Pazmiño, R. A., García-Peñalvo, F. J., & Conde, M. Á. (2017). Is it possible to apply statistical implicative analysis in hierarchical cluster analysis? Firsts issues and answers. In C. Puente (Ed.), *Actas del Congreso Nacional de Ciencia y Tecnología (18-20 de enero, Riobamba, Ecuador)*. Ecuador: Escuela Superior Politécnica de Chimborazo.

Is it possible to apply Statistical Implicative Analysis in hierarchical cluster Analysis? Firsts issues and answers

R. Pazmiño[1], F. J. García-Peñalvo[2], M. Conde[3],
[1] Professor, [2,3] Engineer
[1]Faculty of Science, Polytechnic School of Chimborazo, Riobamba, Ecuador
[2]Education Faculty, University of Salamanca, Salamanca, Spain
[3]Engineering Faculty, Spain University of Leon, Leon, Spain

ABSTRACT

Hierarchical Cluster Analysis is used in clustering, prediction, and decision-making processes. However, in cluster analysis techniques it is difficult to test assumptions, implement the method and interpret the results. In order to enrich the hierarchical clusters techniques, this document proposes a new technique that takes advantage of the noise resistance, speed of calculation and results of easy interpretation that are characteristic of the statistical implicative analysis. To demonstrate how the implicit statistical analysis can be used in hierarchical cluster analysis, the authors have worked with images whose groups are simple to identify and value. This research carried out experiments, randomly grouping images between 2 and 63. These images were grouped hierarchically using two software and then the 35 students read and interpreted the results. Each student indicated the level of agreement on how the images were grouped, then a hypothesis test was performed on the sample of possible clusters to infer the global results. Approximately 70% of the students agree or strongly agree with the groups created with the implicate statistical analysis.

Keywords: Hierarchical cluster analysis, statistical implicative analysis, learning analytics, data mining, data mining education

1. INTRODUCTION

The present paper aims to analyze the use of hierarchy tree with a dendrogram in the hierarchical cluster analysis. Hierarchy tree is based on the implication intensity proposed by I.C. Lerman in [8] in the statistical implicative analysis. SIA was proposed by R. Gras [7]. The origin of this method was to find answers to conditional question: "If an object has a property A, does it also have a property B". The answer is strangely an affirmative one. But, it is possible to observe that there are broad trends. Statistical Implicative Analysis help us to find such tendencies. A dendrogram is a tree diagram used to illustrate the arrangement of the clusters produced by hierarchical clustering analysis. In order to achieve the aims of this research work the authors use a random selection of the feasible groupings with 63 possible images were used. With them a hierarchy tree was defined using chic software and the dendrograms using IBM SPSS Statistics.

Statistical Implicative Analysis is an association rules method [1], SIA aims at finding rules between the variables, subjects or objects. SIA has an interesting property, compared to others association rules methods, because it provides a non-linear measure that satisfies some important criteria. The method is based on the implication intensity that measures the astonishment degree of a rule. The implication index for binary variables is in Equation 1:

$$q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}} \qquad (1)$$

Where: 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. Given a set of data, CHIC extract the association rules. CHIC and SIA have been used in different areas, for review [1,4,5,6]. SIA and CHIC were enhanced by other kinds of variables. Currently, CHIC to handle frequency variables, variables over intervals and interval-variables. CHIC allows us to generate similarity trees, hierarchy trees and implicative graphs. The implication intensity is used to generate an oriented hierarchy tree. The hierarchy need a cohesion index, the index is defined with the implication. The implication is in Equation 2:

$$c(a,b) = \left(1 - \left(-p\ log_2 p - (1-p) log_2(1-p)\right)^2\right)^{1/2} \quad \text{if } p = \varphi(a,b) > 0.5\ and\ c(a,b) = \text{otherwise.} \qquad (2)$$

SIA have the following properties between the variables: the relationships between objects (variables or subjects) are dissymmetric, the association Measures are based on probabilities and are non-linear and the graphical representations follow the semantic of the relationship.

