Article **Journal of Quantitative and Statistical Analysis**
June 2020, Vol.7 No.20 1-11

Clustering in data mining: CJ and K-Medias applied to an opinion survey on bullying at the UTNA

Clusterización en minería de datos: CJ y K-Medias aplicados a una encuesta de opinión sobre acoso escolar en la UTNA

MEDINA, Gricelda*†, LUNA, Francisco, TAVAREZ, Felipe and MARTÍNEZ, Rocío

ID 1st Author: *Sandra Maria, San Miguel-Iza*

ID 1st Co-author: *Roberto Aldo, Gonzalez-Zarazua*

ID 2nd Co-author: *Jose Alfredo, Camacho-Garcia*

ID 3rd Co-author: *Francisco Javier, Martinez-Falcon*

DOI: 10.35429/JQSA.2020.20.7.1.11 Received January 10, 2020; Accepted June 30, 2020

Abstract

In this work we applied unsupervised learning methods, to analyze the results of an opinion survey about bullying in La Universidad Tecnológica del Norte de Aguascalientes (UTNA). This study was applied to 131 students in the career of Communication Technologies at the University. We found the formation of 2 groups with different perceptions about bullying. In the first group of students they believe that bullying is mainly generated because they have family problems, and the other group, believes that bullying originates in the street, as result of, they come together with people who have a negative behavior. The objective of this study, is the use of unsupervised learning methods, applied to data analysis, specifically for this study, data from a survey on bullying perception. The methodology used to develop this study icluded, first the survey design on bullying, consulting many sources dedicates specifically on the bullying topic. After that, the data were introduced in the package data mining platform developed by R program, called R-Comander, and once the data was processed by the tool we proceeded to analyze the results.

Data mining, KDD, Hierarchical clustering, K-medias, Bullying

Resumen

En este trabajo se aplican métodos de aprendizaje no supervisado de minería de datos, para analizar los resultados de una encuesta de opinión sobre el acoso escolar, en la Universidad Tecnológica del Norte de Aguascalientes (UTNA). El estudio fue aplicado a 131 estudiantes en la carrera de Tecnologías de la Información y la Comunicación. En base a los resultados obtenidos, se descubrió la formación de 2 grupos con diferentes percepciones sobre el acoso escolar. Un grupo de alumnos consideran que el acoso escolar se genera principalmente porque tienen problemas en la familia, y el otro grupo, cree que la intimidación se origina en la calle, por juntarse con amigos que tienen un comportamiento negativo. El objetivo del estudio es el uso, de métodos de aprendizaje supervisado aplicados al análisis descriptivo de datos, específicamente para este caso de estudio, los datos de una encuesta sobre percepción del acoso escolar. La metodología empleada para el desarrollo de este estudio incluyó primeramente el diseño de la encuesta, resultado de la consulta de diversas fuentes de décadas al tema del acoso escolar. Después, se ingresaron los datos de la encuesta, en el paquete desarrollado por la plataforma del programa R, llamado R-Comander, y una vez procesados los datos por esta herramienta se procedió al análisis de los resultados.

Minería de datos, KDD, Clusterización jerárquica (CJ), K-medias, acoso escolar

Citation: MEDINA, Gricelda, LUNA, Francisco, TAVAREZ, Felipe and MARTÍNEZ, Rocío. Clustering in data mining: CJ and K-Medias applied to an opinion survey on bullying at the UTNA. Journal of Quantitative and Statistical Analysis. 2020. 7-20:1-11.

* Correspondence to the Author (Email: gricelda.medina@utna.edu.mx)
† Researcher contributing as first Author.

Introduction

At present there is a growing need to generate new theories and computational tools that help to extract useful information and knowledge from the large volumes of existing data, due to the constant use of technological advances in the management and generation of information. These theories and tools are topics concerning the process of knowledge discovery in databases (KDD, Knowledge Discovery in Databases), a term coined for the first time in the first KDD workshop in 1989 (Hand 01, Jiawei 06) and involving since the understanding of the application domain, going through data cleaning and knowledge extraction, to the use and application of that acquired knowledge. Data mining is one of the main stages of this process (KDD) Fig1, in which mathematical, statistical, or algorithmic methods are applied to the data, with the aim of discovering patterns and information hidden in them. This has earned data mining the attention of industry and society, due to the wide range of methods and techniques it offers for this purpose. The information and knowledge or patterns acquired through the data mining process have been used in multiple applications ranging from market analysis (customer segmentation, sales forecasting, risk analysis), fraud detection (credit cards). credit, telephone services, tax payments etc.), customer retention (study of consumption habits), to the exploration of science and medicine (Valenga 07, La Red 14, Jiawei 06) where in genetics, by analyzing changes in DNA sequences, it has been possible to determine the risk of developing diseases such as cancer, which has helped to improve the diagnosis, prevention and treatment of this type of disease (Perez 08).

