

Closures and Partial Implications in Educational Data Mining

Diego García-Saiz¹, Marta Zorrilla¹, and José L. Balcázar²

¹ Mathematics, Statistics and Computation Department, University of Cantabria
Avda. de los Castros s/n, Santander, Spain

garciasad@unican.es marta.zorrilla@unican.es

² LSI Department, UPC, Campus Nord, Barcelona
jose.luis.balcazar@upc.edu

Abstract. Educational Data Mining (EDM) is a growing field of use of data analysis techniques. Specifically, we consider partial implications. The main problems are, first, that a support threshold is absolutely necessary but setting it “right” is extremely difficult; and, second, that, very often, large amounts of partial implications are found, beyond what an EDM user would be able to manually inspect. Our program *yacaree*, recently developed, is an associator that tackles both problems. In an EDM context, our program has demonstrated to be competitive with respect to the amount of partial implications output. But “finding few rules” is not the same as “finding the right rules”. We extend the evaluation with a deeper quantitative analysis and a subjective evaluation on EDM datasets, eliciting the opinion of the instructors of the courses under analysis to assess the pertinence of the rules found by different association miners.

Keywords: Closure Lattices, Partial Implications, Association Rules

1 Introduction

Education is evolving at all levels since the appearance of e-learning environments: Learning Content Management Systems (LCMS), Intelligent Tutoring Systems, or Adaptive Educational Hypermedia Systems. These systems log all the activity carried out by students and instructors, and this raw data, adequately analyzed, might help instructors to obtain a better understanding of the students and of their learning processes. In remote learning, instructors may never see their students in person. Data analysis techniques could help them to detect problems (lack of motivation, under-performance, drop-out. . .) and, possibly, to take action. Yet, unless the course itself is on data mining, it is unlikely that the instructors know much about data mining techniques. If we want to help teachers of, say, philology or law, we need to work out data mining tools that do not require much tuning or technical understanding.

Here we focus on the particular case of mining partial implications [1] (a relaxed form of implication analysis in concept lattices [2]), and their close relatives: association rules [3]. Most of the available algorithms depend on one or more parameters whose value is to be set by the user, and whose semantics are unlikely to be easy to understand by teachers of other disciplines.

We have explored the output of five association algorithms on datasets from educational sources, and evaluated not only the amounts of partial implications found but also the subjective pertinency of the rules obtained. For this last task we kept close cooperation with the end user, namely, the teachers of the online courses from which the datasets were obtained. Our conclusions are in the form of strengths and weaknesses of each of the five algorithms compared.

One of the algorithms participating in the evaluation was a contribution of our group, demonstrated at [4] and described in more detail in [5]: the *yacaree* association miner. This associator extracts partial implications from the “iceberg” (frequent part of the) FCA lattice [6]; it attempts at offering a more user-friendly, parameter-less interface, through self-tuning the support threshold and a threshold on a relative form of confidence studied in [7]: the closure-based confidence boost.

In [8], a two-page poster publication, we have provided a preliminary initial description of this study, containing only the quantitative analysis (a part of Table 2 below) but using a version of *yacaree* which did not report yet rules of confidence 100%. This paper extends it largely with further quantitative analyses and a qualitative, user-based, subjective evaluation of the usefulness of the resulting rules. The main question to study is whether a price, in terms of usefulness of the output for the end user, was being paid for the parameter-less interface. Any parameter-free alternative should stand a comparison of its output with that of other, “expert”-oriented algorithms, to clarify whether, for the subjective perception of the teacher, the outcome does make sense and results useful. Actually, our main conclusion is that they do, and that, developed according to our strategy, a self-tuning associator is able to provide sensible quantities of partial implications that result useful and informative to the end user.

1.1 Related work

In the educational context, data mining techniques are used in order to understand learner behaviour [9], to recommend activities or topics [10], to offer learning experiences [11] or to provide instructional messages to learners [12] with the aim of improving the effectiveness of the course, promoting group-based collaborative learning [13], or even predicting students’ performance [9]. Two interesting papers which detail and summarize the application of data mining to educational systems are [14] and [15].

The FCA community has also contributed in this arena. We must name Romashkin et al. [16] who used closed sets of students and their marks to reveal some interesting patterns and implications in student assessment data, especially to trace dynamic; and Ignatov et al. [17] who showed that FCA taxonomies are a useful tool for representing object-attribute data which helps to reveal some

frequent patterns and to present dependencies in data entirely at a certain level of details. They carried out the analysis of university applications to the Higher School of Economics as case study. Another interesting work in this research line was previously carried out by Belohlávek et al. [18] in order to evaluate questionnaires.

In the particular case of the association rules technique, we find works such as [19] in which association rules are used to find mistakes often made together while students solve exercises in propositional logic, [20] where rules are used to discover the tools which virtual students employ frequently together during their learning sessions, and [21] where association rules and collaborative filtering are used inside an architecture for making recommendations in courseware.

