Use of Statistical Implicative Analysis in Complement of Item Analysis

Raphaël Couturier and Rubén Pazmiño

***Abstract*—In many different situations, students or more generally individuals fill forms or surveys. Such forms could be used to evaluate the knowledge of students after a lesson in a classroom or could form a global evaluation of all the students in a country. More generally a survey aims at gathering the opinion of people on a particular subject. In such a case, item analysis gives interesting information on how the items have been answered. In this paper, we present the Statistical Implicative Analysis (SIA) that produces oriented rules. A survey about the future of students in the different schools of the ESPOCH is studied to highlight the interest of using SIA in order to be able to analyze the general behavior of the population.**

***Index Terms*—Statistical implicative analysis, item response theory.**

I. Introduction

Item response theory (IRT) is used in many situations, especially in scoring tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables. IRT examines the questions in order to assess the quality of the items and the test as a whole. This analysis allows the improvement of tests and also enables us to remove ambiguous or erroneous items. IRT brings very interesting information about the variables one by one but there is no relation between the variables. SIA was created to build quasi implication rules and hence to give sense, or understanding between variables.

In this paper we intend to illustrate that Statistical Implicative Analysis gives complementary information to IRT. CHIC is a software that implements most of the SIA tools.

In Section II, SIA is presented. Section III illustrates the kind of computations that SIA allows us to handle. Section IV presents some examples of analysis with a survey from previous students of the ESPOCH. Then we give a conclusion and perspectives.

II. SIA and CHIC

Statistical Implicative Analysis was initiated by Gras [1]. The first goal of this method was to define a way of answering the question: "If an object has a property A, does it also have a property B". The answer to this question is rarely a positive one. Nevertheless it is possible to notice that there are general trends. SIA aims at discovering such tendencies in a set of properties.

As classical association rules methods [2], SIA aims at finding rules between the variables. Nevertheless SIA has a very interesting property, compared to other methods because it provides a non linear measure that satisfies some important criteria. First of all, the method is based on the implication intensity that measures the astonishment degree of a rule. To present the implication index for binary variables, we need to define some notations. In the following we consider that: n represents the total number of subjects, n_a represents the number of subjects having the property a, n_b represents the number of subjects having the property b, and $n_{a\wedge\bar{b}}$ represents the number of subjects having the property a and not b. The implication index is defined by:

$$q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$$

In the general case, when we can approximate the Poisson law with a normal law, the implication intensity is defined by:

$$\varphi(a,b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

Hence, trivial rules that are potentially well known to an expert are discarded. In fact, a rule of the form $A \Rightarrow B$ is considered trivial if almost all objects of the population have property B. For further information the reader is invited to consult [1]. Based on that original measure, CHIC, given a set of data, enables one to extract association rules. CHIC and SIA have been used in wide domain areas, for example [1], [3], [4].

Initially SIA, as CHIC, was designed to handle binary variables. Later, SIA was enhanced by other kinds of variables and so was CHIC. Currently, CHIC allows the user to handle binary variables, frequency variables, variables over intervals and interval-variables. The case of binary variables is obviously the simplest one. Ordinal variables (also called nominal ones) can be coded using as many binary variables as number of categories. Frequency variables take a real value between 0 and 1. This kind of variables allows the user to include the case of discrete variables which only take a fixed number of values (or modalities) ranging between 0 and 1. Of course, the way of defining modalities is very important, because it strongly affects the results of CHIC whether the values of modalities are close to 0 or to 1. This remark is also true concerning the frequency variables. It

Manuscript received June 20, 2014; revised September 15, 2014.

Raphaël Couturier is with the FEMTO-ST Institute, University of Franche-Comte, Belfort, France (e-mail: raphael.couturier@univ-fcomte.fr).

Rubén Pazmiño is with the Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador (e-mail: rpazmino@espoch.edu.ec).

should be noticed that ordinal variables are also coded using frequency ones. The user must pay attention to the way real variables are transformed into frequency ones.

