When Data Exploration and Data Mining meet while Analysing Usage Data of a Course[1]

André Krüger[1,2], Agathe Merceron[2] and Benjamin Wolf[2]
{akrueger, merceron, bwolf}@beuth-hochschule.de
[1]Aroline AG, Berlin, Germany
[2]Beuth University of Applied Sciences, Berlin, Germany

1   Introduction

Learning Management Systems (LMS) are web-based systems that are increasingly used in education. If lecturers want to gain a deeper insight on whether and how students use the learning resources they put at their disposal in a course, user data stored by the LMS needs to be explored. To help explore these data, we have designed and implemented an application called *ExtractAndMap* [1] that structures and exports all data stored at various places and under various forms by a LMS into a consistent data base that can be used for numerous queries and further data mining. Using queries on the data base it is possible to answer questions like: "how many students have attempted self-evaluation exercise of week 1?", "how many students have attempted self-evaluation exercise of week 2?" and so on, till the end of the semester. When looking at the results it may well be that these numbers diminish as shown in Figure 3 p. 7 in [1].

An experienced lecturer will read in those figures the following familiar experience: at the beginning of the semester students are enthusiastic and are not yet overloaded with a lot of homework in different courses. Therefore, many of them attempt self-evaluation exercises. As the semester progresses, always less students attempt the exercises till a stable group remains that sticks to its working habits and consistently attempt all evaluation-exercises. Is this hypothesis correct? The sole answers to the above questions tackling each exercise separately cannot tell for sure. In this contribution we show that a deeper exploration with queries handling several exercises together like "how many students have attempted exercise A and exercise B?" can be enough to infer the association rule "if students attempt exercise A, then they also attempt exercise B".

2   Inferring Association Rules from Data Exploration

For definitions about association rules and interestingness measures like *confidence*, *lift* and *cosine* the reader is referred to [2]. Data exploration involves quite often simple counting, like how many transactions contain X, Y or X *and* Y, thus, giving $P(X)$, $P(Y)$ or $P(X,Y)$. If exploration shows that $P(X)$ and $P(X,Y)$ are equal, the we have:

$$conf(X \rightarrow Y)=1, \; lift(X \rightarrow Y)=\frac{P(X,Y)}{P(X).P(Y)}=\frac{1}{P(Y)} \; \text{and}$$

$$cosine(X \rightarrow Y)=\frac{\sqrt{(P(X))}}{\sqrt{(P(Y))}}, \text{ where } X \rightarrow Y \text{ is the association rule "if } X, \text{ then } Y\text{". In}$$

practice $P(X)$ and $P(X,Y)$ are going to be almost equal, not exactly equal. Table 1

---

shows how confidence, lift and cosine evolve. The rows show $P(X,Y)$ as a fraction of $P(X)$. $P(X,Y)=0.97\,P(X)$ means that $P(X,Y)$ equals 97% of $P(X)$. $P(Y)$ is taken to be 0.8 in the third column and 0.7 in the fourth column. $P(X)$ is taken to be first 2/3 of $P(Y)$ in column 5 and 3/4 of $P(Y)$ in the last column. We recall that confidence is a number between 0 and 1 (highest is 1), that lift rates a rule as interesting if its value is above 1, and that cosine is a number between 0 and 1 and rates a rule as interesting if its value is above 0.66.

Table 1. Evolution of confidence, lift and cosine

	conf.	*lift* $P(Y)=0.8$	*lift* $P(Y)=0.7$	*cosine* $2/3\,P(Y)$	*cosine* $3/4\ P(Y)$
$P(X,Y)=P(X)$	1	1.25	1.43	0.81	0.87
$P(X,Y)=0.97\,P(X)$	0.97	1.21	1.39	0.79	0.84
$P(X,Y)=0.95\,P(X)$	0.95	1.19	1.36	0.77	0.82
$P(X,Y)=0.90\,P(X)$	0.90	1.12	1.29	0.65	0.78

Summing up, when $P(X)$ is almost equal to $P(X,Y)$, data exploration is enough to infer the association rule $X \to Y$, there is no need to use a data mining algorithm for association rules extraction in such a case.

3   Conclusion and Future Work

We have used this result while analyzing the data of the course "Programming 1" in our university, see [1]. The associations found show that a group of students emerges that keep doing self-evaluation exercises during the semester. A future work is to continue conducting case studies with courses taught in different topics and designed in different ways to further enhance our catalogue of questions that can be interesting for teachers, and investigating further connections between data exploration and data mining.

References

[1] Krueger A., Merceron, A., Wolf, B. A Data Model to Ease Analysis and Mining of Educational Data. *Third International Conference on Educational Data Mining,* 2010. Pittsburgh, USA.

[2] Merceron, A., Yacef, K. Interestingness Measures for Association Rules in Educational Data. *First International Conference on Educational Data Mining,* 2008. Montreal, Canada, p. 57-66.