However, association rule algorithms still have some drawbacks, as analyzed in [22]: mainly, first, as most often the instructors are not data mining experts, the decisions about setting to useful values the parameters of the algorithms present difficulties. Then, a second difficulty is the large number of rules often obtained as output, most of which are redundant and non-interesting for decision making and, in many occasions, exhibit low understandability. The authors of [22] offer some solutions although none of them is automatized or gathered in an algorithm. For example, they propose to use Predictive Apriori, rather than the implementation of Apriori in Weka [23], since it only requires one parameter which is the number of rules that the user wants to obtain. In [24], it is argued that cosine and added value (or equivalently lift) are well suited to educational data, and that instructors can interpret their results easily. In our opinion, these measures lack actionability since they are symmetric, which reduces the use of the rules in decision making tasks. Orientation is a crucial and very suggestive property of association rules and partial implications, and we consider that it must be preserved in an effective but asymmetric measure, as close as possible to confidence. Many measures of intensity of implication are described e.g. in [25],[26].

2 Case Studies

This section contains our major contributions: we compare the output of five well-known association rule miners on five educational datasets and evaluate the subjective pertinency of the rules obtained in close cooperation with the teachers involved in the two virtual courses analyzed.

2.1 Association rule miners

There is a long list of association rule miners; large sets of references and surveys appear e.g. in [http://michael.hahsler.net/research/bib/association rules/](http://michael.hahsler.net/research/bib/association%20rules/) and in all main Data Mining reference works. Among them, we have chosen the following algorithms for our comparison: the implementation of Apriori by Borgelt [27], the implementation of Apriori in the Weka package [23], the Predictive Apriori

implementation in Weka [28], the implementation of ChARM [29] available in the Coron System [30], and our own closure-lattice-based associator *yacaree* [4].

The implementation of Apriori by Borgelt [27] is a representative of the standard usage of association rules in data mining, as per [3], particularly in the way support and confidence parameters are handled, as well as in the restriction to association rules with a single item in the consequent. In this fully standard approach, first, one constructs all frequent sets, and then each item in each frequent set is tried as consequent with the rest of the frequent itemset as antecedent, and the confidence of the rule evaluated; the rule is reported if its confidence is high enough. This implementation is amazingly well streamlined for speed. It offers, additionally, an ample repertory of additional evaluation measures (lift, normalized chi-square...), and we must warn that a specific flag must be set (as we did, “-o”) so that support is computed accordingly with the notion of support in other tools.

Weka is one of the oldest and most extended open-source data mining suites, and all implementations there are widely used. The implementation of Apriori in the Weka package is similar to the one just described, employing confidence and support constraints; it departs slightly from [3], though. First, the rules generated can have more than one item in the consequent. Also, instead of fixing the support at the given threshold at once, the user is requested to indicate a number of rules and a “delta” parameter. Then, support is set initially at 100% and iteratively reduced by “delta” until either the support threshold is reached or the requested number of rules is collected.

The Predictive Apriori implementation in Weka follows [28]. The advantage of this algorithm is that it only requires from the user to set the number of rules to be discovered, which is appropriate for users that are not data mining experts, provided that, in some sense, “the right rules” are found. The algorithm automatically attempts at balancing optimally support and confidence on the basis of Bayesian criteria related to the so-called expected predictive accuracy. A disadvantage of this method is that it often requires longer running times than the previous ones.

These three implementations construct partial implications on the basis of all frequent itemsets. Our other two systems work on the basis of frequent closures, which allow one to know the support of any frequent itemset without storing all of them. The Coron system [30] offers several implementations of different closed-set-based algorithms. These methods return the same set of closure-based partial implications, although they compute them in different ways. We have used ChARM [29], but the specific method is not relevant here because we do not include yet running times in our evaluation: we concentrate on the usefulness of the output.

The fifth implementation is our own association miner *yacaree* [4]. Like ChARM, it is based on closures, and allows for several items in the consequent of the partial implications. In the partial implications output by this system, both antecedent and total set of items in each rule will be closed sets. The currently most recent version 1.2.0 is the first to report rules of confidence 100%.

First, it constructs the Closure Lattice up to a support bound that is adjusted autonomously during the run, on the basis of the technological limitations, so that the user does not need to select it. Second, it constructs a basis of partial implications out of these closures. Third, it filters the partial implications along the way, on the basis of the closure-based confidence boost [7], whereby the confidence of an association rule is compared to that of other similar rules: a rule must offer a clear improvement on similar ones to be considered useful.

2.2 Datasets

For the case studies, we used the data from two courses offered in the University of Cantabria. Both courses are eminently practical. The first one, entitled “Introduction to multimedia methods”, has the objective of teaching the students how to use a particular multimedia tool (in what follows, we refer to it as the multimedia dataset) and the second one, “Basic administration of a UNIX-LINUX system” (the Linux dataset) teaches the students the basic utilities and tools to install and configure correctly a LINUX operating system.

The multimedia course is designed by means of web pages and includes some video tutorials, flash animations and interactive elements. The students must perform 4 exercises, 2 projects and one final exam online. The course is open to all degrees and the number of students enrolled was 79.