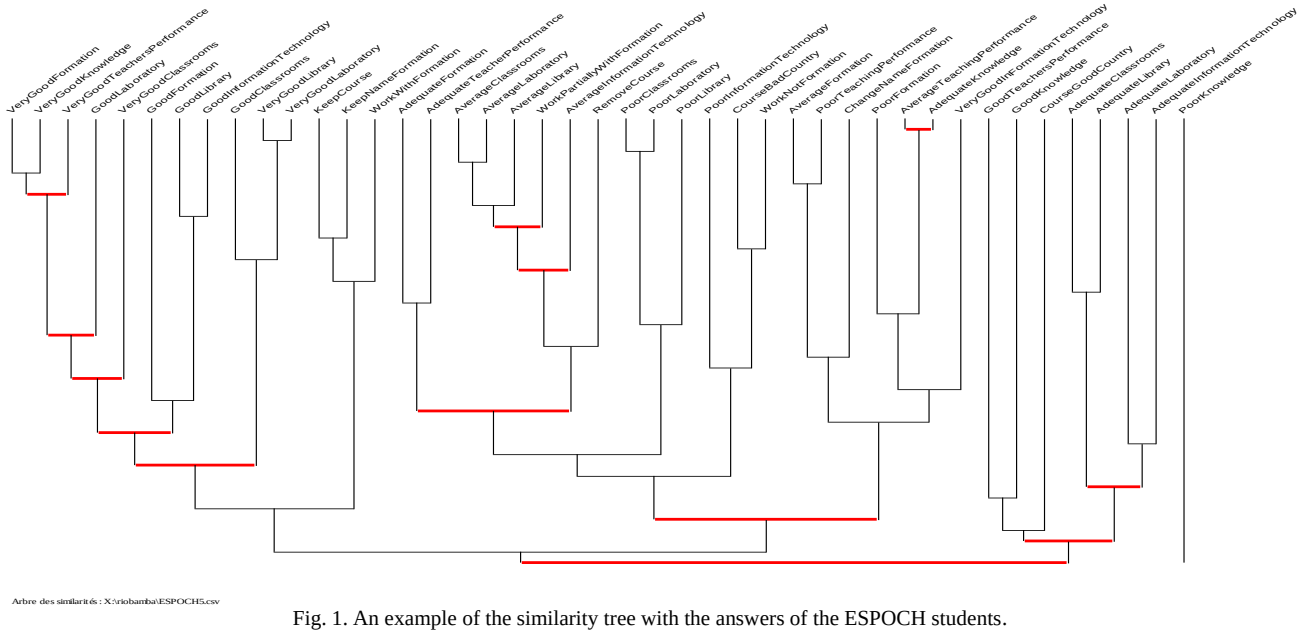

Fig. 1. An example of the similarity tree with the answers of the ESPOCH students.

