edutec
XVIII Congreso Internacional
RIOBAMBA-ECUADOR 2015
"Educación y Tecnología desde una visión Transformadora"

Statistical implicative analysis for educational data sets: 2 analysis with RCHIC

Análisis Estadístico Implicativo para datos de educación: 2 análisis con RCHIC

Eje temático: Ciencia, Tecnología e Innovación

Raphaël Couturier
University of Franche-Comté / ESPOCH
Raphael.couturier@univ-fcomte.fr
Rubén Pazmiño
Escuela Superior Politécnica de Chimborazo, Ecuador
rpazmino2009@gmail.com
Miguel A. Conde
University of Leon, Spain
mcong@unileon.es
Francisco J. Garcia
University of Salamanca, Spain
fgarcia@usal.es

Resumen

En este trabajo mediante dos ejemplos mostramos nuestro interés en la utilización del Análisis Estadístico Implicativo (SIA) en la comprensión de relaciones entre datos en Educación. Con SIA y la herramienta RCHIC es posible construir, gráficos (árbol de jerarquía, grafo implicativo) en los cuales el profesor o experto pueden visualizar y comprender las implicaciones entre los datos. Recomendamos a los profesores e instituciones utilizar SIA, debido a que ésta es una herramienta que permite encontrar posibles soluciones para mejorar evaluaciones, encuestas, etc.

Abstract

In this paper we show through two examples the interest of using Statistical Implicative Analysis (SIA) to understand the relations between data in Education. With SIA, it is possible to build, with the RCHIC tool, graphics (hierarchy tree, implicative graph), in which a teacher or an expert can visualize and understand implications between the data. We strongly encourage teachers and institutions to use SIA because this is a possible solution to improve evaluations, surveys, etc.

Palabras clave (máx. 5 palabras)

Análisis Estadístico Implicativo, Árbol de Jerarquía, Gráfico Implicativo

Keywords

Statistical implicative analysis, Hierarchy tree, Implicative graph

ISBN: 978-84-608-3627-8

1. Introduction

The goal of this paper is to show that Statistical Implicative Analysis (SIA) is very useful to analyze educational data and to understand the links between variables. SIA was created by Régis Gras more than 40 years ago. This method was designed to answer the question: “If an object has a property, does it also have another one?”. In practice, there are often some counter-examples. Nevertheless, this method aims at measuring this trend. In the following, even if there are counter examples for an implication, we will use the term implication instead of quasi-implication, for the sake of simplicity. SIA can be considered as a method that produces association rules Agrawal, R. and Srikant, R. (1994). However, its particularity is to be non-linear and robust to noise (when the number of counter-examples is low). Finally this method is defined with statistics and it measures the surprise of having so few counter-examples when considering two independent variables. The implication is inversely proportional to the surprise of having a small number of counter examples. Let us take an example. Suppose we have a population of 100 elements. If the cardinalities of A, B and the number of counter examples (A and not B) are respectively equal to 8, 9, 2, then the implication is very strong. In fact, if we take two random sets of the same size than A and B, then the probability of having 2 counter examples is very small, which is surprising. On the contrary, if the cardinalities of A, B, and the number of counter examples are respectively equal to 80, 90, 20, the implication is very small. In fact, if we take two random sets of the same size as A and B, then the probability to have 20 counter examples is not small at all. The implication between A and B, noted A => B is taken into account if the cardinality of A is smaller than B’s one. Otherwise, we have the implication B => A. For more information about SIA, interested readers are invited to consult Gras et al. (2008). SIA allows us to manipulate different kinds of variables: binary variables, categorical variables, frequency variables, fuzzy variables and numeric variables. The latter ones are partitioned using an algorithm initially presented in Diday, E. (1971).

The goal of this paper is to show that in many situations SIA can provide very interesting information for education datasets Couturier, R., Pazmiño, R. (2016). In order to use all the concepts provided by SIA, a library called RCHIC, is

XVIII Congreso Internacional EDUTEC "Educación y Tecnología desde una visión Transformadora" – Noviembre 2015
ISBN: 978-84-608-3627-8

available for R. RCHIC is the next step of CHIC, it stands for Cohesion Hierarchical Implicative Classification, Couturier, R. (2008). The R environment is a statistical framework in which many statistical methods and tools are available.

