

Bayesian Networks for Web Usage Mining

Reti Bayesiane per il Web Mining

Paola Cerchiello

Dipartimento di Statistica Università degli Studi Milano-Bicocca

paolacerchiello@libero.it

Paolo Giudici

Dipartimento di Ec. Pol. e Met. Quant.- Università degli Studi di Pavia

giudici@unipv.it

Riassunto: Questo articolo è lo sviluppo di un contributo esposto nella precedente conferenza della Società Italiana di Statistica (Giudici, 2002). L'obiettivo del lavoro è quello di offrire nuovi spunti nel campo del web usage mining attraverso il confronto dei risultati precedentemente ottenuti (tramite l'utilizzo di alberi decisionali e sequenze) con nuovi risultati, ottenuti dall'applicazione di modelli grafici discreti. Ricordiamo che l'obiettivo del web usage mining consiste nel mostrare come l'informazione, riguardante l'ordine in cui le pagine di un sito web sono visitate, possa essere impiegata utilmente per prevedere il comportamento di visita relativo a un sito.

Keywords: Graphical models, Sequence rules, Classification trees.

1. Sequence rules

We first briefly recall what a sequence rule is. For more details the reader can consult a recent text on data mining, such as Han and Kamber (2001) or, from a more statistical viewpoint, Hand et al (2001), Hastie et al. (2001) and Giudici (2003).

An *association rule* is a statement between two sets of binary variables (itemsets) A and B , that can be written in the form $A \rightarrow B$, to be interpreted as a logical statement: *if A , then B* . If the rule is ordered in time we have a sequence rule and, in this case, A preceeds B .

In web clickstream analysis, a sequence rule is typically *indirect*: namely, between the visit of page A and the visit of page B other pages can be seen. On the other hand, in a *direct* sequence rule A and B are seen consecutively.

A sequence rule model is, essentially, an algorithm that searches for the most interesting rules in a database. In order to find a set of rules, statistical measures of "interestingness" have to be specified. The measures more commonly used in web mining to evaluate the importance of a sequence rule are the indexes of support and confidence.

In this paper we shall consider mainly the confidence index. The confidence for the rule $A \rightarrow B$ is obtained dividing the number of server sessions which satisfy the rule by the number of sessions containing the page A .

Therefore, the confidence approximates the conditional probability that, in a server session in which page A has been seen, page B is subsequently requested.

We finally remark that classification trees represent a global model to obtain sequence rules (see e.g. Giudici, 2003).

2. Graphical models

A graphical model (see for example Lauritzen, 1996) is a family of probability distributions incorporating the conditional independence assumptions represented by a graph. It is specified via a graph that depicts the local relations among the variables (that are represented with nodes). In this paper we will focus on the case in which all considered variables are discrete, and the graph is undirected (symmetric).

It is known that recursive graphical models consist of a powerful tool for predictive data mining, because of their fundamental assumption of probability dependency between variables. However, symmetric graphical models can be considered as an important and valid tool in the preliminary phase of analysis because they can show the main relevant association, useful to construct a subsequent recursive model.

3. Web usage mining

The data set that we consider to illustrate our methodology is the result of the elaboration of a log file concerning a site of e-commerce, described, for instance, in Giudici (2003). The whole data set contains 250711 observations, each corresponding to a click, that describe the navigation paths of 22527 visitors among the 36 pages which compose the site of the webshop. For illustrative purposes, Table 1 reports a very small extract of the available dataset, corresponding to the user session of one visitor

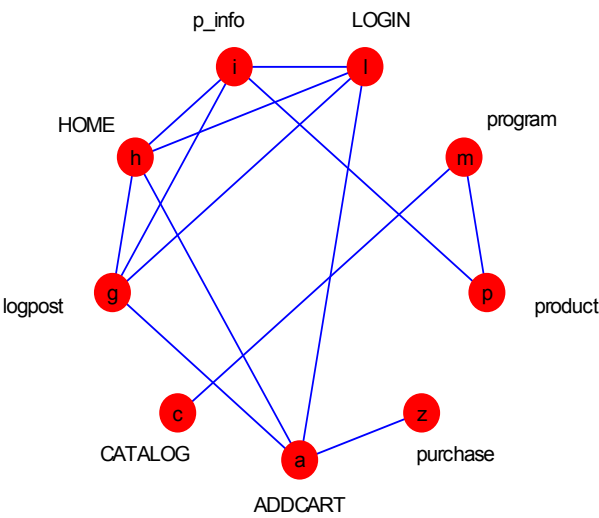
Table 1: Extract of the considered dataset

<i>c_value</i>	<i>c_time</i>	<i>c_caller</i>
70ee683a6df...	14OCT97:11:09:01	home
70ee683a6df...	14OCT97:11:09:08	catalog
70ee683a6df...	14OCT97:11:09:14	program
70ee683a6df...	14OCT97:11:09:23	product
70ee683a6df...	14OCT97:11:09:24	program

In order to model the above data we have considered three main classes of statistical models: sequence rules, decision trees and symmetric graphical models. We underline that for the construction of the symmetric graphical models we have used the software MIM (www.hypergraph.com), instead for the association rules and the decision trees we have considered SAS Enterprise Miner.