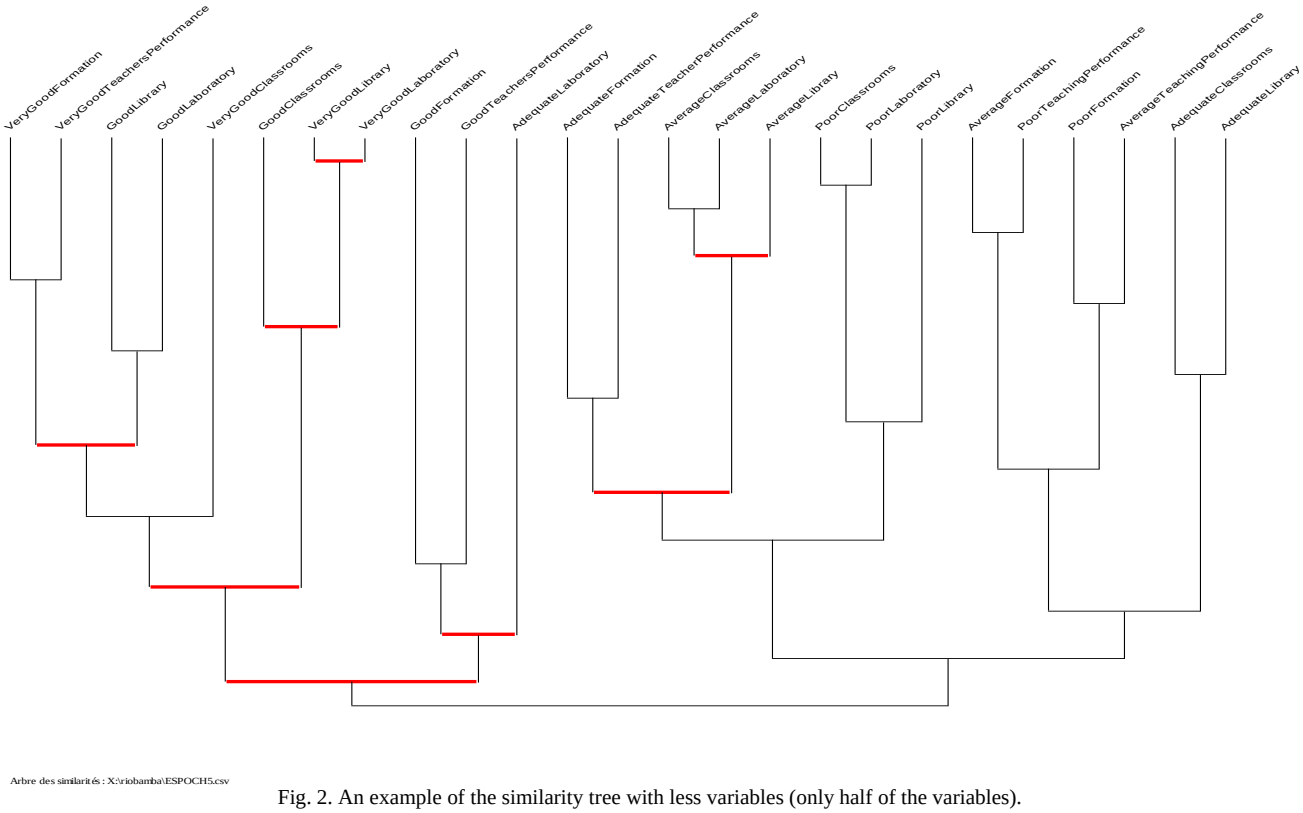

Fig. 2. An example of the similarity tree with less variables (only half of the variables).

Based on the implication intensity and the similarity intensity, proposed by I.C. Lerman in [5], CHIC allows us to build a similarity tree, a hierarchy tree and an implicative graph. The most classical tree is a similarity tree (also known as dendogram). It is based on the similarity index. In a similar way, the implication intensity can be used to build an oriented hierarchy tree. It should be noted that for the hierarchy tree a cohesion index is defined with the implication. More precisely this index is based on the Shannon entropy and the implication and is defined by

$$c(a,b) = \left(1 - \left(-p\,log_2 p - (1-p)log_2(1-p)\right)^2\right)^{1/2}$$

if $p = \varphi(a,b) > 0.5$ *and* $c(a,b) = 0$ otherwise.

The implication intensity can also be used to define an implication graph, which lets the user select the association rules and the variables he or she wants. In opposition to most of the other multidimensional data analysis methods, SIA establishes the following properties between the variables it handles:

- relationships between variables are dissymmetric
- the association measures are non linear and are based on probabilities
- the user can use graphical representations which follow the semantic of the relationship

The different graphs will be shown in the next section.