Intuitively a dendrogram is a graphic of the tree that shows a group of objects in different levels of similarity. Dendrograms can be applied to a variables and subjects. The variables can contain the information of images. The usage of simple images helps to understand the basic objective of the dendrograms and interpret them easily. Figure 1 is a hierarchy tree it shows different levels of cuasi-implication between variables that represent images. The hierarchy tree was realized in the software chic version 6.0. The images considered are Img01, Img02, Img03, Img04 and Img05. Let's notice its similarity with a traditional hierarchy dendrogram.

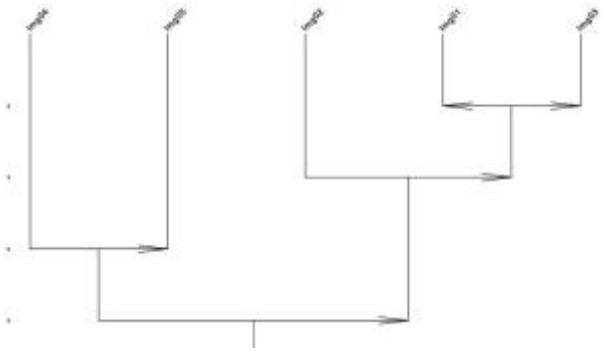

Figure 1. Hierarchy tree of Img01, Img02, Img03, Img04 and Img05 (generated by CHIC)

In the superior part of the figure the names of the variables are shown: Img04, Img05, Img02, Img01 and Img03. To facilitate the interpretation let's get closer to the term similarity for the cuasi-implication utilized such as a theory in the previous graphic. The images more similar are the ones that are grouped by the horizontal line higher than the one that unites Img01 and Img03. The double arrow when using images that indicates that they are the similar. This detail is not shown in general in the dendrograms generated by traditional techniques. The second horizontal line (from the top of the graphic) shows us that Img02 is the most similar to Img01 (or Img03) and this can be seen in the respective graphics that are shown in Figure 2.

Figure 2. Images Img02 and Img01

Figure 2. shows that the only difference between the two images is the spot that is on the chest of the puppy in Img02. The arrow that goes from left to right represents that there is a cuasi-implication be-tween Img02 and the class of images Img01 and Img03. Working with images means that the first image (Img02) contains almost all the elements that make it different from the second image (Img01). This detail is also not shown the dendrograms generated by traditional techniques.

The third horizontal line, shows that are like images Img04 and Img05. This can be observed in its respective images (Figure 3).

Figure 3. Images Img04 and Img05

Img04 contains two more spots at the end of the right posterior paw than in Img05. The arrow comes out from Img04 to Img05 that shows us that the details that makes them different are almost all found in Img04.

Finally, the last line of the hierarchy tree shows a similarity between the class formed by the images (Img04 and Img05) and the class formed by the images (Img02 and (Img01 and Img03). To say: $[\text{Img04} \rightarrow \text{Img05}] \rightarrow \big[\text{Img02} \rightarrow [\text{Img01} \leftrightarrow \text{Img03}]\big]$. Due to that the Img01 and Img03 are the same, any of them keeps all the details and it has: $[\text{Img04} \rightarrow \text{Img05}] \rightarrow [\text{Img02} \rightarrow \text{Img01}]$. Image 2 keeps almost all the details of Img01 and Img02 therefore: $[\text{Img04} \rightarrow \text{Img05}] \rightarrow \text{Img02}$. In the same way image 4 keeps almost all the details of images 4 and 5 therefore: $\text{Img04} \rightarrow \text{Img02}$. Finally, the last equation indicates that Img04 should keep almost all of the details of Img02 and therefore almost all the details of the 5 images as seen in the following.