In order to convince the readers of the interest of using SIA to analyze educational data sets, we take two classical data sets and we highlight the results provided by RCHIC. For that we use the hierarchy tree and the implicative graph, two functionalities of SIA that will be explained.

2. Analysis of the exam scores from inner London

This data set was built in order to analyze the exam scores from inner London. This data frame contains 4,059 observations on the following 9 variables. This data set was created by Goldstein, H., Rasbash, J., et al (1993). In R, this data set is available with the mlmRev package with the name Exam. In the following we present the variables we have used (in fact we have removed the school identificator which is not interesting for this analysis and the variable intake because the authors of the paper removed it):

normexam: it is the normalized exam score.

schgend: it represents the school gender, there are 3 different types: mixed, boys and girls.

schavg: it is the school average of intake score.

vr: it represents the student level verbal reasoning (VR) score band at intake, there are 3 different types: levels are bottom 25%, mid 50%, and top 25%.

standLRT: it represents the standardised London Reading Test score.

sex: sex of the student, there are two types F for female and M for male.

type: it is the school type, there are two levels: Mxd for mixed and Sngl for single.

Among these variables, only normexam, schavg and standLRT are real variables that take their values respectively between [-3.7, 3.7], [-0.76, 0.64] and [-2.94, 3.02]. As previously mentioned, real variables are partitioned, in our case, we choose to use 3 partitions. Hence, for example normexam1 represents the students with the lower results, normexam2 represents the students with middle results and normexam3 represents the students with the best results. The other variables are binarized in order to split all the possible cases.

ISBN: 978-84-608-3627-8

So with the hierarchy tree, which makes a tree aggregating the different implications, many implications are not visible. Nevertheless the tree is built with the strongest implications. For example, in Figure 1, we can observe that the strongest implication is (sholavg 1 => vr bottom 25). This rule means that the school with the lowest intake score generally implies that the level of the student's verbal reasoning is also the lowest. The second strongest rule is (schavg 3 => vr top 25). This rule means, in opposition to the previous rule, that schools with the best intake score generally have students whose level in verbal reasoning is also the highest. The third rule (vr mid 50 => schavg 2) is similar since it means that middle students' level verbal reasoning generally implies middle school intake score. Even if these rules may be normal, we can clearly see that in the hierarchy tree. The fourth implication is (type.Sngl => schgend.girls), it means that if a school is not mixed, then generally it is a school with girls. Some implications are in red, it means they are significant. Such an implication is more significant than its previous implication and its next ones. According to the intensity of rules, the expert must decide when to stop his or her analyze because the last implications are less significant (this parameter is given in RCHIC).