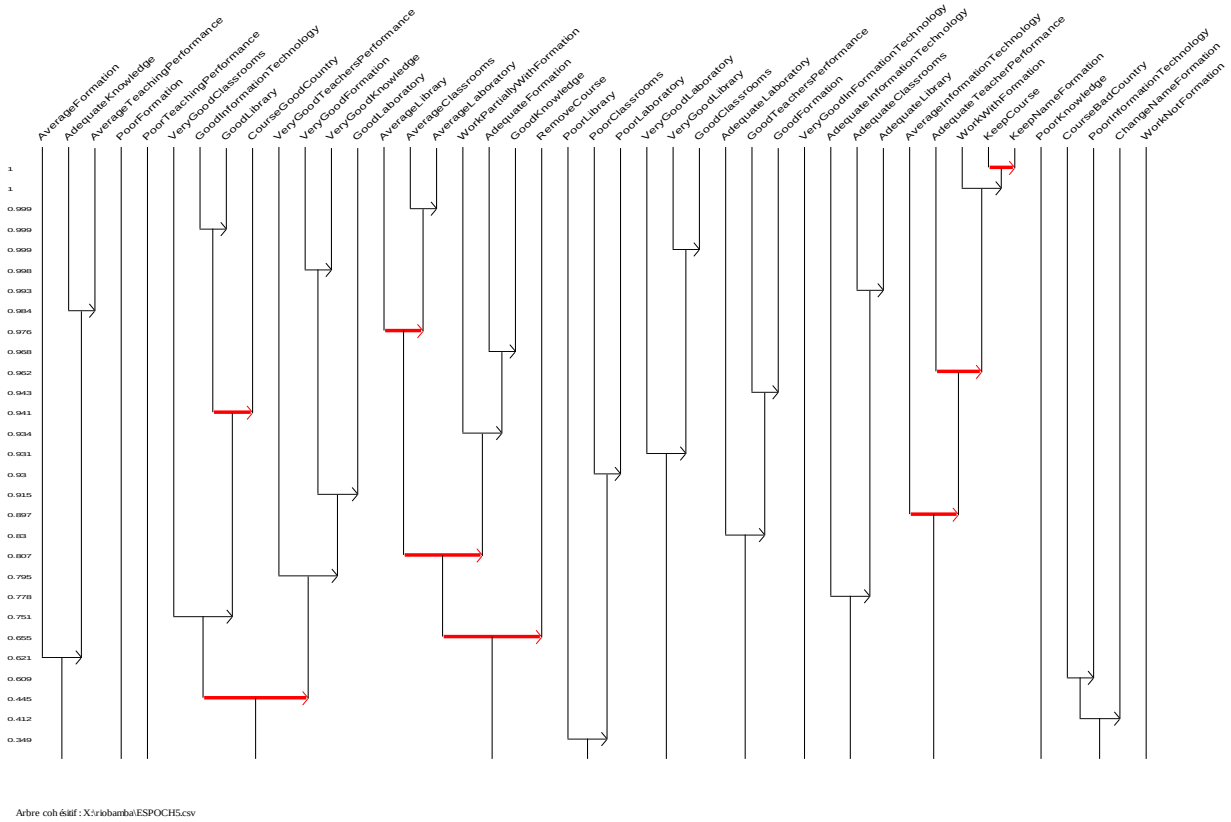

Fig. 3. An example of the hierarchy tree.

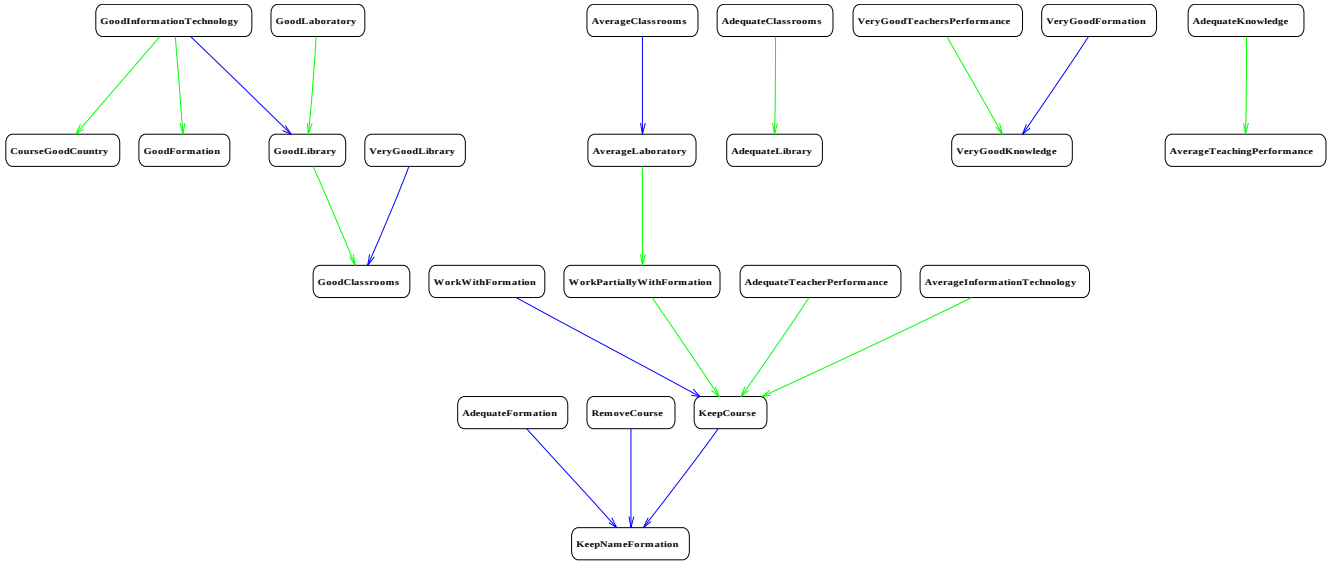

Fig. 4. An example of the implicative graph.