This document is divided into 6 sections that contain the following: The first section talks about the results found in the application of the national survey on exclusion, intolerance and violence in upper secondary schools in Mexico. Section 2 contains a general description of the meaning of data mining and its classification, based on the type of tasks it handles. The third section describes the methodology that was used for the development of this study, detailing each technique used. In section 4 are the results obtained from the application of the survey and processed by the R-Comander program. Similar works found in the bibliography are mentioned in section 5, and finally in section 6 the final conclusions of the project are defined.

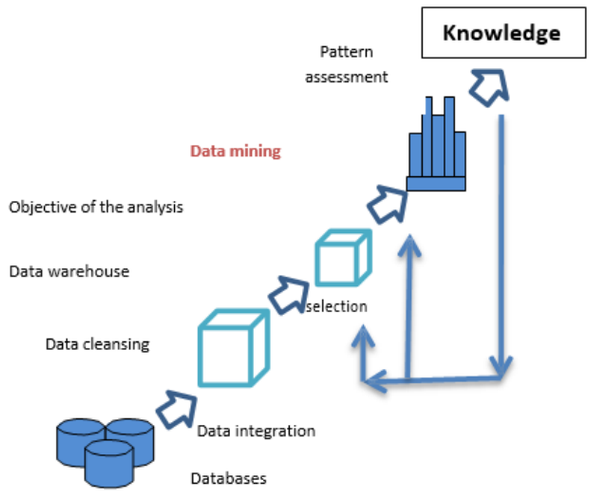

Figure 1 KDD process
Source: (Fayyad 96)

Bullying

Last year the Third National Survey on Exclusion, Intolerance and Violence in Higher Secondary Education Schools was applied by the Educational System of Mexico.

The survey was carried out with a probability sampling in 150 higher education schools, to obtain a sample size of 1,500 students.

The survey addressed issues related to interpersonal relationships in the space of the school community and provides a brief description of the social relationships between classmates, and of events that show signs of aggression or violence in these social relationships.

The data obtained allowed to establish the presence of psychological violence and situations of physical or verbal violence, to students at school, in 72% of the men and 65% of the women surveyed.

Although most of the reported cases were sporadic and only part of the students, they reported acts of violence that occur on several occasions. There are indications that students who had recurrent violence registered an increase in absenteeism, higher than 30% compared to students who have not suffered. Furthermore, 30.3% of the students surveyed do not consider school as a safe place.

ISSN: 2410-3438

That is why we consider bullying as an especially important issue in the educational field and that requires analysis tools that help us detect signs of violence in institutions, to implement actions that help us reduce acts and the consequences of these on students at any educational level.

Data mining

The purpose of applying mining techniques to data is to generally seek two types of tasks or objectives:

The description or prediction and each of these tasks are described below: (La Red 14, Shu 2012)

Descriptive or exploratory tasks (unsupervised learning)

The objective of this type of task is to partition or segment a set of data or individuals into groups. The groups are formed based on the similarity of the data or individuals in certain variables.

They are also known as unsupervised learning because they look for things in the data without guidance, they are exploratory and descriptive, and they involve looking in the data for common patterns of behavior (similar characteristics, preferences, behaviors, habits, etc.). In this type of task, the method autonomously discovers characteristics, correlations, and similar categories between them in the input data. They are techniques that start from a measure of proximity between individuals and based on a total population, seek to group the individuals most like each other, according to a series of measured variables. The characteristics to be covered are that the groups or clusters must be as tightly coupled or like the cluster (group), and the clusters as separate or different as possible from each other. There are several descriptive methods in data mining, focused on this type of task, among the most common are:

- Hierarchical Clusterization (CJ).
- K-means.
- Principal Component Analysis (PCA).
- OLAP (Online Analytical Processing).
- Factorial methods.

Predictive Tasks (Supervised Learning)

These types of tasks are intended to predict future or unknown values of the variables (sales volumes, potential fraudulent customers, good paying customers or not, etc.). They are also called supervised learning and their objective is to create a function capable of predicting the value corresponding to a variable, after having analyzed a series of examples (the training data). There are also several predictive-type methods in data mining, among the most common are:

- Time series.
- Discriminant analysis.
- Regression.
- Decision trees.
- Suport Vectors Machine.
- Bootstrapping methods.

Methodology

This work has focused mainly on the study of clustering methods: Hierarchical Clusterization (CJ) and K-means, using them to analyze the results obtained from an opinion survey on bullying. The survey was applied to 131 students of the career of Technologies of the Aguascalientes, and through the clustering methods, similar groups were found in terms of their perception about the issue of bullying.