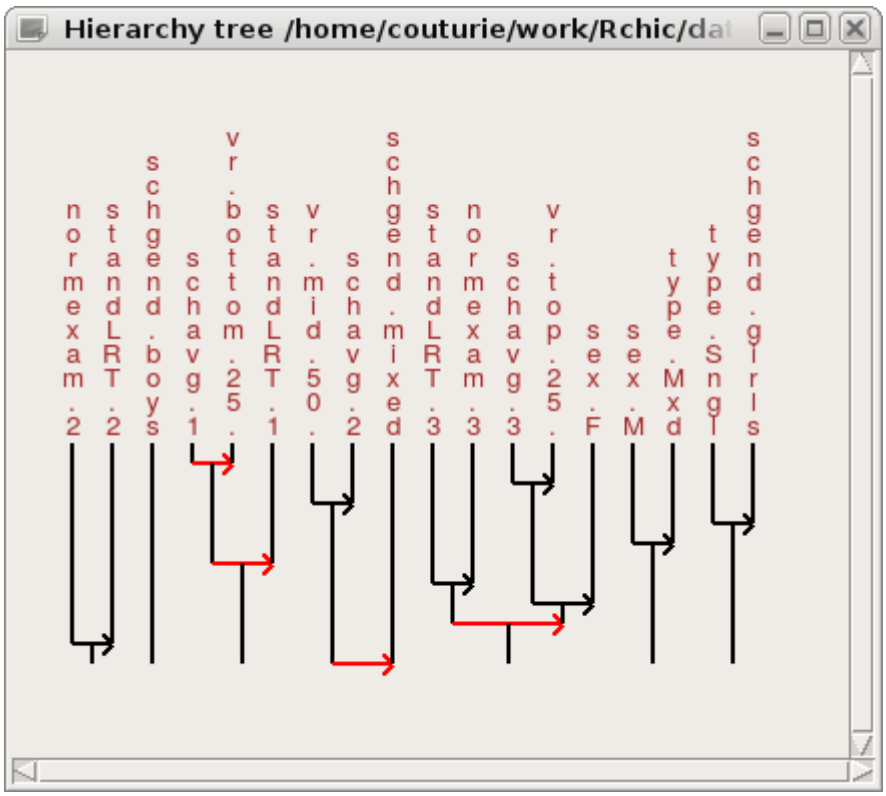

Figure 1 Hierarchy tree for the London data set

XVIII Congreso Internacional EDUTEC "Educación y Tecnología desde una visión Transformadora" – Noviembre 2015
ISBN: 978-84-608-3627-8

With the implicative graph, we can see all the implications and not only the most significant ones. Implications are of different colors according to their strength. With RCHIC, the threshold values can be changed. The previous implications seen in the hierarchy tree are visible in Figure 2. So, in order to display more things, in Figure 3, we remove some variables.

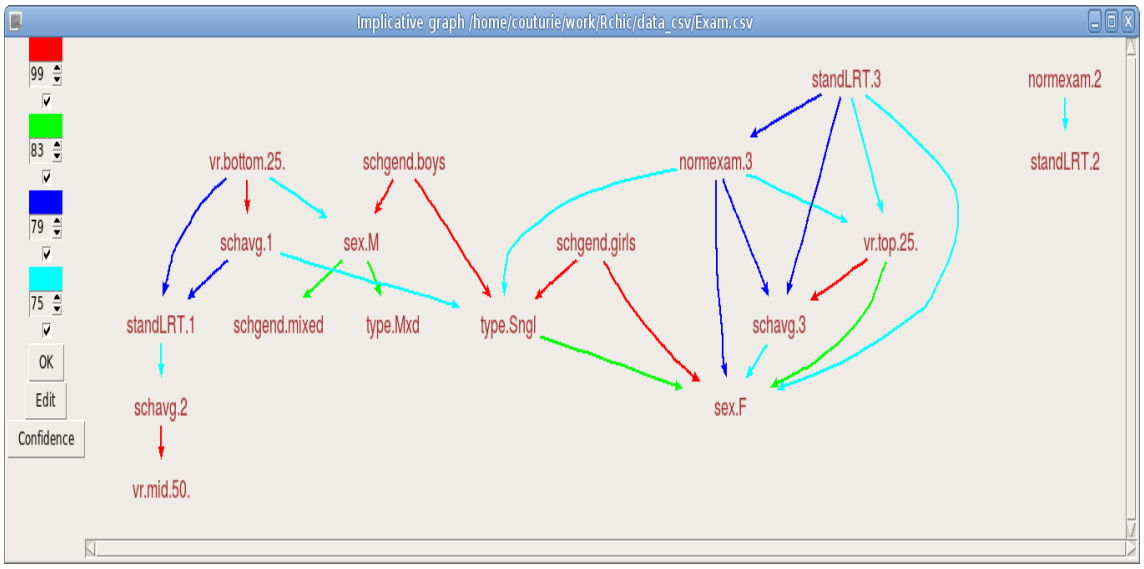

Figure 2 Implicative graph for the London data set with all the variables

Figure 3 shows the same implicative graph in which the 3 variables for schgend and the 3 variables for vr have been removed. So there is no more rule with high intensity. Besides, the confidence for each rule has been displayed. The confidence is the conditional probability that gives us the percentage of students that validate the rules. For example, we can observe the rules (normexam 3 => sex F). This rule has an implication intensity between 0.79 and 0.83, as the rule is in green. Moreover we can see that 64% percent of the students that have good results for normexam are girls. Combining SIA and confidence is very interesting because SIA selects the most surprising rules whereas the conditional probability gives us its percentage of validation.