Unlike the multimedia course, the Linux course only allows 24 students to be enrolled, all of them from a telecommunications degree. All materials of the course are available since the first day of the course. Furthermore, the contents of a previous edition of the course is also offered in pdf; these files have the advantage that they can be kept locally and used for study in case any technical problem would prevent access to the updated files, but do not include all the contents of the present edition. Additionally, during the course, the students must deliver 6 practical exercises and pass two online exams. The course includes 38 self-tests, one for each topic of the course. The instructor indicates the topics and self-tests that they must perform every week on the calendar.

We worked with five datasets. The first one, “linux materials”, gathers the access logs to materials prepared by the instructor (html pages, pdf files, tests, and so on) as used by each student in each learning session of the Linux course. The datasets “linux resources” and “multimedia resources” are the session-wise log of the resources and tools used by each student in each learning session (assessment, content-pages, forum, and so on). It was immediately apparent that, in these datasets, one specific resource led to some “noise”: the “organizer” resource acts as front page of most sessions (near 84% in Linux and 85% in multimedia, as the only other alternative is the access through the forum) and hence it appears in many rules and creates many variants, mostly of low information contents. Thus, we prepared two datasets, named “linux resources reduced” and “multimedia resources reduced” respectively, which are identical to the second and third dataset, except that the “organizer” resource is fully removed. The number of different items and transactions of each dataset is shown in Table 1. For the

sake of better understanding, we show a diagram of the intents of the concept lattice of the linux dataset above 13% support in Fig. 1.

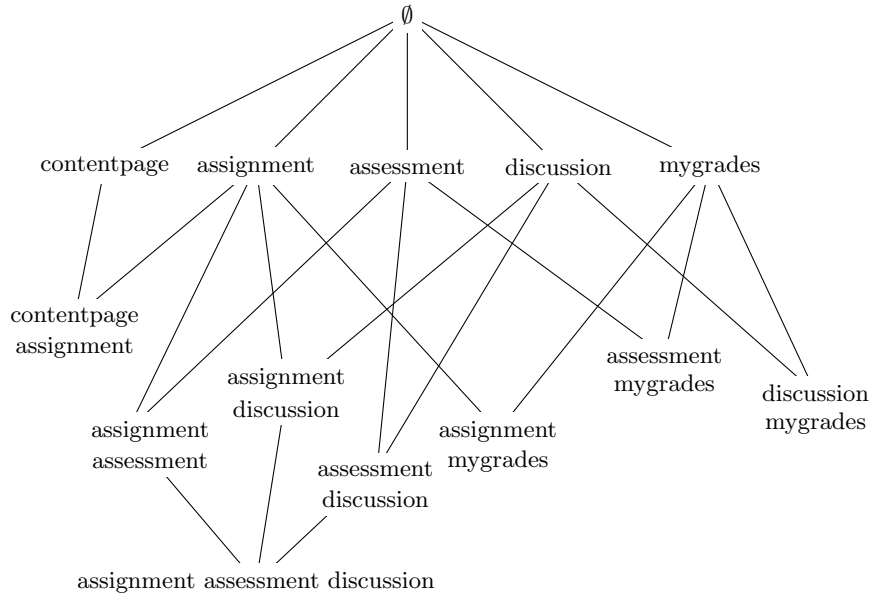


Fig. 1. Intents of at least 13% support.

Table 1. Datasets description

Name	Transactions	Items
Dataset1 (linux materials)	407	22
Dataset1 (linux resources)	2486	27
Dataset2 (linux resources reduced)	2346	26
Dataset4 (multimedia resources)	5892	27
Dataset5 (multimedia resources reduced)	5643	26

2.3 Datasets results

With the aim of comparing several association programs, one difficulty is always the setting of the parameters, particularly the support, as the value chosen might favor one particular algorithm in larger degree. In our case, there is an extra level of difficulty, as one of the participating algorithms, *yacaree*, self-tunes the

support on itself. In order to find fair comparison grounds, we performed a brief preprocessing.

Running on one of the “Linux resources” dataset, *yacaree* took about four minutes (a bit long for a non-expert to wait) and delved down to 0.02% support; however, for this low threshold, both Weka alternatives were substantially worse (Predictive Apriori took 40 minutes and Apriori led to overflow even when given 2GB of memory). Similar facts happened for the other datasets.

Given this information, we decided to fix at 1% the support threshold for all the computations, and at 66% the confidence threshold (initial value set up by *yacaree*). In all the runs, we left unbounded, or, in the case of Weka tools, we set very high (10000) the number of rules to be found, even if this meant overriding their default value for this quantity. We show the number of rules obtained utilizing the different algorithms on our datasets in Table 2. The entries marked “—” on the table are cases where the corresponding algorithm was unable to complete in 6 hours.