The operation of both methods used are described below:

Hierarchical Clusterization

The objective of Hierarchical Clusterization or Automatic Classification is that, from a data table, where the columns represent the variables and the lines the individuals, a dendrogram is constructed, which is cut to identify the clusters or clusters and thus find information from them. Fig 2 (Hand 01). This clustering method is based on the idea of calculating the distances or dissimilarity indices of all against all (variables or individuals) in a table.

ISSN: 2410-3438

The dissimilarity index is a mathematical function that takes two individuals and assigns them a number between 0 and more infinite (0, + ∞), which has to fulfill the property of being symmetric, that is, that the distance between x and, is the same as between y and x.

$$d:IxI \rightarrow [0,+\infty] \quad (1)$$

And

$$d(x,y) = d(y, x) \text{ for all } x, y \epsilon I \quad (2)$$

Symmetry property

For the calculation of these dissimilarity indices, there are several formulas, such as the Euclidean distance formula, the squared Euclidean distance, or the Manhattan distance formula, among others, but the most common is the formula for the Euclidean distance, which was used to analyze the results of the survey (Hand 01, Mirkin 05).

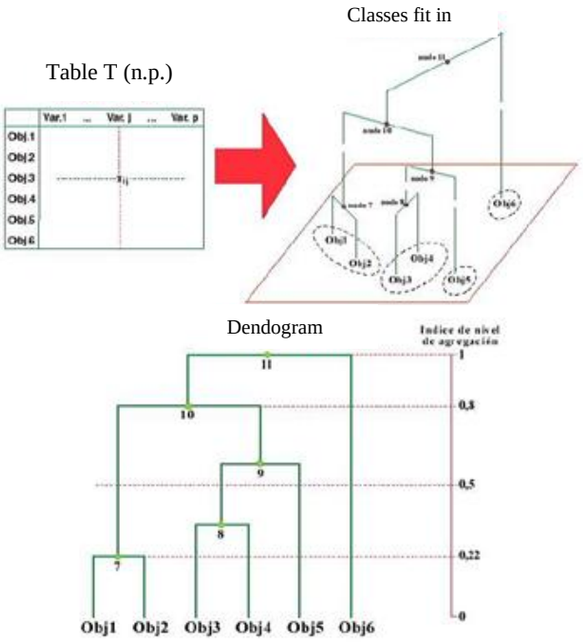

Figure 2 Scheme of the Hierarchical Clusterization (CJ) method

Euclidean Distance

$$d(Xi,Xs) = \sqrt{\sum j=1}(Xij - Xsj)^2 \quad (3)$$

Once the dissimilarity indices have been calculated through the calculation of the distances of all the individuals, a symmetric matrix of distances is obtained as a result, which is used to unite in pairs, those individuals whose dissimilarity indices are lower, then join the pair with the next lowest dissimilarity indices and so on to unite all the individuals.

Once the individuals have been joined, now, through an aggregation index, groups of individuals are joined, for which, there are also several formulas, such as Ward's aggregation index, the simple link formula, the link formula half, or complete, or McQuitty's among others, but the most common is Ward's aggregation index formula (Jiawei 06, Chambers 09).

The objective of the K-means method is the same as that of the CJ method, or the Ward's aggregation.

Once the groups of individuals are united

$$\delta_w(x,y) = \frac{(x)*(y)}{(x)+(y)} \left\| g_x - g_y \right\|^2 \quad (4)$$

The dendrogram structure is finally generated, and from this, the clusters or conglomerates of the data table can be found, to finally analyze them and find the information they provide (Valenga 07, Jiawei 06, Yakushev 14).

K-medias

It is the most used method for clustering and is generally used when considerable volumes of information are analyzed, this because the CJ method has an exponential growth problem in the calculations, since when performing the operations of the distances of all against all the individuals in a table, it can become a very computationally heavy process. For example, when analyzing a table when there are thousands of records.

ACP method, etc. Find clusters as homogeneous as possible between the individuals of each cluster and that the clusters between them are as different as possible from each other. The K-means method begins by randomly assigning each individual to a cluster, once all the individuals have been assigned, the next step is to calculate the center of gravity of each cluster, and then calculate the distances of all the individuals to their center of gravity. (Jiawei 06).

ISSN: 2410-3438

If some individuals are closer to the center of gravity of another cluster than to the center of gravity of the assigned cluster, a reallocation is made (they change clusters) and the clusters are rearranged again.

And the same procedure is carried out again, that is, the centers of gravity of each cluster are calculated again and if again there are individuals who are closer to the center of gravity of another cluster than at the assigned cluster, they are reassigned (they change clusters) rearranging the clusters. And so on, it continues to iterate, until there are no changes, or a maximum number of iterations indicated at the beginning of the process is exceeded, because, if there are, for example, a million records, the changes cannot be stabilized, so a maximum number of iterations is assigned for the algorithm to stop at a given time. Fig 3.