XVIII Congreso Internacional EDUTEC "Educación y Tecnología desde una visión Transformadora" – Noviembre 2015
ISBN: 978-84-608-3627-8

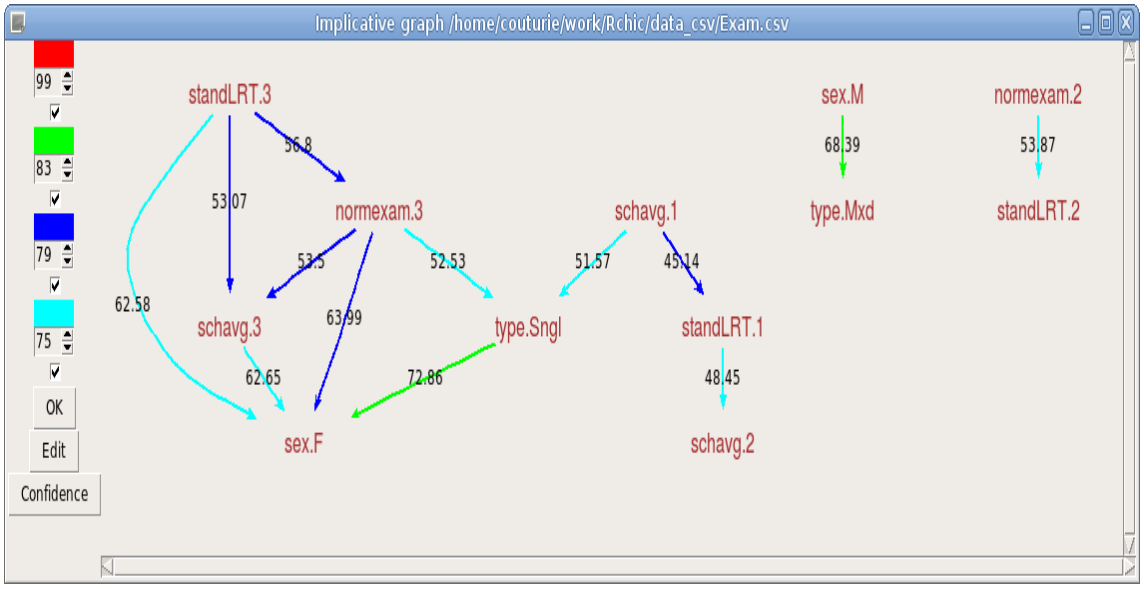

Figure 3 Implicative graph for the London data set with less variables

3. Analysis of the Student Teacher Achievement Ratio (STAR)

This data set was built in order to analyze the Tennessee's Student Teacher Achievement Ratio. This data set contains 26,796 observations on the following 18 variables. The results of this study are explained in Kaminski, R. et al (2003). In R, this data set is available with the mlmRev package with the name star. We choose to suppress some variables: id of the student, id of the school, id of the teacher, teacher race, student's ethnicity and the student's birth year. So the variables we have are described in the following:

gr: represents the grade, there are 4 ordered levels K < 1 < 2 < 3

cltype: represents the class type, there are 3 types: small, reg and reg+A. The last level indicates a regular class size with a teacher aide.

hdeg: represents the highest degree obtained by the teacher, there are ordered levels: ASSOC < BS/BA < MS/MA/MEd < MA+ < Ed.S < Ed.D/Ph.D

clad: represents the career ladder position of the teacher, the different types are NOT APPR PROB PEND 1 2 3

exp: represents the total number of years of experience of the teacher

read: represents the student's total reading scaled score

math: represents the student's total math scaled score

ses: represents the socioeconomic status, there are 2 possibilities: F and N representing eligible for free lunches or not eligible

XVIII Congreso Internacional EDUTEC "Educación y Tecnología desde una visión Transformadora" – Noviembre 2015
ISBN: 978-84-608-3627-8

schtype: represents the school type, there are 4 possibilities: inner, suburb, rural and urban

sx: represents the student's sex: M for male or F for female

yrs: number of years of schooling for the student, it is a numeric version of the grade gr with Kindergarten represented as 0.