Table 2. Number of rules obtained on our datasets with the five algorithms

Dataset	Number of rules s=1% c=66%				
	Weka Apriori	Predictive Apriori	Borgelt Apriori	ChARM	<i>yacaree</i>
Dataset1 (linux materials)	2272	1730	524	366	40
Dataset2 (linux resources)	7523	over 10000	3751	5610	255
Dataset3 (linux resources reduced)	4249	over 10000	1876	2586	93
Dataset4 (multimedia resources)	1442	—	1023	1427	182
Dataset5 (multimedia resources reduced)	488	—	404	469	46

Results from “resources reduced” datasets If we analyze the results obtained with Apriori from Weka, we can see that the number of rules is unmanageable, e.g. 4249 rules for Linux resources reduced dataset. The first 243 are implications of full confidence, 100%, low support, and high redundancy: see rules 2 and 3 and 235 and 236 and the followings in Table 3. Had we used the tool’s default settings of the parameters, we would have found essentially no information. The same happens with multimedia dataset (we do not show the table for space reasons).

The analysis of the results obtained from Predictive Apriori is very costly, as it generates as many rules as we allow it to. With 10000 rules required, they are obtained on dataset2 and dataset3 waiting for more than 20 minutes, and the accuracy is still high, so that many further rules could be obtained. If we restrict ourselves to the first few rules returned, they turn out to offer a very low support and quite some redundancy (see Table 4).

The output offered by Borgelt’s implementation presents a large number of rules: 1876 and 404 rules in Linux and multimedia reduced datasets respectively,

Table 3. Subset of association rules obtained with Apriori from Weka on the "Linux resources reduced" dataset

No.	Association rule	(Sup., Conf.)
2	announcement tracking \Rightarrow assessment	(1.7, 100)
3	announcement mygrades tracking \Rightarrow assessment	(1.6, 100)
235	assignments calendar contentpage discussion medialibrary syllabus \Rightarrow assessment	(1.0, 100)
236	assessment calendar contentpage discussion medialibrary syllabus \Rightarrow assignments	(1.0, 100)
2523	announcement assessment calendar syllabus \Rightarrow assignments contentpage	(1.2, 78.0)
2524	announcement assessment calendar syllabus \Rightarrow assignments discussion	(1.2, 78.0)
2530	announcement calendar mail \Rightarrow contentpage	(1.0, 78.0)
2534	announcement assignments calendar chat \Rightarrow contentpage	(1.0, 78.0)

Table 4. Subset of association rules obtained with Predictive Apriori from Weka on the "linux resources reduced" dataset

No.	Association rule	(Support, Accuracy)
122	assignments calendar search \Rightarrow syllabus	(0.85, 0.95439)
123	assignments chat weblinks \Rightarrow assessment syllabus	(0.85, 0.95439)
124	assignments chat weblinks \Rightarrow discussion syllabus	(0.85, 0.95439)
125	assignments discussion search \Rightarrow assessment syllabus	(0.85, 0.95439)

of which 141 and 2 are implications. Coming up with specific conclusions becomes harder. The rules tend to be small, exhibit high redundancy and involve low-support tools that are almost never used, so that they offer little interest to the instructor. As shown in Table 5, where the rules 11, 12, 13 differ slightly from the rules 99, 100 and 101 which contain the announcement tool in the antecedent with a very low support and similar confidence.

Table 5. Subset of association rules obtained with Borgelt's apriori implementation on the "linux resources reduced" dataset

No.	Association rule	(Supp. , Conf.)
11	chat \Rightarrow discussion	(3.7, 84.9)
12	chat \Rightarrow assignments	(3.7, 75.6)
13	chat \Rightarrow assessment	(3.7, 81.4)
99	chat announcement \Rightarrow discussion	(2.0, 84.8)
100	chat announcement \Rightarrow assignments	(2.0, 87.0)
101	chat announcement \Rightarrow assessment	(2.0, 93.5)

ChARM returns a higher number of rules, 2586 and 469 with 193 and 2 implications in Linux and multimedia resources reduced datasets respectively.

As in previous cases, the rules also present high redundancy (see rules 3 to 6 and 7 and 8 in Table 6 and rules 10,11,12 and 31,32,33 in Table 7).

Table 6. Subset of association rules obtained with ChARM on the “linux resources reduced” dataset

No.	Association rule	(Supp. , Conf.)
3	announcement, contentpage, medialibrary, syllabus \Rightarrow assessment	(1.02, 96.00)
4	announcement, assessment, medialibrary, syllabus \Rightarrow contentpage	(1.02, 88.89)
5	announcement, assessment, contentpage, medialibrary \Rightarrow syllabus	(1.02, 70.59)
6	announcement, medialibrary, syllabus \Rightarrow assessment, contentpage	(1.02, 82.76)
7	announcement, medialibrary, syllabus \Rightarrow contentpage	(1.07, 86.21)
8	announcement, contentpage, medialibrary \Rightarrow syllabus	(1.07, 67.57)

Table 7. Subset of association rules obtained with ChARM algorithm on the “multi-media resources reduced” dataset

No.	Association rule	(Supp. , Conf.)
10	chat, contentpage, discussion \Rightarrow assessment	(1.13, 81.01)
11	assessment, chat contentpage \Rightarrow discussion	(1.13, 94.12)
12	chat, contentpage \Rightarrow assessment, discussion	(1.13, 71.91)
31	contentpage, discussion, syllabus, \Rightarrow assessment	(1.12, 84.00)
32	assessment, discussion, syllabus, \Rightarrow contentpage	(1.12, 66.32)
33	assessment, contentpage, syllabus, \Rightarrow discussion	(1.12, 79.75)