Figure 4. Images of Img04, Img02, Img05 and Img03

In the Figure4. you can notice that Img04 contains almost all the details of the 5 images. The word almost is to indicate that the minor details are not considered such is the case of the spot in the chest of the puppy in Img02. One characteristic of the Statistical Implicative Analysis is the resistance to noise, is to say that it considers the most notorious characteristics and not the external factors or insignificant characteristics. For example, the spot on the chest of the puppy is a small characteristic considered with other spots. In the following it is shown the dendrogram applied to the previous 5 images. The software IBM SPSS Statistics version 21 was used that in a predetermined way works with the square of Euclidian distance.

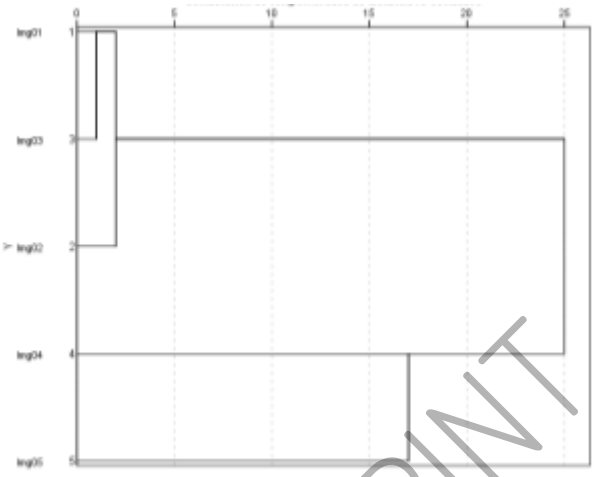

Figure 5. Dendrogram of images Img01, Img02, Img03, Img04 and Img05 (generated by IBM SPSS Statistics version 21)

Observing Figure 1. and Figure 5., we can notice that the dendrogram is the same permitting the same interpretation realized in the case of the 5 images. They only differentiate in the details of form. Clusters Analysis have been used to group students, for example [2] show that data sources can be incorporated and used to classify students in based of the set of beliefs that they bring to the learning situation. Clusters Analysis have been used in learners actions too, in [4] found some characteristic patterns in how learners use exploratory learning environments, and used this information to recognize more and less useful learners strategies.

2. METHODS

The purpose of this paper is to establish that the Statistical Implicative Analysis and in particular the hierarchy Tree can be utilized as a dendrogram. We will prove statistically the hypothesis that the proportion of students that feel they are strongly adequate or adequate is approximately 65%. We will prove statistically if the numbers of images that are part of the dendrogram influence the level of perception of adaptation. Furthermore, we will prove statistically if the software used to elaborate the dendrograms influences in the level of perception of the adaptation. All the tests will be done with a level of significance of 95%. The assurance that proves the previous hypothesis makes it viable that a study with more detail is needed to determine similarities, differences, advantages and disadvantages with other cluster methods that are similar. The population was formed by all the possible dendrograms that can be part of the 63 images. The dendrograms can be formed from 2 to 63 images. For example, if we take 2 images there will be a total of 1953 different dendrograms. With 33 different images taken from the 63 there will be 8,60778E+17 dendrograms. Adding all of the dendrograms that can be formed with 63 images in groups from 2 to 63 it has a population of 9,22337E+18. Because of time and human capacity is difficult that the students can interpret many dendrograms. We took random samples of 20 dendrograms, 10 to work them with the software Chic and 10 to work them with the software IBM SPSS Statistics. The formula 2 was utilized to calculate the sample.

$$n = \frac{Z_{\frac{\alpha}{2}}^2 pqN}{E^2(N-1)+Z_{\frac{\alpha}{2}}^2 pq} \qquad (2)$$

In the formula were entered, confidence level of 95%, p and q are 0.5, N is 9,22337E+18, margin of error is 0.31. The sample size is 10 dendrograms. The independent variables are numbers of images of the dendrogram and the method of generation of the dendrogram utilized by the software. The dependent variable is called adaptation and it measures the perception of the students over the adequate representation of the dendrogram. This variable is qualitative and it uses a scale of Likert. The 5 values utilized in the scale are: strongly adequate. For a better comprehension 3 new categories were grouped. The first formed by a strongly Agree and Agree, the second formed by neither agree nor disagree and the third formed by strongly disagree or disagree. The images were equalized before using them. All of them, are in white and black, their size is of 530 pixels in width by 759 pixels in height. They are all in the png format and they were added noise. In figure 3 shows some examples.