In Figure 4, the hierarchy highlights the strongest implications for the STAR data set. We can see the following rules:

(hdeg Ed.D/Ph.D => exp 2) it means if the teacher has a high degree, then generally he/she has a middle experience, (clad APPR => yrs 0) it means that if a teacher has not a permanent position, then generally the student has 0 year of schooling in Kindergarten, (hdeg MA => math 1) it means that if the teacher has a master, then generally the students have bad results in math, (clad APPR => exp 1), it means that if a teacher has not a permanent position, then generally he/she has not a long experience, (schtype inner => ses F), it means that if the school in the inner city, then generally a student is eligible for free lunch, (clad.PROB hdeg.BS.BA) it means that if a teacher has not a permanent position, generally he/she has a small degree.

Figure 5 shows the first part of the graph from threshold .95 to threshold 0.80. As there are too many links, it is not possible to show the graph in one part.

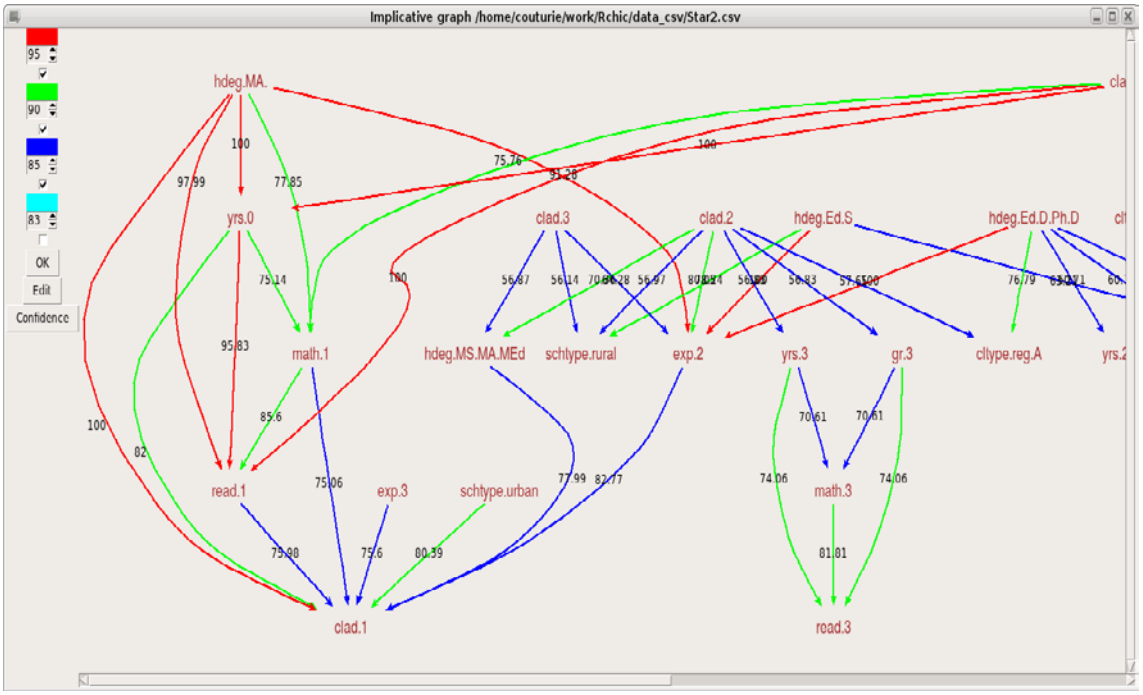

Figure 4 First part of implicative graph with the Star data set

XVIII Congreso Internacional EDUTEC "Educación y Tecnología desde una visión Transformadora" – Noviembre 2015
ISBN: 978-84-608-3627-8