Despite the fact that the number of rules obtained with *yacaree* on reduced resources datasets is a bit high, 93 for dataset3 and 46 for dataset5, it is possible to discover the resources which students use frequently together in each learning session and, at the same time, the kind of sessions which they perform. It is remarkable the reduction in the number of rules due to the use of confidence boost parameter. A subset of the most relevant rules obtained with *yacaree* on Linux resources reduced dataset is shown in Table 8. However, there appear as well quite a few trivial and non-interesting rules for the instructor. For instance, rule 1 is trivial because it is obvious that to send a task is necessary to use the file manager tool. The rules 6, 18 and 19 do not offer new information to the instructor given that he uses the forum in order to establish the date of the exams. So that these kind of sessions are known to the instructor. The rules 7, 12, 36 and 50 gather sessions in which students want to know specific dates: deadlines for tasks or assessments, exam dates. Rule 16 indicates quite a few sessions in which the students are interested in knowing their progress, and rules 8 and 10 gather the study sessions in which the students combine reading of content pages with tackling self-tests.

Table 9 depicts a subset of the most relevant rules obtained with *yacaree* on multimedia resources reduced dataset. As in the previous result, there are

Table 8. Subset of association rules obtained with *yacaree* on the “linux resources reduced” dataset

No.	Association rule	(Supp., Conf., Lift, Cboost)
1	filemanager \Rightarrow assignments	(4.6, 93.9, 1.908, 1.908)
6	discussion whoisonline \Rightarrow assessment	(3.0, 75.5, 1.648, 1.379)
18	discussion mail \Rightarrow assessment	(3.2, 72.1, 1.574, 1.268)
19	announcement mail \Rightarrow assessment discussion	(1.6, 80.9, 3.381, 1.267)
7	announcement \Rightarrow assessment	(7.6, 88.1, 1.923, 1.369)
12	calendar \Rightarrow assessment	(9.1, 75.9, 1.656, 1.337)
36	calendar \Rightarrow assignments	(8.1, 67.0, 1.362, 1.219)
50	announcement calendar \Rightarrow assessment assignments	(2.6, 77.2, 2.941, 1.200)
16	tracking \Rightarrow mygrades	(6.8, 80.3, 2.409, 1.272)
8	contentpage mygrades \Rightarrow assessment	(3.8, 84.8, 1.850, 1.369)
10	contentpage discussion \Rightarrow assessment	(7.3, 75.1, 1.639, 1.339)

some trivial and non-interesting rules for the instructor. For example, rule 1 already explained, and rule 2 and 40 which gather sessions in which students wanted to know specific dates for assignments. Instead, other rules as rule 7, 14 and 36 allowed the teacher to discover the students visited the content pages and the forum in working sessions with the aim at solving problems or doubts in the resolution of the tasks. Furthermore, she was happy when observed that learning objectives tool was used while studying the contents (rule 3). This means that students played the videotutorials which she had recorded with great effort. Additionally, rule 4 informed her about the joint use of contents and weblinks tools. This last one contains the links to downloadable material. This reinforced her idea that the material should be presented in both formats, online and downloadable.

Table 9. Subset of association rules obtained with *yacaree* on the “multimedia resources reduced” dataset

No.	Association rule	(Supp., Conf., Lift, Cboost)
1	filemanager \Rightarrow assignments	(5.1, 71.5, 1.871, 1.871)
2	calendar \Rightarrow assignments	(6.1, 74.9, 1.961, 1.610)
40	announcement \Rightarrow assignments	(3.9, 67.2, 1.759, 1.153)
3	weblinks \Rightarrow contentpage	(3.7, 78.2, 2.105, 1.588)
4	learningobjectives \Rightarrow contentpage	(4.5, 81.4, 2.192, 1.530)
7	contentpage mygrades \Rightarrow assignments	(2.7, 66.7, 1.746, 1.421)
14	assignments whoisonline \Rightarrow discussion	(1.7, 72.5, 1.612, 1.301)
36	discussion weblinks \Rightarrow assignments	(1.9, 73.4, 1.923, 1.180)

Results from “resources” datasets, not reduced From the point of view of a virtual course instructor who is not an expert in Data Mining, the decision

of removing the “organizer” item from the “resources” dataset is debatable. This would be rather an action typical of a Data Mining expert. We consider that it was appropriate to do it, as the designers of the e-learning platform could easily predict that this “organizer” item was to be extremely frequent, and thus the option of discarding it could be incorporated by design into a set of related data mining tools ahead of time. However, we briefly discuss now what happens if one works with the complete “resources” dataset.

With *yacaree* we obtain 255 and 182 rules in dataset2 and dataset4 respectively. In both cases, one of them indicates that “organizer” is used in near 84% and 85% of the sessions respectively (see Tables 10 and 11). For this format of rule, with empty antecedent, support and confidence clearly must coincide. Essentially, the output of *yacaree* is not that different from the previous cases: many rules from the previous analysis reappear now in pairs, once with “organizer” and once without; when such a pair appears, the rule having “organizer” may look sometimes redundant, but its confidence boost value shows that it has high enough confidence so as to make it nonredundant (see Tables 10 and 11).