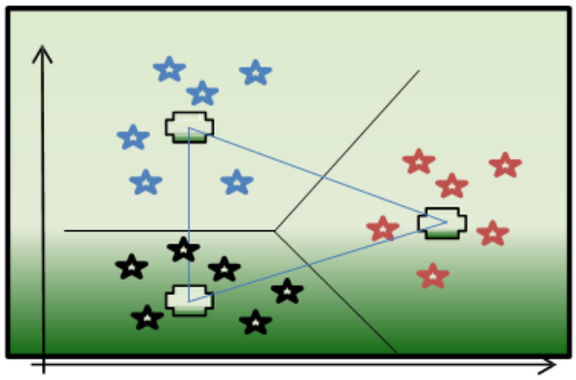

Figure 3 Scheme of the K-means method

The center of gravity of a cluster is calculated, with the vector average of the individuals that belong to the cluster, by means of the following formula.

Calculation of the total center of gravity

$$g_k = \frac{1}{[C_k]} \sum i \in C_k X_j \quad (5)$$

Where n = amount of data in the data table.

$$g = \frac{1}{n} \sum_{i=1}^{n} X_i \quad (6)$$

The term total inertia is a value that is calculated by averaging the distances of the vectors from the total center of gravity, which indicates the standard deviation of each vector with the general mean of inertia.

Calculation of total inertia

$$I = \frac{1}{n} \sum_{i=1}^{n} \|X_i - g\|^2 \quad (7)$$

The inter-class inertia term is the value which indicates how distant or different the classes are from each other, and its calculation is performed by calculating the distance from the center of gravity of each class to the total center of gravity, and then make a weighted average of those distances and square it.

$$B(P) = \sum_{k}^{k} = 1 \frac{[C_k]}{n} \|g_k - g\|^2 \quad (8)$$

Where Ck = Number of elements.

Its calculation implies the sum of the distances of the individuals that belong to a class to its center of gravity, carrying out the same procedure for the n classes, finally the averages are added and they are divided by the number of classes (Bullyinformate.org 15). Where: k = number of clusters. And gk = represents the center of gravity of each class. To calculate the total center of gravity, what is done is, average all the vectors and divide them by n, where n is the amount of data in the database (Husson 10, Mirkin 05).

$$W(P) = \sum_{K=1}^{K} I(C_k) = \frac{1}{n} \sum_{K=2}^{K} = \sum I \in \|Xi - Gk\| \quad (9)$$

The main objective of this, is maximize the distance between classes B (P) and minimize the distance between classes W (P), to leave clusters as different from each other B (P) and for the individuals of a cluster to be as homogeneous as possible W (P). With the K-means method it is possible to solve the combinatorial problem of the hierarchical classification method (CJ), since in the latter, the calculations grow exponentially, and the k-means method reduces its calculations to a polynomial time which makes it feasible to calculate (Husson 10, Graham 11, Fayyad 96).

The R-Commander program

The analysis of both methods was processed in the R statistical program through the graphical interface of the R Commander package, designed by John Fox, of the University of Hamilton, Ontario, Canada, which covers most of the most common statistical analyzes of R, through drop-down menus (Arriaza 08, Artime 13, Cena 15, Le 08, Torsten 09).

ISSN: 2410-3438

Results obtained

The opinion survey on bullying applied to the 131 students included 7 questions handling 26 variables to consider. This survey was designed considering the approaches and proposals of some social and governmental organizations specialized in the subject of bullying (Bullyinformate.org 15, Montaño 14, Pang-Ning 06, Merino 08, OCSE 15). For each question, students could only choose a single answer. Table 1 shows the content of the survey that was applied:

1. Bullying is: p1a) A hobby p1b) Something normal happens. p1c) It is an abuse and causes pain. P1d) Show strength and leadership.
2. Select the one that you consider the main consequence of bullying. p2a) Feeling fear. p2b) Low grades, drop out of school. p2c) It has no consequences. p2d) It can cause someone to commit suicide.
3. What would have to happen to fix this problem? p3a) It can't be fixed. p3b) That the teachers and families do something. p3c) That the classmates do something.
4. Why do you think that some bully others? p4a) For playing a joke or teasing. p4b) Because they mess with them. P4c) Because they are stronger. P4d) Problems in your family.
5. Where do you think bullying mainly originates? p5a) In the house. p5b) Outside with friends. p5c) It is the personality of each person.
6. What do you think of boys or girls who bully others? p6a) Nothing, I pass the topic. p6b) It seems wrong to me. p6c) It is normal for it to happen between colleagues. p6d) They do
7. Do you know anyone or know of any case of bullying in this institution? p7a) No. p7b) Yes.