III. Some Possibilities of CHIC for Item Analysis

In order to show the possibility of SIA and CHIC for item analysis we have taken data issued from a survey given to former students of the ESPOCH in Ecuador. These students were asked to answer different questions about their studies in the ESPOCH. As explained previously the first step to be able to use CHIC consists in formatting the data. Many items ask the opinion of students with 4 or 5 possible answers: from poor to very good. Then, as many students did not answer some questions, we chose to remove these questions from the analysis because according to the missing values, ASI could produce very different implications.

The first computation is the similarity tree. This is a similar variables are. In this Fig. 1, there are some links in red. They represent the most significant levels. A significant level is more significant than the previous level and the next one. A mathematical criterion is defined to measure the quality of the partitions obtained at each level of the similarity. For more details, interested readers are invited to consult [1].

When many variables are involved a user can simply choose to remove some variables from the tree. In this case, the tree may be completely different since the associations between the classes depend on the variables.

The hierarchy tree presents implication between the variables. From Fig. 1 and Fig. 3 we can remark that even at the first level of the trees, the first classes are not the same. It is often the case. In fact, in the similarity tree, similar classes are built whereas in the hierarchy tree classes with the higher cohesion are chosen. These indexes are very different. As for the similarity tree, it is possible to see significant levels in the hierarchy tree. With SIA there are some specific issues that enable us to understand how the classes or the graph are built. In fact, when a user observes an interesting class in the similarity tree or the hierarchy tree or an implication in the implicative graph, he or she can compute the contribution of all the subjects on the creation of this class or implication. The most contributive subjects are those who have 1 for two variables involved in the implication or the similarity. The most typical subjects are those who are more similar to the class or the implication. For example, when a class has a middle cohesion or similarity or when an implication in the graph is not very strong, the most typical subjects are those who have not so similar variables or those for which implication or cohesion is not so strong. In addition to these features, it is possible to define supplementary variables. These variables do not change the creation of classes or graphs but they can be used to compute the contribution of these variables to the creation of classes or graphs. For example, sometimes we know if subjects are men or women. If we consider that this information is not correlated with the variables we study, we can define these variables as supplementary variables. In this case, we can use the SIA. Next with the results of SIA we may wonder if men or women are responsible or not for the creation of a given rule or implication. SIA allows us to compute the most contributive or typical supplementary variables. This feature will be illustrated in the analysis.

IV. Example of Analysis of a Form for the Future of ESPOCH Students

The Ecuadorian University schools are entering a process of assessment and accreditation of courses. In particular to meet criterion 2 (Relevance) and criterion 24 (Institutional Environment) meetings have been organized with graduates of years 2011, 2012 and 2013 of the Faculty of Sciences (ESPOCH). The meeting was held in November 2013.

Six schools of the ESPOCH are considered: Chemistry, Biotechnology, Biochemistry, Statistics, Biochemistry and Pharmacy, Chemistry and Engineering. Stratified random sampling was used. The final sample was of 154 graduates.

The self-administered questionnaire was applied, with 22 questions:

- 1 questions about informative data.
- 12 questions about training.
- 8 questions on labor situation.
- 2 questions about labor market.

Former students of the ESPOCH were given instructions on how to answer the survey. We also know if the students are men or women. All these variables are supplementary variables because we do not want these variables to be involved in the computation. Then students were asked to grade the quality of their formation.

Many conclusions can be drawn from this survey. Our goal is not to give a complete and exhaustive analysis of the survey. This would be very interesting but our goal in this paper is to show that using only item analysis response is not sufficient. So in this section we want to show examples of what SIA can provide. From Fig. 1, we can observe that there are similarities between some variables concerning the satisfaction of students for some criterion. For example there are strong similarities between:

- very good formation, very good knowledge and very good teaching performance.
- average teaching performance and average knowledge
- very good library and very good laboratory.

Fig. 2 illustrates an example of a similarity graph in which some variables have been removed temporarily to allow a user to focus on some variables. As explained previously this feature can be very interesting when there are many variables to enhance the understanding. With the similarity tree and the hierarchy tree, it should be noted that the tree may be completely different when the set of variables is reduced because a tree is built with the totality of available variables. It is not the case with the implicative graph as shown in Fig. 5.

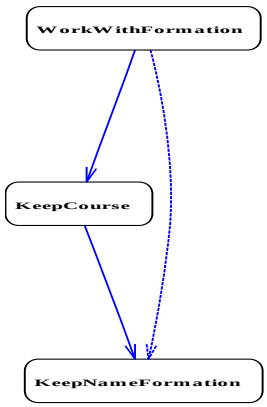

Fig. 5. An example of the implicative graph with only the three variables involved in the 2 rules with the most cohesion in the hierarchy tree.