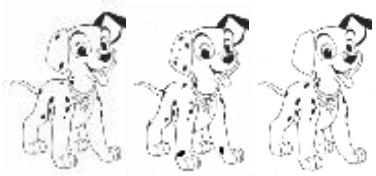

Figure 6. Examples of Images used in hierarchy tree and dendrogram

To study if there is a relation between the number of images in the dendrogram and the qualification of the dendrograms 5 groups where created in the following way: from 2 to 14, from 15 to 27, from 28 to 40, from 41 to 53 and from 54 to 63 images. The observation of the images, the dendrograms and the determination of the level agreed was in charge of 35 university students. It was carefully considered the group most homogeneous in the capacity of interpretation of the dendrograms. The students were capacitated in the reading of dendrograms for a total of 8 hours. Before answering the questionnaire, it was clearly explained to all of them the purpose of the study. The students worked in the same computer laboratory, with the same equipment for an approximate time of 1 hour. They all used the same software and digital resources. The surveys were realized using the questionnaire resources from the virtual class based on the Moodle system. Later all of the answers were exported to a digital sheet and this way the date base of statistics was formed. The analysis of the data was done using the software IBM SPSS Statistics version 21 and R version 3.0 to demonstrate the first hypothesis the parametric test t for proportions. The 35 students were in the same capacity of analysis of the dendrogram, but it cannot be guaranteed that they all worked with the same precision or motivation. They took as long as they wanted to answer the questionnaire; it took the majority approximately 1 hour.

3. RESULTS

The 69.14% of the students are agree or strongly agree in favor with the groups created by the dendrogram based on the hierarchy tree. Only 24.49% are strongly disagreeing or disagree. The rest 6.38% comment that they do not agree or disagree. The following diagram divides in detail the 5 initial scales.

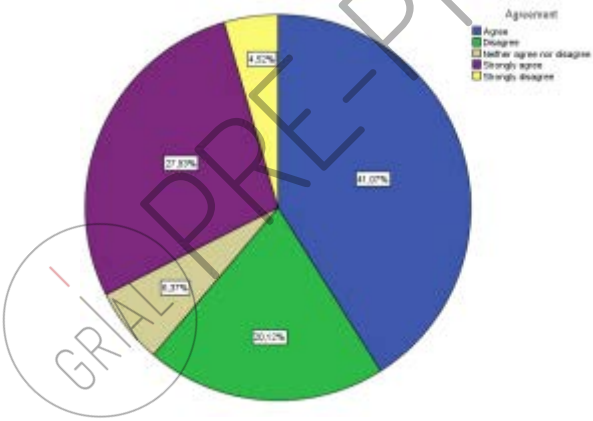

Figure 7. Pie Chart of percentages of the responses to the questionnaire

We did the hypothesis test of Agree and Strong Agree proportions.

$$H_0: \pi_{\text{agreeORstrongagree}} \leq 65\%$$
$$H_1: \pi_{\text{agreeORstrongagree}} > 65\%$$

We used the one sample proportions t-test without continuity correction in R software.

Table 1. t-test of Agree and Strong Agree proportions

X-squared:	3.654
Df:	1
p-value:	0.02797
95 percent confidence interval:	(0.6559121, 1.0000000)
sample estimates:	p=0.691358

p-value = 0.02797 is less than 0.05, then the null hypothesis is rejected and the alternative hypothesis is true, therefore the Agree and Strong Agree proportions is greater than 0.65.

$$H_0: \text{Dendrograms and Agreement are independent}$$
$$H_1: \text{Dendrograms and Agreement are dependent}$$

We used the chi-square test in IBM SPSS Statistics software.