Table 10. Subset of association rules obtained with *yacaree* on the “Linux resources” dataset

No.	Association rule	(Supp., Conf., Lift, Cboost)
2	\Rightarrow organizer	(83.9, 83.9, 1.000, 1.982)
158	mygrades tracking \Rightarrow assessment organizer	(4.6, 71.7, 1.888, 1.109)
287	mygrades tracking \Rightarrow assessment	(5.0, 78.6, 1.818, 1.096)

Table 11. Subset of association rules obtained with *yacaree* on the “multimedia resources” dataset

No.	Association rule	(Supp., Conf., Lift, Cboost)
1	\Rightarrow organizer	(84.9, 84.9, 1.000, 2.421)
9	chat \Rightarrow discussion organizer	(2.0, 77.6, 2.324, 1.283)
113	chat \Rightarrow discussion	(2.2, 84.2, 1.954, 1.085)

The extra effort to be spent on the *yacaree* output is not that high compared with the alternative algorithms. ChARM and Borgelt’s Apriori runs into the same difficulties indicated for the reduced datasets, increased by the fact that the number of rules is, with ChARM, 5610 in dataset2 and 1427 in dataset4, and with Borgelt, 3751 in dataset2 and 1023 in dataset4, which include a considerable number of rules whose only consequent is “organizer”. Intuitively, all of them are pointing out to the fact that this item is so prevalent. Similarly, Weka Apriori obtains over 7000 rules in dataset2 and 1442 in dataset4, of which the first 568 are implications of 100% confidence, 474 of which are again rules that only have “organizer” as consequent. Predictive Apriori, beyond taking 45 minutes

to complete, also generates a large amount of rules (which we limited to 10000 again); and again the first ones have as single consequent “organizer”, and the next ones are long rules of very low support.

Results from the “linux materials” dataset We show in the Table 12 some of the most relevant rules among the 40 rules, of which 16 are implications of confidence 100%, selected by *yacaree* on this dataset. Such a limited output size allows for easy inspection by the instructor.

Table 12. Subset of association rules obtained with *yacaree* on the “materials” dataset

No.	Association rule	(Supp., Conf., Lift, Cboost)
1	topic6 \Rightarrow topic-pdf	(13.3, 1.0, 2.544, 2.544)
2	topic7 \Rightarrow topic-pdf	(9.8, 1.0, 2.544, 2.500)
3	topic4 topic-pdf \Rightarrow topic5	(6.4, 76.5, 5.764, 2.266)
18	topic1 topic3 \Rightarrow topic2	(3.9, 72.7, 4.055, 1.377)
6	topic9 \Rightarrow topic10 topic-pdf	(0.057, 1.0, 7.537, 1.917)
7	topic10 topic7 \Rightarrow topic8 topic-pdf	(0.037, 1.0, 14.536, 1.875)
23	topic-pdf topic10 topic6 \Rightarrow topic8	(2.9, 66.7, 9.690, 1.286)
40	exam2 topic-pdf \Rightarrow topic10	(1.7, 77.8, 5.862, 1.167)
9	test2 \Rightarrow test1 test3	(4.9, 71.4, 13.844, 1.667)
10	test9 \Rightarrow test6 test7 test8 topic-pdf topic10	(2.5, 66.7, 27.133, 1.667)
14	test7 topic-pdf topic10 \Rightarrow test6 test8 test9	(2.5, 76.9, 31.308, 1.538)
23	test9 \Rightarrow test8 topic-pdf topic10	(3.4, 93.3, 23.742, 1.273)
28	test3 test4 \Rightarrow test5 topic-pdf	(2.7, 73.3, 14.213, 1.222)

The rules show that the course is divided clearly in two parts, up to topic and test number 5 and the followings (see rules 2 and 18 and 6, 7 and 23 as well as the set of rules from 9 to 28). The instructor observed that not all topics get really studied: some are worked out only through self-tests (set rule from 9 to 28 with a higher support than the corresponding to topic rules). He was very interested by these rules: first, as many of them indicate that students do not really study their assigned materials, but rather they undertake the tests and only look at the study materials when they do not know the answer, hence reversing the intended order of use of the materials; second, because they show that the outdated, incomplete materials from the earlier edition of the course (topic-pdf appears in most rules), which were thought of as a remedial offer for cases of technical connectivity difficulties only, were actually used much more than intended, even in sessions devoted to learning through self-tests. The first seven rules shown in the table also seems to suggest that students checked at what extent the contents of each topic differs from the old compiled version and as it was easier to manage and carry out searches, they frequently used it with tests. Another piece of interesting information, as judged by the teacher, is the fact that the topics in the second half of the course were consulted in more sessions than the first; this did match his perception that he had had to offer

more “moral support” to students on the brink of failure towards the end of the course. Rule 38 shows a good support for exam2, which is not the case for exam1; in fact, the exams are one-shot events. This unexpected support for exam2 was due to technical problems: half the students lost their connections and had to reconnect later in order to finish their exams, accounting for a misleadingly high number of sessions. (The instructor was surprised that our association rules could detect this.).