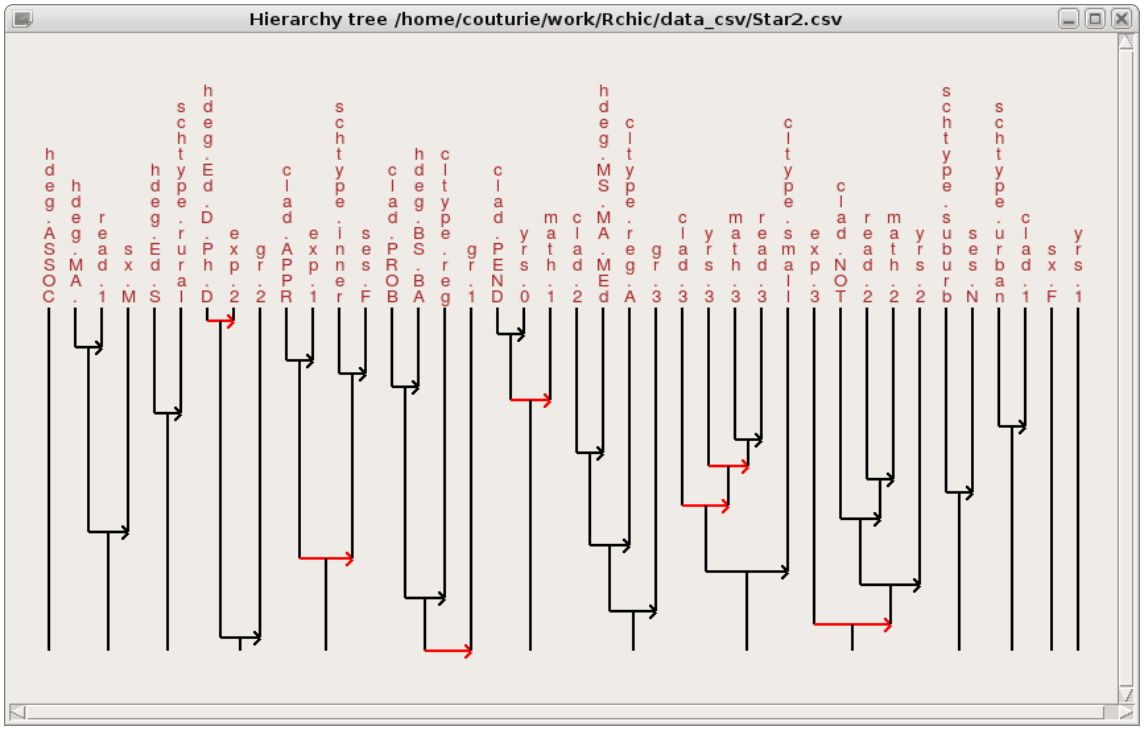

Figure 5 Hierarchy tree with the Star data set

There are many links in Figure 5, so we cannot comment all of them. Of course, in practice we can remove the variables we want in order to see the links more clearly. Here we keep all the links to show that there are many implications. All the links in red have very strong implications and confidences. For example, we can see hdeg Ma implies read 1, math 1, yrs 0, clad 1. This means that if a teacher has this degree, then generally a student is not strong in reading, math, he has 0 year of schooling and generally the teacher is in position 1 of his/her career. We can also see that hdeg Ma, hdeg Ed S and hdeg Ed D Ph D implies exp 2. It means that teachers with these degrees have generally a middle experience. We can also see that (math 1 => read 1) and (math 3 => read 1). The link between the level in math and in reading is quite strong, both for students with low and high results. Likewise, in order to have good results in math and reading, the grade is also important, we can see the implication (gr 3 => math 3) and (gr 3 => read 3).

In Figure 6, we present the other part of the implicative graph. So the complete implicative graph is composed of both figures. We can also see the implication (read 2 => math 2) that confirms what we mentioned just before. We can see that teachers without permanent position (clad APPR and PROB) have

ISBN: 978-84-608-3627-8

generally a small experience (exp 1) and a small degree (hdeg BS BA). Other interesting implications can be observed but we do not have enough space to comment all of them.