Table 2. Chi-square test of independency between Simplified Agreement and layer

Simplified Agreement and Layer	Chi square gl Sig.	83.845 12 0.000

p-value = 0.000 is less than 0.05, then the null hypothesis is rejected and the alternative hypothesis is true, therefore Layer and Simplified Agreement are dependent.

Table 3. Percentage of answers in Agreement by software (1% did not answer)

	Software	
Agreement	**CHIC**	**IBM SPSS Statistics**
Agree	24%	15%
Disagree	12%	9%
Neither agree nor disagree	4%	2%
Strongly agree	16%	11%
Strongly disagree	3%	5%

We continue with the hypothesis test of independence between Software and Agreement.

H_0: Software and Agreement are independent

H_1: Software and Agreement are dependent

Table 4. Chi-square test of independency between Software and Agreement

Agreement and Software	Chi cuadrado gl Sig	20.934 5 0.001

p-value = 0.001 is less than 0.05, then the null hypothesis is rejected and the alternative hypothesis is true, therefore Software and Agreement are dependent.

Table 5. Comparisons of proportions of columns Agreement by Software

	Software	
Agreement	**CHIC (A)**	**IBM SPSS Statistics (B)**
Agree		
Disagree		
Neither agree nor disagree		
Strongly agree		
Strongly disagree		A

In the dendrograms population, percentage of IBM SPSS Statistics (5%) is greater than percentage of CHIC (3%) in the category Strongly disagree.

4. CONCLUSIONS

The 69.14% the participants in the experiment agrees or strongly agrees with the kind of grouping presented by the hierarchy trees. Only the 24,49% disagrees or strongly disagrees with this kind of grouping. We may assert, with 95% of confidence that, from the 8.60778e17 of different groupings with 163 possible images, the 65% of them are those best grouped by the hierarchy tree. From the research work carried out, we should note that the 69.14% of the persons that take part in experiment they understand it as a good way to carry out groupings. Dendrograms and agreement are dependent, and software and agreement are dependent too. The number of images grouped by using the hierarchy tree have impact in how easy it can be understood. That means that how it can be understood depends on the number of images. This has some associated questions, such as:

1) What is the number of images that produces a change in how a hierarchy tree is understood;
2) Are there other issues that affect how the hierarchy trees are understood?
The working time, the type of image, the preliminary analysis, etc.

5. REFERENCES

[1] Agrawal R., Imieliński T., Swami A.1993. *Mining association rules between sets of items in large databases*. ACM SIGMOD Record. vol. 22. No. 2. ACM.

[2] Amershi, S., Conati, C. 2009. *Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments. Journal of Educational Data Mining*, (1), 71-81.

[3] Beal, C.R., Qu, L., Lee, H. 2006. *Classifying learner engagement through integration of multiple data sources*. Paper presented at the 21st National Conference on Artificial Intelligence (AAAI-2006), Boston, MA.

[4] Couturier, R., Gras, R., Guillet, F. 2004. *Reducing the number of variables using implicative analysis. In International Federation of Classification Societies, IFCS 2004, Classification, Clustering, and Data Mining Applications,* pp. 277-285, Springer.
[5] Couturier, R., Pazmiño, R. 2014: *Use of Statistical Implicative Analysis in complement of Item Analysis*. International Journal of Information and Education Technology. IJIET. vol. 6, No. 1.
[6] Couturier, R. 2008. *CHIC: Cohesive Hierarchical Implicative Classification. In Statistical Implicative Analysis*, vol. 127 of Studies in Computational Intelligence, pp. 41-52. Springer.
[7] Grass, R., Suzuki, E., Guillet, F., Spagnolo, F. 2008. Statistical *Implicative Analysis. Theory and Applications*. Springer-Verlag, Berlin.
[8] Lerman, I. 1981. *Classification et analyse ordinale des données*. Dunod.