Table 1 Opinion questionnaire on bullying

Figure 4 shows the binary data matrix, generated from processing the results of the application of the surveys. For reasons of simplicity, an identification key was used for each possible answer in each question, and as the survey was anonymous, no names of students were used, only a nomenclature that identifies the group to which each student belongs and a consecutive number to control the number of students who responded to the survey per group.

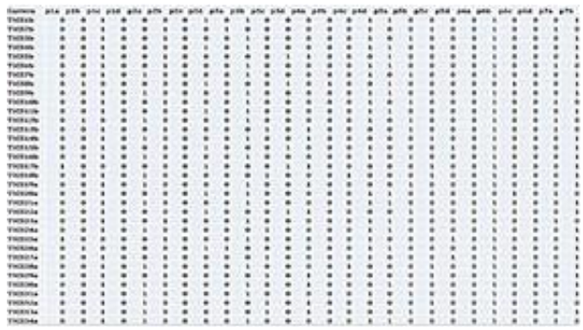

Figure 4 Binary data matrix

The results obtained by each clustering method are detailed below, which was used to analyze the information from the surveys.

Results obtained with the Hierarchical Clusterization (CJ) method

By means of the Hierarchical Clusterization method, the following was obtained: The binary data matrix was entered into the program and the dissimilarity indices of all individuals were calculated, using the Euclidean distance formula, generating as a result the symmetric matrix of distances shown in Fig. 5.

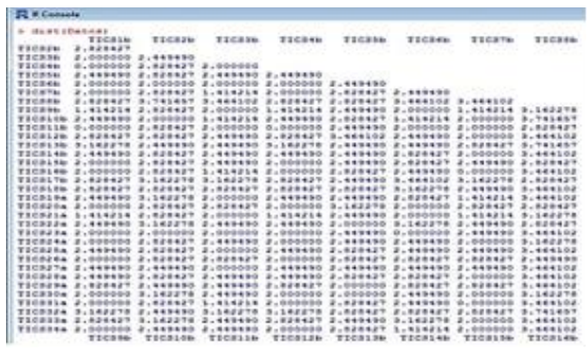

Figure 5 Symmetric matrix of distances

By using the Ward's aggregation index formula, to unite the groups of individuals, the following dendograms were obtained as a result, in 3D and 2D Fig. 6 and Fig. 7 respectively.

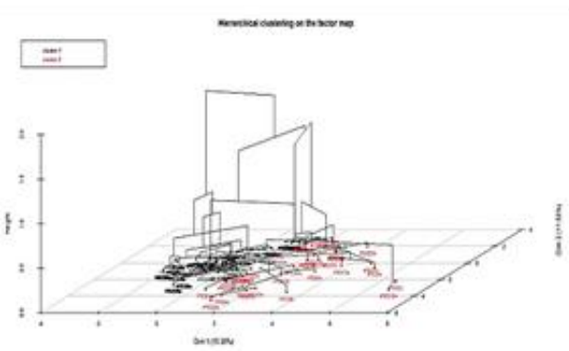

Figure 6 3D dendogram

ISSN: 2410-3438

Figure 7 Dendrogram

Once the dendograms were created in the CJ method, it can be clearly seen that the system generated 2 clusters or groups, group number 1 with 86 students and number 2 with 46.

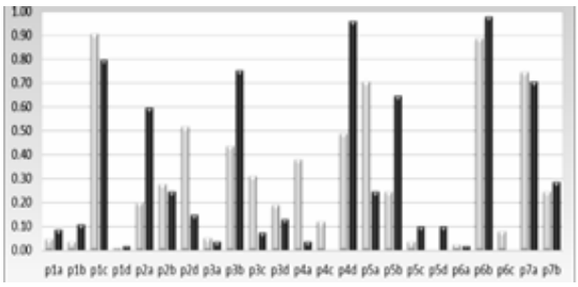

Figure 8 General graph

Fig. 8 shows the general graph where the characteristics that define each cluster are observed, finding very marked differences in questions 2, 3, 4 and 5.

Analyzing the results separately for each of these questions, it is necessary to: In question 2 (Select the one that you consider the main consequence of bullying), cluster 1 thinks that it can lead to suicide (52%) and cluster 2 considers that it causes fear mainly (60%). Fig. 9.

Question 2: Consequences of Bullying

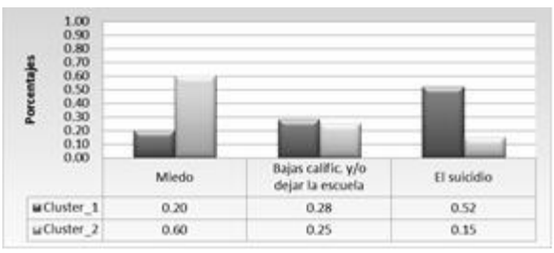

Figure 9 Histogram from question 2

In question 3 (What would have to happen to fix the problem of bullying?), Cluster 1 thinks that teachers and families (67%), including classmates (18%), should do something, while cluster 1 2 believes that the problem can be stopped by applying legal sanctions (39%) and is more in favor of peer intervention (30% compared to 18% in cluster 1). Fig. 10.

