Deliverable 2. Data Analysis.

Detailed insight on available data and first statistic analysis.

Grupo de Sistemas Inteligentes

Departamento de Ingeniería de Sistemas Telemáticos

Universidad Politécnica de Madrid.

Project Report

Madrid, October 2012

Authors:

Adrián Pérez Orozco

Álvaro Carrera Barroso

Carlos A. Iglesias Fernández

Executive Summary

Contents

List of Figures

List of Tables

1 Evaluation criteria

The predictive information obtained in the data mining process, will lead to the implementation of systems which will give us a prediction using current events as its input. This prediction will be given in the form of an alarm or set of alarms, which are likely to be raised within a given prediction period. In this section we will approach the problem of *evaluation* of these predictive information.

As a first thought, it might seem appropriate to evaluate our predictions by how true they actually are. We can measure the *accuracy* of a prediction rule system easily by checking how often it becomes true and how often it does not. This is an important factor to take into account, but is however not completely significant of the overall quality of the system. In a limit case in which we only attained a trivial but highly accurate rule which gives valid but trivial predictions all the times, we would have an accuracy of 100%, while the overall quality of the system would be none. We must actually check not only the accuracy of our predictions, but also their relevance against the whole situation.

Therefore, we will have two different evaluation parameters: one related to the accuracy of our predictions, and other related to the fraction of events we are able to predict. In first place, we will define *precision* as the fraction of our predictions which are accurate. In the case of evaluating a rule against a test set, $P_{accurate}$ would be the times that both the antecedent and consequent of the given rule have happened within the stipulated time window; while P_{total} would be the times that the antecedent of the given rule has happened, whether the consequent has or hasn't happened. Prediction can be as well calculated for a whole rule set, or for any kind of system which gives a predicted event based on other input events.

$$Prec_i = \frac{P_{i,accurate}}{P_{i,total}} \tag{1}$$

On the other hand, we will define *recall* as the relation between events which have successfully been predicted by our system ($E_{predicted}$) and the

total number of events (E_{total}).

$$Rec_i = \frac{E_{i,predicted}}{E_{i,total}} \tag{2}$$

Notice that the number of events which have been predicted ($E_{predicted}$) is, in fact, the number of accurate predictions as calculated in the definition of *precision*, ($P_{accurate}$)

In other words, precision is the ratio between accurate predictions and the total number of predictions; while recall is the ratio between accurate predictions and the total number of events.

It is important to notice that in our context, an event can't be *wrongly* predicted. Our prediction can be either true or false, but if we make a prediction of the type $\{A, B\} \longrightarrow \{C\}$ and instead we observe that $\{A, B\} \longrightarrow \{D\}$; it does not mean in any way that we predicted C instead of D, but that our prediction of C was false and we did not predict D. As a result, some other tools generally used to complement values of precision and recall (such as *confusion matrices*) cannot be applied in our case.

Taking a further step, we can merge both indicators in a single one, obtaining a single indicator for a much easier evaluation. Precision and recall are often merged in the called *F-measure*, defined as:

$$F = \frac{(\beta^2 + 1) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{3}$$

where $\beta \in [0, 1]$ balances the importance between recall and precision.