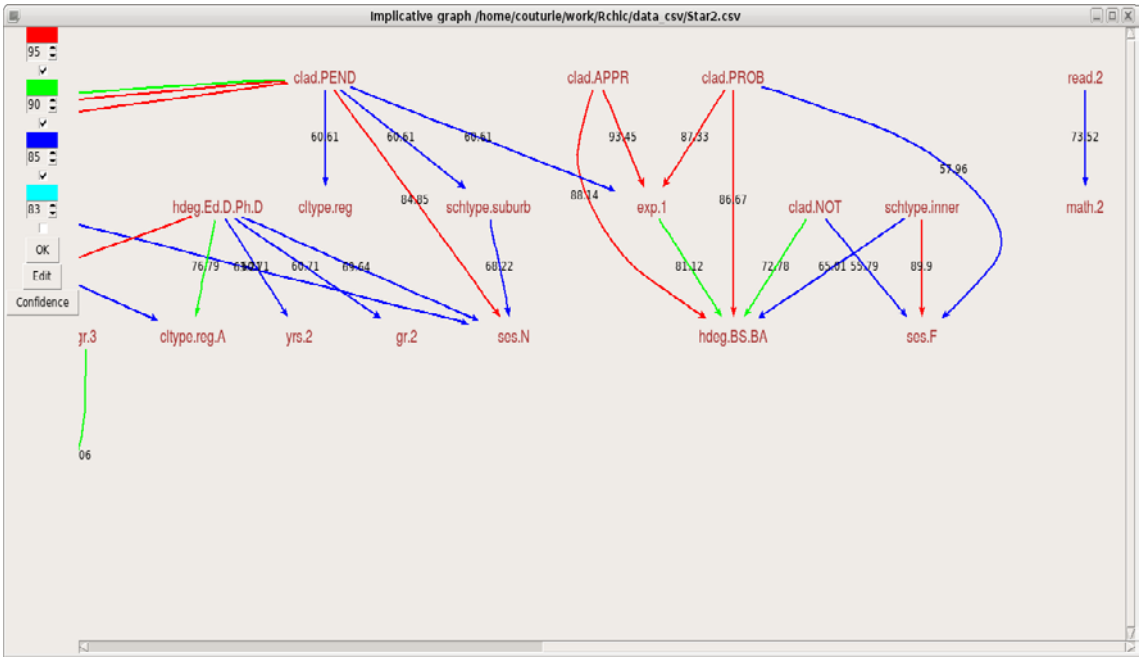

Figure 6 Second part of implicative graph with the Star data set

4. Conclusion and perspectives

In this paper we have presented the interest of using the Statistical Implicative Analysis (SIA) for educational data, of course it is possible to use it for other domains. SIA enables us to highlight the most significant implications in data sets.

With RCHIC, everyone can use the tools of SIA easily. Moreover, it is possible to see the percentage of people that are in accordance with the rules. Hence it is possible to have a better understanding of the relations between the variables of data sets. That is why we encourage the education community to use SIA in order to be able to highlight interesting concepts inside educational data and consequently to be able to improve future scores, surveys, etc.

In future works, we are interested to use other tools and method available in R, complementary to SIA, for example, in order to make for example predictions, dimension reduction with big data sets.

Acknowledgement

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

ISBN: 978-84-608-3627-8

References

[1] Agrawal, R. and Srikant, R. (1994), Fast algorithms for mining association rules, *Proceedings of the 20th VLDB Conference, Santiago, Chile*.

[2] Gras, R. et al. (2008), Statistical Implicative Analysis Theory and Application. *Book Springer.*

[3] Couturier, R. (2008). CHIC: Cohesive Hierarchical Implicative Classification, In Statistical Implicative Analysis, volume 127 of Studies in Computational Intelligence, pages 41-52. Springer.

[4] Diday, E. (1971), La méthode des nuées dynamiques, *Journal Revue de statistique appliquée*, vol 19, nb 2, pages (19-34).

[5] Goldstein, H., Rasbash, J., et al (1993). A multilevel analysis of school examination results. Oxford Review of Education 19: 425-433

[6] Couturier, R., Pazmiño, R. (2016). Use of Statistical Implicative Analysis in Complement of Item Analysis, International Journal of Information and Education Technology 6 (1), 39

[7] Kaminski, R., Stormshak, E. A., Good, R., & Goodman, M. R. (2003). "Prevention of substance abuse with rural Head Start children and families: Results of Project STAR." Psychology of Addictive Behaviors, 16, S11–S26.

ISBN: 978-84-608-3627-8