1st Cladag Young Researcher Data Mining Prize

Sponsored By

SAS
THE POWER TO KNOW.

On the occasion of CLADAG 2009 hosted in Catania, we announce the first edition of "CLADAG Young Researcher Data Mining Prize", sponsored by SAS Institute.

The aim of the Competition is to promote the transfer of young researchers' new scientific knowledge to the operative world of data mining users.

This Competition asks for a solution to a real data mining problem from groups composed of at most 3 young researchers.

CLADAG Young Researcher Data Mining Prize will consist of 2 prizes:

1) 4000 euro for the absolute best solution
2) 1500 euro for the best solution in terms or re-applicability

The grants will be paid as funds to the institutions that winners belong to, or, if the winners prefer, they will be paid directly, in this case all taxes on the prize are the sole responsibility of the winners

Eligibility:
The Competition is open to anyone

1) **under 35 years of age** at the moment of the registration and who has at least a masters degree, for foreign competitors, or at least either a "laurea specialistica" or a "laurea magistrale" or a "laurea di vecchio ordinamento" for Italian competitors;
2) currently employed by a university, in this case only those **at assistant professor** level (*ricercatore*) or below are eligible;
3) currently employed by any Public Administration or Private Business, in this case anyone working as a research director is **not** eligible;
4) who wishes to work in a team, in this case only groups composed of **at most 3 people** with all the above requirements are eligible;
5) who registers for the Competition, paying a fee of 20 euro for each team component.

After registration each team will nominate a correspondent author who will receive the data set on which the analysis process has to be developed and an analytical problem description. It is not possible to change the team's composition once registration for the competition has been completed.

The deadline for registration is May 31, 2009. Competitors must send an email to cladag09@unict.it clearly specifying the composition of the team, together with the correspondent author, and any institutional affiliation.
The registration fee must be paid with a bank transfer:

Bank name: **Credito Siciliano - Sede di Catania**

Account holder: **Comitato Organizzatore Cladag 2009**

SWIFT code: **RSANIT3P**

IBAN code: **IT77T0301916903000006000578**

The bank transfer must be labeled **"Cladag Data Mining Competition 2009"**.

The registration is not valid until the Cladag Prize Organization receives the payment.
Registered teams will submit their solution before July 31, 2009. Components of the registered teams will have right to register for the CLADAG meeting with a special fee ("Young Competitor") of 100 euro before June 14, 2009. After this date the "Young Competitor" fee will be 120 euro.

The problem is described briefly at the end of this document, registration for the Competition is necessary before receiving the analytical description.

Any software can be used. Technical reports submitted for the competition have to spell out **algorithm and analysis process**, with no software reference: this is to encourage mass dissemination of proposed solutions.

The technical report has to be written up in English following CLADAG paper style and it has not to exceed 20 pages.

From all the solutions submitted, the 4 best will be selected according to the criteria indicated below and they will be presented in a specific **plenary session** of the CLADAG meeting. Afterwards, the authors will be invited to submit the scientific papers for publication in the *CLADAG selected papers* volume edited by Springer.

Evaluation criteria of proposed solution:
Innovation level of the proposed techniques, scientific regard to the application, precision in the presentation of results, accuracy and care in the numerical results obtained, solution ability to be re-applied or applied to massive data in an operative context.

Brief description of the problem:
The given data shows an operative form and its structure is fully fashioned to the data mart of real problems.
A set of k variables (qualitative and quantitative) is given, which synthesizes socio-demographic and behavioral information of each subject. Monitoring of these variables set up what we call **X** matrix (**status matrix**) where each row is the observation status.
Behavioral information about each subject is also given. In order to furnish some examples you can think of a service provider studying the behavior of its customers, or a bank which detects how many and which particular one among its financial services its clients have requested, or the case of a mobile telephone company seeking to discover which services have been activated on clients' accounts.

This matrix which we will indicate with **S**, can be called **active behaviour matrix** and it is a qualitative variables matrix. Probably among these matrix variables you can find some kind of association.
You have also a quantitative variable which can be considered as the outcome generated from each observation: this will be the **target** variable *y*.
Dichotomic variable **h** collects information about **actions** (impulses) that each subject has received in the unit time.
Data System (DS) described above is observed in a time series of 36 lags, which means that if DS is collected monthly, data available are referred to 3 years.
The problem idea consists of a variable *y* which depends on the subject status, from the impulse received as well as on the activated behaviour and that this relationship is not time invariant.

Competition aim:
Implementing an analytic system which detects and expresses, even only in descriptive terms, the relationship set present in the Data System and at the same time that gives specification of a predictive system for the variable *y*. Such system will produce the prediction of the variable *y* as will be requested by the competition text problem.