Towards Parameter-Free Data Mining: Mining Educational Data with *yacaree*

MARTA E. ZORRILLA and DIEGO GARCÍA-SAIZ and JOSE L. BALCÁZAR, University of Cantabria

The need of parameter-free alternatives for data mining algorithms is widely recognized, and is particularly acute in the web-based educational field, where instructors involved in the teaching process are interested in improving their virtual courses and adapting these to the learners' behavior: most often, these instructors are not expected to know about data mining technologies. We report on a quantitative comparison of several algorithms for association rules on educational datasets, including in the comparison both well-established implementations and a recent parameter-free association miner; our purpose is to clarify whether this newer approach is actually useful for the educational data mining field. Our results indicate that it has indeed a high potential, and allows us to identify some important aspects that must be improved.

1. INTRODUCTION

In the educational arena, Data Mining techniques are acquiring a major importance since the appearance of the e-learning environments. These systems log all the activity carried out by students and instructors, and this raw data, adequately processed, may offer useful knowledge about the learning process for instructors.

But data mining techniques are out of the reach of most teachers, e.g., for humanities or law studies. Thus, if we want to help users of all disciplines, we need to work out data mining tools that do not require much tuning or technical understanding from the user. In particular, this is relevant for the case of association rules: all the available algorithms up to recent work depend on one or more parameters (confidence, support,etc) whose value is to be set by the user, and whose semantics may not be easy to grasp. Likewise, the number of rules which obtain as output is often large, and most of them are redundant and non-interesting for decision making [García et al. 2007].

There is, thus, a clear need to design and implement parameter-free data mining algorithms addressed to "non-experts", and they must stand reasonably well a comparison with other "expert"-oriented algorithms. To the best of our knowledge, *yacaree* [Balcázar 2011] is the first parameter-free association miner implemented. Here, we compare this system with other three well-known association rule miners: the Apriori [Agrawal and Srikant 1994] and Predictive Apriori [Scheffer 2001] tools from Weka, and Borgelt's Apriori implementation [Borgelt 2003].

2. MAIN RESULTS

Yacaree is a parameter-free algorithm which mines frequent closed itemsets and constructs association rules from them; a main property is that it reduces the number of rules shown to the user by means of a parameter, called confidence boost, which eliminates redundant rules. Essentially, a rule is considered redundant with respect to another if it has larger antecedent or smaller consequent, and simultaneously the ratio of their confidences falls below a given threshold. This algorithm is available at http://sourceforge.net/projects/yacaree/.

Thus, this algorithm tends to generate rules with small antecedents and larger consequents, unlike the original association rules and their implementation by Borgelt. In fact, whereas a rule like $X \to AB$ implies

This work is partially financed by the Spanish Government grants TIN2007–66523 and TIN2008–05924.

Author's addresses: M. E. Zorrilla, D. García-Saiz and J. L. Balcázar, Mathematics, Statistics and Computation Dept., Univ. of Cantabria.

Table I. Number of rules obtained on our datasets with the four algorithms

	Number of rules at support 1% and confidence 66%			
	Weka Apriori	Predictive Apriori	Borgelt Apriori	*yacaree*
Dataset1	2272	1730	617	24
Dataset2	7523	over 10000	3751	214
Dataset3	4249	over 10000	1876	88

both rules $X \rightarrow A$ and $X \rightarrow B$, the converse is not true, and, therefore, *yacaree* choses to keep larger consequents whenever possible as they furnish the most informative configuration. This program keeps a fixed confidence threshold, defaulting at 2/3, and it reports only rules that belong to a certain "basis" of irredundant rules corresponding to the threshold. It has very generous start values for support and confidence boost thresholds; but a key property is that the algorithm "tunes" on itself these two thresholds along the computation, by monitoring lift, memory consumption, and other parameters.

Our three educational datasets come from logs of a virtual course, and relate sessions and materials employed by the students; they have, respectively, 407 transactions on 22 items (Dataset1), 2486 transactions on 27 items (Dataset2), and 2346 transactions on 26 items. As *yacaree* self-tunes the support, we performed a brief preprocessing to tune manually the rest of algorithms and guarantee a fair comparison. We decided to fix at 1% the support threshold for all the computations, and at 2/3 (or 66%) the confidence threshold. The limit on the number of rules in the Weka tools was set very high (at 10000 rules), and left unbounded in Borgelt's Apriori and *yacaree*. We show the number of rules obtained for each case in Table I.

The first consideration that we can highlight is that *yacaree* provides a reasonable size of the output. In general these rules contain good knowledge without overwhelming the user. Furthermore, these rules are, intuitively, reasonably irredundant ("they say different things"). Instead, both Apriori implementations in Weka and the one by Borgelt lead to more voluminous and redundant output. Predictive Apriori tends to choose first rules of a support rather lower than the user would like to, tends to create overwhelming output sizes, and leaves room for quite a degree of redundancy. Its running time tends to be unacceptably high, and the "expected predictive accuracy" parameter is less interpretative than support and confidence for the end-user. On the other hand, all of these well-established algorithms do return rules of 100% confidence, something that *yacaree* does not. We hope to add this feature in the near future, as this experiment clearly marked this as the issue that needs remedy most urgently.

In the opinion of the instructor involved in the virtual course analyzed (prof. Rafael Menéndez), the results of *yacaree* are superior in comparison with the rest of the algorithms used in our case study, in terms of subjective usefulness for the teacher. In summary, *yacaree* seems particularly well-suited to educational datasets which seem to require a low support threshold, but do include items of rather high support, as this combination seriously hinders the ability of traditional association miners to offer interesting output.

REFERENCES

Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules in large databases. In *VLDB*, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 487–499.

Balcázar, J. L. 2011. Parameter-free association rule mining with yacaree. In *EGC*, A. Khenchaf and P. Poncelet, Eds. Revue des Nouvelles Technologies de l'Information Series, vol. RNTI-E-20. Hermann-Éditions, 251–254.

Borgelt, C. 2003. Efficient implementations of apriori and eclat. In *FIMI*, B. Goethals and M. J. Zaki, Eds. CEUR Workshop Proceedings Series, vol. 90. CEUR-WS.org.

García, E., Romero, C., Ventura, S., and Calders, T. 2007. Drawbacks and solutions of applying association rule mining in learning management systems. In *Proceedings of the International Workshop on Applying Data Mining in e-Learning.* 13–22.

Scheffer, T. 2001. Finding association rules that trade support optimally against confidence. In *In: 5th European Conference on Principles of Data Mining and Knowledge Discovery.* 424–435.