Question 3: How to stop the problem?

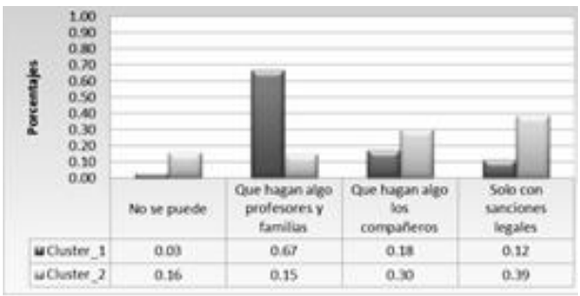

Figure 10 Histogram for question 3

Regarding the results of analyzing the answers to question 4.

(Why do you think that some bully others?), Almost 100% of cluster 1, thinks that some bully others because they have problems at home (98%), while cluster 2 thinks that they do it mainly to annoy or play a joke (71%) Fig. 11.

Question 4: Why do you think some people intimidate others?

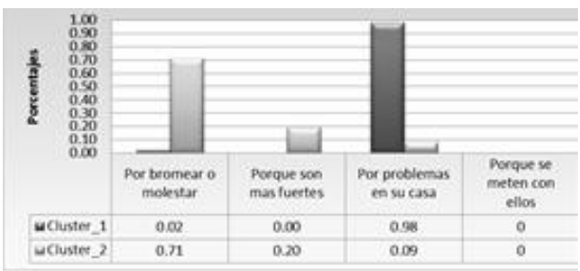

Figure 11 Histograma de la pregunta 4

Analyzing the answers to the question 5 (Where do you think bullying originates?), Cluster 1 thinks that it is at home (71%) and sometimes with friends (25%), while those in cluster 2 think the opposite, that it is mainly with friends (65%) and only some of them at home (25%). Fig. 12.

ISSN: 2410-3438

Question 4: Where do you think bullying originates?

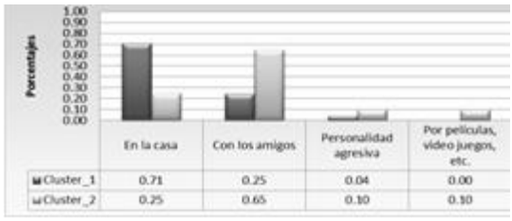

Figure 12 Histogram for question 5

Regarding the results of questions 1 (Is bullying?) And 7 (Do you know of any case of bullying in this institution), in both clusters the results are remarkably similar, as can be seen in Fig 13.

At least 80% of both clusters believe that bullying is abuse and causes pain. And 25% of all respondents in both clusters consider that they have suffered bullying at some point in the institution, which represents 32 students of the respondents.

Survey Question 1 and 7

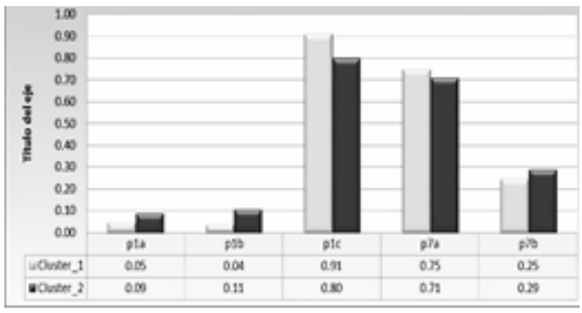

Figure 13 Histogram for question 6

Results obtained with the K-means method.

The binary matrix data was loaded, and the R-Comander program was told that by means of the K-means method, 2 clusters should be generated, just as it was generated automatically in the Hierarchical Clusterization method. In this method the clusters were formed as follows: Cluster 1 was created with 44 students and cluster 2 with 87, a result very similar to that obtained with the CJ method. For reasons of the tool and as the assignment of each individual to the cluster is random, the program inverted the number of clusters, that is, cluster 1 of the hierarchical classification method is represented by the number 2 in this method, and vice versa, the cluster 2 of the CJ method is represented by the number 1 in the K-means method.

In this method, the system generated the following biplot Fig. 14, where it indicates the trends of the responses in each question for each cluster.