With Coron’s ChARM many of the rules generated are somewhat redundant variants of the rules found by *yacaree*. Many other rules are also found: essentially, longish rules of confidence 100% (see Table 13). The task of browsing through the hundreds of rules, however, is slow and not user-friendly, and we do not believe a regular instructor would display enough patience to find out the most instructive rules among those returned by the algorithm.

Table 13. Subset of association rules obtained with Coron’s ChARM implementation on the “materials” dataset

No.	Association rule	(Supp. , Conf.)
6	topic7 topic9 topic10 topic-pdf \Rightarrow topic8	(1.23, 100.00)
7	topic7 topic8 topic9 topic-pdf \Rightarrow topic10	(1.23, 100.00)
8	topic7 topic8 topic9 topic10 \Rightarrow topic-pdf	(1.23, 100.00)
9	topic7 topic9 topic-pdf \Rightarrow topic8 topic10	(1.23, 100.00)
65	test5 test7 test8 test9 topic10 topic-pdf \Rightarrow test6	(1.47, 100.00)
66	test5 test6 test8 test9 topic10 topic-pdf \Rightarrow test7	(1.47, 100.00)
67	test5 test6 test7 test9 topic10 topic-pdf \Rightarrow test8	(1.47, 100.00)

This objection also happens in Borgelt’s implementation and worsens with the Weka Apriori, which produces 2272 rules, of which 1522 are again longish implications of confidence 100%. Still, one can see that some of the rules having several items as consequent subsume into a single line several rules that the classical scheme separates into one rule per consequent item. Predictive Apriori generates 1730 rules, of which the first handful are 100% confidence implications with topic-pdf (the old material) as consequent, and the rest consists mostly of rules of rather low support.

3 Conclusions

One of the drawbacks of some data mining algorithms is a dependence on suitable parameter settings which can be difficult for “non-expert data miners” to determine. Another aspect is the degree of difficulty of interpretation of the results. Although the results obtained by association rule miners can be considered easy to interpret by end-users, the large number of rules generated by the more commonly used algorithms, most of which contain facts that, intuitively, will be seen as redundant by users, makes their interpretation and comprehension difficult.

Our comparison of different associators shows that they are vastly different in mere quantitative terms (already advanced in [8] and confirmed in this work); most associators lead to voluminous output; on the other hand, *yacaree* provides several dozen rules that may contain good knowledge yet will not overwhelm the user.

The main question, then, is: are they “the right ones?” Our educational datasets seem to require a low support threshold, but do include items of rather high support; and this combination seriously hinders the ability of traditional association miners to offer interesting output. On the other hand, the most recent version of *yacaree*, which includes implications of confidence 100%, seems particularly well-suited to these cases, and finds rules of both high and low supports; and indeed we find that in most cases these rules “say different things”. All our conclusions have been thoroughly discussed with the instructors of the virtual courses to which the datasets refer.

Summarizing, we can say that *yacaree* offers several advantages for non-expert data miners. First, it offers a parameter-less interface, which makes its usage easier. Second, it generates a reduced number of rules, as it works with closed frequent itemsets, mines only a rule basis, and prunes the rules through the confidence boost parameter. Third, it shows the support, confidence, lift and confidence boost in the output at the same time, which allows end-users to better assess the rules, once these measures are conveniently explained.

The current (and previous) versions of *yacaree* present a limitation: by default, it sets up the number of output rules to 50; our study reveals that this condition should be removed or, at least, relaxed. Previous versions did not search for full implications, and only the latest current version (1.2.0) does; our studies confirm that this must be maintained, as a number of interesting implications for our external user were missed in previous versions.

As final conclusion, our interaction with the instructors involved in the virtual courses analyzed indicates that the results of *yacaree* are superior, in the case of analyzing datasets coming from logs of educational learning systems, in comparison with the rest of the algorithms used in our case study. This program can be freely downloaded from SourceForge, and a link has been provided in the web page on FCA software kindly maintained by prof. Uta Priss.

References

1. Luxenburger, M.: Implications partielles dans un contexte. *Mathématiques et Sciences Humaines* **29** (1991) 35–55
2. Ganter, B., Wille, R.: *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag (1999)
3. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules. In: *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press (1996) 307–328
4. Balcázar, J.L.: Parameter-free association rule mining with *yacaree*. In Khenchaf, A., Poncelet, P., eds.: *EGC. Volume RNTI-E-20 of Revue des Nouvelles Technologies de l’Information.*, Hermann-Éditions (2011) 251–254