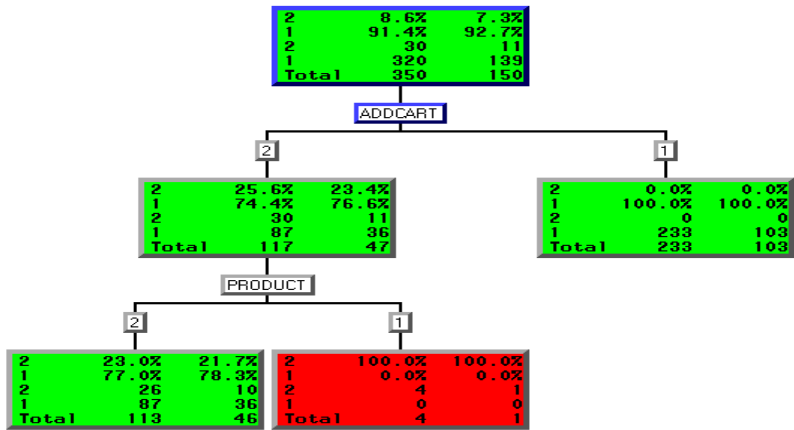
We considered two contexts of analysis: in the first one, in order to obtain a useful comparison between graphical model and decision trees, we introduced a target variable 'purchase' that describes whether a client buys or not the products sold. A forward selection procedure, using a significance level of 5 %, gave us the structure in Figure 1:

Figure 1: Graphical model with the variable 'purchase'



From Figure 1 the variable 'purchase' has only one link (one direct association), with the variable 'addcart'. In Figure 2 we compare this result with a classification tree model.

Figure 2: Decision tree with the variable 'purchase'

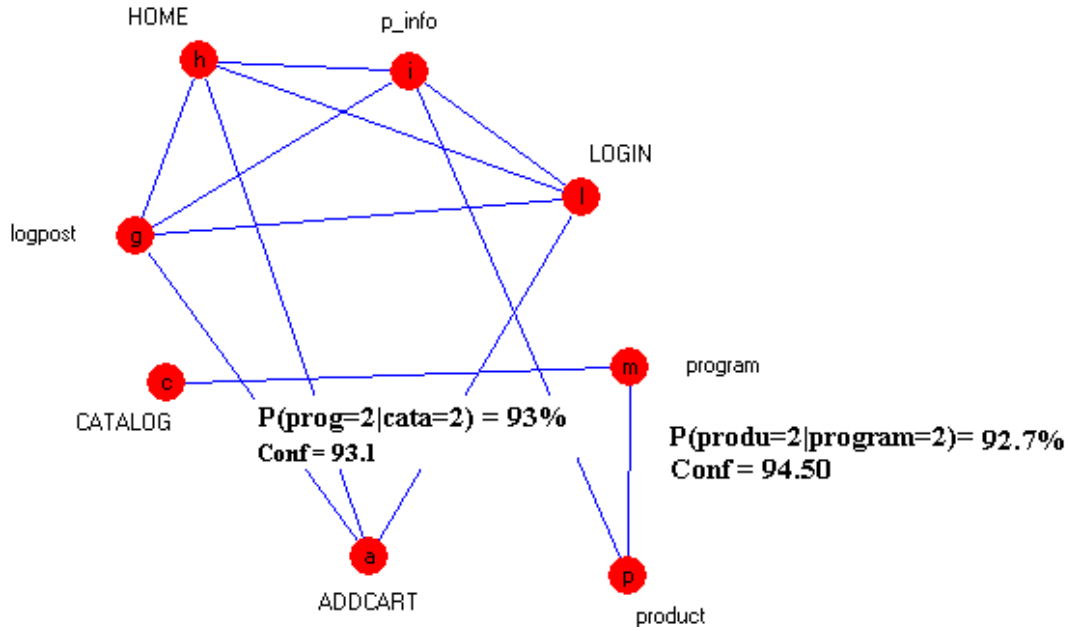


The decision tree in Figure 2 gives us a similar result, in fact 'addcart' represent the variable most important in the first split of the tree. Therefore we can conclude that the results from graphical models and decision trees give consistent results.

In the second context of analysis we excluded the target variable 'purchase' in order to obtain a comparison between graphical models and sequence rule models. In order to summarise the main results of this analysis, in Figure 3 we present the selected undirected graphical model (according to a forward procedure and 5% significance level). Furthermore, we attach to each edge the highest of the two confidence indexes

and conditional probabilities associated to it. For lack of space, in Figure 3 we represent such statistics only for the two strongest associations.

Figure 3: Graphical model with confidence indexes and conditional probabilities.



The graphical model in Figure 3 shows, as the strongest associations, the relationship between the variables catalog-program and program-product. We remark that, while confidence indexes are based on the results from the Apriori algorithm of SAS Enterprise Miner (see for instance Giudici, 2003) conditional probabilities have been derived from the fitted contingency tables produced by the software MIM.

Comparing confidence indexes with conditional probabilities it turns out that the results from graphical models are quite similar with those from sequence rules. This means that, for this kind of data, sequence rules, based on marginal associations, constitute a valid approximation to detect association structures well described by a graphical model. On the other hand, they are much easier to calculate and interpret.

Essential references

Giudici, P. (2002) Association models for web mining. *Atti della XLI riunione scientifica della Società Italiana di Statistica*, Cleup, Milano, pp. 329-338.

Giudici, P. (2003) *Applied Data mining, statistical methods for business and industr.* Wiley, London.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The elements of statistical learning: data mining, inference and prediction*. Springer-Verlag, Berlin.

Han, J. and Kamber, M. (2001) *Data mining: concepts and techniques*. Morgan Kaufman, New York.

Hand, D.J., Mannilla, H. and Smyth, P (2001) *Principles of Data Mining*, MIT Press, New York.

Lauritzen, S.L. (1996). *Graphical models*, Oxford, Oxford University Press.