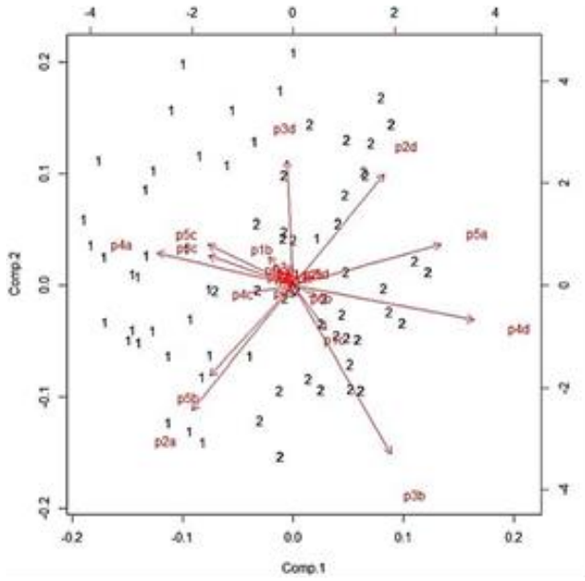

Figure 14 K-Means Biplot

The distances of the arrows on the biplot indicate the representation of the answers to each question. The closer the arrow is to the periphery of the square, it means that it was a highly selected response in the survey, and in its environment, the cluster number to which said selection belongs is found. The smaller the distance between the dates or they are closer to the center, it means that these answers were not among the most selected by the students.

Question	Cluster1 with K- means	Cluster2 with CJ
2. Main consequence of bullying	P2a) Feeling fear.	P2a) Feeling fear. P2a) Low grades.
3. What should be done to fix this problem?	P3c) The classmates must do something.	P3c) The classmates must do something. P3d) With legal sanctions.
4. Why do some bully others ?	Q4a) For playing a joke or teasing	Q4a) For playing a joke or teasing
5. Where does bullying mainly originate?	P5b) and P5c) With friends and because of the personality of each person.	P5b) With friends and because of the personality of each person.

Table 2 Cluster 1 Characteristics Comparison

ISSN: 2410-3438

As can be seen, cluster 1 chose the responses p2a, p3c, p4a, p5b and p5c, detailed in Table 2. On the other hand, in cluster 2 it can be seen that the most selected responses are: p2d, p3c, p4d and p5a, similar responses to cluster 1 of the CJ method.

Table 3 details each response, specifying the clustering method applied

Question	Cluster2 with K-Means	Cluster2 with CJ
2. Main consequence of bullying	P2d) can cause someone to commit suicide.	P2d) May cause suicide. P2c) Low qualifications.
3. What is due do to fix this problema	P3c) should do rather teachers and families.	P3c) they must do something. teachers and families.
4. Why Some intimidate others?	P4d) Because they have problems in your family.	P4d) Because they have problems in your family.
5. Where is primarily causes harassment.	P5a) It originates mainly in the house.	P5a) It originates mainly in the house.

Table 3 Cluster 2 characteristics comparison

The results of the answers in questions 1 and 7 are not visible in the biplot, which indicates that they are very close to the center, that is, that they are very similar in their answers and do not highlight considerable differences between them. Regarding the inertias obtained, the total inertia indicated a value of 2,856. The value of the inertia between classes B (P) was 0.473 and the inertia between classes W (P) was 2.383, adding both inertia gives a total of 2.856, which is the value of the total inertia. Thus proving Fisher's duality theorem, which indicates that the sums of the inter-class and intra-class inertia is equal to the total inertia of the point cloud.

Related jobs

There are several works related to the application of clustering methods used to analyze groups of individuals with similar characteristics. Among the works analyzed are the use of exploratory data mining methods to identify students at risk of dropping out or school failure (La Red 14, Márquez 12). The use of clustering to find patterns of criminal behavior, or patterns of drug use (Valenga 07), (La Red 14), (Yakushev 14).

Clustering work for Customer Segmentation (Jo-Ting 13), to support decision-making in business processes (Pérez 12, Pinzón 11, Sadat 15) and patterns for analysis of perception of corruption (Paulus 15) , or to analyze the quality of the information producers (Dinner 15).

Conclusions and future work

As can be seen, the results obtained in each question by both clustering methods are very similar, which has allowed us to obtain perception patterns on bullying that specifically define each cluster. This type of information can be extremely important for the Tutoring area or psychopedagogical counseling area within the institution. With the objective of establishing programs of talks and conferences focused on each group of students with similar characteristics. For cluster 2, awareness programs about the consequences of bullying and group integration techniques can be organized for those students who consider that bullying is generated only by making a joke or annoying, and who think that this type of behaviors, is generated mainly in the street with friends.

For the students that make up cluster 1, who think that bullying is generated mainly at home and that it occurs mainly due to family problems, family talks or guidance of special help can be focused on them when the case requires it. All this considering that, analyzing the last percentages of school dropouts, in the institution, they are not given by low academic performance, but by personal problems that affect the student to such a degree, that they decide to leave school, due to the consequences that these problems imply for them.