5. Balcázar, J.L., García-Sáiz, D., de la Dehesa, J.: Iterator-based algorithms in self-tuning discovery of partial implications. ICFCA, Supplementary proceedings (2012)
6. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with Titanic. *Data Knowl. Eng.* **42**(2) (2002) 189–222
7. Balcázar, J.L.: Formal and computational properties of the confidence boost in association rules. Available at: [<http://personales.unican.es/balcazarjl>]. Extended abstract appeared as [31] (2010)
8. Zorrilla, M.E., García-Sáiz, D., Balcázar, J.L.: Towards parameter-free data mining: Mining educational data with yacaree. [32] 363–364
9. Hung, J.L., Zhang, K.: Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *Journal of Online Learning and Teaching* **4**(4) (2008) 426–436
10. Zaïane, O.R.: Building a recommender agent for e-learning systems. In: Proc. of the International Conference on Computers in Education (ICCE), Washington, DC, USA, IEEE Computer Society (2002) 55–59
11. Au, T.W., Sadiq, S., Li, X.: Learning from experience: Can e-learning technology be used as a vehicle? In: Proceedings of the fourth International Conference on e-Learning, Toronto: Academic Publishing Limited (2009) 32–39
12. Ueno, M., Okamoto, T.: Bayesian agent in e-learning. *IEEE International Conference on Advanced Learning Technologies* (2007) 282–284
13. Perera, D., Kay, J., Koprinska, I., Yacef, K., Zaïane, O.R.: Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering* **21**(6) (2009) 759–772
14. Romero, C., Ventura, S.: Educational data mining: A review of the state-of-the-art. *IEEE Transactions on Systems, Man and Cybernetics, part C: Applications and Reviews* **40**(6) (2010) 601–618
15. Castro, F., Vellido, A., Nebot, A., Mugica, F.: Applying data mining techniques to e-learning problems. In Kacprzyk, J., Jain, L., Tedman, R., Tedman, D., eds.: *Evolution of Teaching and Learning Paradigms in Intelligent Environment*. Volume 62 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg (2007) 183–221 10.1007/978-3-540-71974-8_8.
16. Romashkin, N., Ignatov, D.I., Kolotova, E.: How university entrants are choosing their department? mining of university admission process with fca taxonomies. [32] 229–234
17. Ignatov, D.I., Mamedova, S., Romashkin, N., Shamshurin, I.: What can closed sets of students and their marks say? [32] 223–228
18. Belohlávek, R., Sklenar, V., Zaczal, J., Sigmund, E.: Evaluation of questionnaires supported by formal concept analysis. In Eklund, P.W., Diatta, J., Liquiere, M., eds.: *CLA*. Volume 331 of *CEUR Workshop Proceedings*., CEUR-WS.org (2007)
19. Merceron, A., Yacef, K.: Mining student data captured from a web-based tutoring tool: Initial exploration and results. *Journal of Interactive Learning Research* **15**(4) (2004) 319–346
20. Zorrilla, M.E., García-Saiz, D.: Mining service to assist instructors involved in virtual education. In Zorrilla, M.E., Mazón, J.N., Óscar Ferrández, Garrigós, I., Daniel, F., Trujillo, J., eds.: *Business Intelligence Applications and the Web: Models, Systems and Technologies*. Information Science Reference (IGI Global Publishers) (September 2011)
21. García, E., Romero, C., Ventura, S., de Castro, C.: An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Model. User-Adapt. Interact.* **19**(1-2) (2009) 99–132

22. García, E., Romero, C., Ventura, S., Calders, T.: Drawbacks and solutions of applying association rule mining in learning management systems. In: *Procs of the International Workshop on Applying Data Mining in e-Learning*. (2007) 13–22
23. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques* (2ed). Morgan Kaufmann (2005)
24. Merceron, A., Yacef, K.: Interestingness measures for associations rules in educational data. In de Baker, R.S.J., Barnes, T., Beck, J.E., eds.: *EDM*, www.educationaldatamining.org (2008) 57–66
25. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. *ACM Comput. Surv.* **38**(3) (2006)
26. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research* **184**(2) (2008) 610–626
27. Borgelt, C.: Efficient implementations of apriori and eclat. In Goethals, B., Zaki, M.J., eds.: *FIMI*. Volume 90 of *CEUR Workshop Proceedings.*, CEUR-WS.org (2003)
28. Scheffer, T.: Finding association rules that trade support optimally against confidence. In: *5th European Conference on Principles of Data Mining and Knowledge Discovery*. (2001) 424–435
29. Zaki, M.J., Hsiao, C.J.: Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering* **17**(4) (2005) 462–478
30. Kaytoue, M., Marcuola, F., Napoli, A., Szathmary, L., Villerd, J.: The Coron System. In Boumedjout, L., Valtchev, P., Kwuida, L., Sertkaya, B., eds.: *8th International Conference on Formal Concept Analysis (ICFCA) - Supplementary Proceedings*. (2010) 55–58 (demo paper).
31. Balcázar, J.L.: Objective novelty of association rules: Measuring the confidence boost. In Yahia, S.B., Petit, J.M., eds.: *EGC*. Volume RNTI-E-19 of *Revue des Nouvelles Technologies de l'Information.*, Cépaduès-Éditions (2010) 297–302
32. Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C., Stamper, J.C., eds.: *Procs of the 4th International Conference on Educational Data Mining*, Eindhoven, The Netherlands, July 6-8, 2011. In Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C., Stamper, J.C., eds.: *EDM*, www.educationaldatamining.org (2011)