 12th International Conference on Artificial Intelligence in Education, AIED05, Amsterdam, July 22, 2005

[8] **E. Hüllermeier**:- Fuzzy Association Rules: Semantic Issues and Quality Measures, In: B. Reusch (Ed.), Proceedings of the International Conference, 7th Fuzzy Days on Computational Intelligence, Theory and Applications, Springer-Verlag, Lecture notes in Computer Science, number 2606, 2001, 380-391

[9] **I. Iancu, M. Gabroveanu, A. Giurca**:- A Pair of Confidence Measures for Association Rules, In: A. Wnuk, V. Koppen, A. Omelchenko (Eds), Proc. of 30th Annual Conference of the German Classification Society (GfKl) - Advances in Data Analysis, March 8-10, 2006, Berlin, Germany, Freie Universitat Berlin, 93 (short paper)

[10] **I. Iancu, M. Gabroveanu, A. Giurca**:- A General Quality Measure for Association Rules, Annals of the University of Bucharest, Mathematics and Computer Science Series, vol. LV, 2006, ISSN 77179

[11] **Tianyi Li**:- Web-Document Prediction and Presending Using Association Rule Sequential Classifiers, Masters Thesis, 2001

[12] **Zhiyong Lu, Yiyu Yao, Ning Zhong**:- Web Intelligence, chapter Web Log Mining, Springer, Berlin Heidelberg New York, 2003, 173-194

[13] **M. Sayal, P. Scheuermann**:- Distributed Web Log Mining Using Maximal Large Itemsets, Knowledge and Information Systems, 3(4), 2001, 389-404

[14] **G. Shafer**:- A Mathematical Theory of Evidence, Princeton University Press, 1976

[15] **G. Wagner, A. Giurca, S. Lukichev**:- R2ML: A General Approach for Marking up Rules, Dagstuhl Seminar Proceedings 05371, In: F. Bry, F. Fages, M. Marchiori, H. Ohlbach (Eds.), Principles and Practices of Semantic Web Reasoning, ISSN:1862-4405, `http://drops.dagstuhl.de/opus/volltexte/2006/479/pdf/05371.GiurcaAdrian.Paper.479.pdf`

[16] **G. Wagner, A. Giurca, S. Lukichev**:- A Usable Interchange Format for Rich Syntax Rules Integrating OCL, RuleML and SWRL, In: Proceedings of Reasoning on the Web 2006, RoW2006, May 22, 2006, Edinburgh, Scotland, `http://www.aifb.uni-karlsruhe.de/WBS/phi/RoW06/procs/wagner.pdf`

[17] **G. Wagner**:- How to Design a General Rule Markup Language?, Invited Talk, Workshop XML Technologien fr das Semantic Web (XSW 2002), Berlin, June 2002, Lecture Notes in Informatics, Gesellschaft f. Informatik, `http://oxygen.informatik.tu-cottbus.de/rewerse-i1/GRML.pdf`