Fig. 3 shows a hierarchy tree in which rules are oriented. The strongest rule is: if students think their courses were good, then they tend to think that the name of their formation is also good (KeepCourse -> KeepNameFormation). The second strongest rule is: WorkWithFormation -> (KeepCourse -> KeepNameFormation). This kind of rule is more difficult to understand. It can be something like: if students have a job related to their formation, then they tend to think that if their courses were good, then the name of their formation is also good. Then we can observe that the following and strongest rules are related to relations between students' satisfactions for some criteria (as with the similarity tree but with orientation). For example we can see:

- AverageClassrooms -> AverageLaboratory
- GoodInformationTechnology -> GoodLaboratory
- VeryGoodLibrary -> GoodClassrooms.

Fig. 4 represents an implicative graph with all the implications greater or equal to 0.97 in green and greater or

equal to 0.99 in blue. In opposition to the two previous trees, the implicative graph shows us all the implications and not only the strongest ones at a given level. With the hierarchy tree we observed the rule WorkWithFormation -> (KeepCourse -> KeepNameFormation). This rule is also present in the implicative graph. In Fig.4, we can only see the implication: KeepCourse -> KeepNameFormation. By default, in CHIC, quasi transitive closures are not displayed. We can display quasi transitive closures, the equivalent to transitive closure in logic with a simple option in CHIC. Of course there is no transitivity as in logic. That is why the term quasi transitive closure is maybe ambiguous. In Fig. 5, only the 3 variables involved in the previous rules are displayed. Moreover, the quasi transitive closures are displayed with a dot line. So in the graph we can see exactly the same information than in the hierarchy tree. And of course, we can find information that is not visible with the two previous trees.

V. Conclusion and Perspectives

In this paper we have presented the statistical implicative analysis and the CHIC software. The great advantage of SIA compared to other approaches is to give a clear sense of results because they are based on statistics and the meaning is very intuitive. The CHIC software implements most of the theory of SIA. It produces very intuitive results that allow users to look for interesting knowledge in their data. We applied an analysis on a survey given to former students of the ESPOCH in Ecuador who are now working. The goal is to see that SIA and CHIC give complementary information to classical item response analysis. In future work, as CHIC has been developed in C++ and is not portable, we plan to port it in the R framework. This will allow many users to use it and customize it more easily.

Acknowledgment

This work is partially funded by the "Prometeo Scholarships" and SENESCYT.

References

[1] R. Gras, E. Suzuki, F. Guillet, and F. Spagnolo, *Statistical Implicative Analysis*, *Theory and Applications*, *Studies in Computational Intelligence*, Springer, 2008.
[2] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, no. 2, ACM, 1993.
[3] R. Couturier, "CHIC: Cohesive hierarchical implicative classification," *Statistical Implicative Analysis*, vol. pp. 41-52, Springer, 2008.
[4] R. Couturier, R. Gras, and F. Guillet, "Reducing the number of variables using implicative analysis," *International Federation of Classification Societies*, Springer, pp. 277-285, 2004.
[5] I. C. Lerman, *Classification et Analyse Ordinale des Données*, Dunod, 1981.

Raphaël Couturier received his Ph.D. degree in 2000 in computer science from the Henri Poincare University in Nancy, France. From 2000 to 2006 he was an assistant professor at the University of Franche-Comte. Then he has been a professor at the same university. His research interests include parallel and distributed algorithms with a strong knowledge on asynchronous iterative methods, GPU and FPGA computing and data mining. Raphaël authored or co-authored more than 80 papers in conferences and journals and two books. He has also served in many program committees for conferences.

Rubén Pazmiño is preparing his P.h.D. thesis in the Salamanca University, Spain. He received his bachelor of mathematics in the ESPOCH University, Riobamba, Ecuador. He received a master degree in education and computer science in the University of Los lagos, Osorno, Chile and he also received his master in educational research in the UNL University, Loja, Ecuador. He has been a principal professor in ESPOCH since 1992. Rubén Pazmiño has created a statistical and mathematical modeling master degree and then he has created a master in statistics. His research interests include educational data mining, learning analytics, statistical implicative analysis and computational statistics. He has authored or co-authored some papers.