In relation to future work, the application of data mining supervised learning methods has an infinite field of application, since the importance of analyzing data and information to identify groups with similar characteristics or simply to describe the characteristics of the themselves, it is of utmost importance. Specifically for the issue of bullying, it is planned to analyze the issue, but now from the use of social networks, where through supervised learning methods, the comments made by students, in their publications towards other classmates, will be analyzed.

ISSN: 2410-3438

References

(Arriaza 08) Arriaza Gómez A. J., Fernández Palacín, F. López Sánchez M. A., Et al., Estadística Básica con R y R-Commander, (2008) Publicaciones de la Universidad de Cadiz. Recuperado: 10/06/2015, URL: http://knuth.uca.es/moodle/mod/url/view.php?id=1126

(Artime 13) Artime Carleos C., Corral Blanco N., Paquetes estadísticos con licencia libre, (2013), Revista electrónica de metodología aplicada, Vol. 18 No 2, pp. 12-33. Recuperado 9/06/2015. http:// www. Unioviedo .es /reunido/index.php/Rema. Departamento de Estadística e I. O. y D.M. Universidad de Oviedo.

(Bullyinformate.org 15) Fundación en Movimiento (Respetar para mejor convivir), A.C. http://bullyinformate.org/tests/test-escuela-segura.

(Cena 15) Cena A., Gagolewski M., Mesiar R., Problems and challenges of information resources producers' clustering. (2015). Journal of Informetrics, recuperado: 16/04/2015. www.elsevier.com/locate/joi, 273–284 ELSEVIER.

(Chambers 09) Chambers J. M., Software for Data Analysis: Programming with R (Statistics and computing), (2009). Stanford, ca. USD: Springer-Verlag.

(Graham 11) Graham W., Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!), (2011). New York USD: Springer-Verlag.

(Hand 01) Hand D., Mannila H. & Smyth P., Principles of Data Mining, (2001). A Bradford Book The MIT Press Cambridge, Massachusetts London England. Massachusetts Institute of Technology.

(Husson 10) Husson F., Le S., Pages J., Exploratory Multivariate Analysis by Example Using R, (2010). Chapman & Hall/CRC Computer Science & Data Analysis, Taylor & Francis Group, an inform business, Boca Ration London New York, CRC Press.

(Jiawei 06) Jiawei H., Kamber M., Data Mining Concepts and Techniques, (2006). Second Edition. Morgan Kauffman Publishers.

(Jo-Ting 13) Jo-Ting W., Ming-Chun L., Hsuan-Kai Ch., Hsin-Hung W., Customer relationship management in the hairdressing industry: An application of data mining techniques. (2013). Expert Systems with Applications, Elsevier. Recuperado: 22/02/2015 www.elsevier.com/locate/eswa

(La Red 14) La Red, D. L. & Podestá, C. E., Metodología de Estudio del Rendimiento Académico Mediante la Minería de Datos, (2014). Campus virtuales, 3(1), Revista Científica de Tecnología Educativa, Argentina.

(Le 08) Le S., Josse J., Husson F., "FactoMineR: An R Package for Multivariate Analysis", (2008). Volume 25, Issue 1, Journal of Statistical Software, American Satistical Association.

(Márquez 12) Márquez Vera C., Romero Morales C., Ventura Soto Sebastián. Predicción del Fracaso Escolar mediante Técnicas de Minería de Dato. (2012.). IEEE-RITA Vol. 7, Núm. 3, Nov.

(Merino 08) Merino González J., Revista de estudios de la violencia. Núm. 4, (Ene-Mar 2008). Instituto Catalán de Estudios de la Violencia (ICEV).

(Mirkin 05) Mirkin B., Clustering for Data Mining: A Data Recovery Approach, (2005) Chapman & Hall/CRC Computer Science & Data Analysis, Taylor & Francis Group, a Chapman & Hall Book, Boca Ration London New York: CRC Press.

(Montaño 14) Montaño J., Gervilla E., Et al. Técnicas de clasificación de data mining: una aplicación al consumo de tabaco en adolescentes. (2014). Anales de Psicología, vol. 30, núm. 2 633-641, May-Ago. Murcia, España.
(OCSE 15) Observatorio Ciudadano de la Seguridad Escolar. http://www.iea.gob.mx/ocse/default.aspx. (OCSE).

ISSN: 2410-3438

(Paulus 15) Paulus M., Kristoufek L., Worldwide clustering of the corruption perception, (2015). Physica A, www.elsevier.com/locate/physa 351–358, Procedia Computer Science Volume 29, ICCS 14th International Conference on Computational Science.

(Pang-Ning 06) Pang-Ning T., Steinbach M., Kumar V., Introduction to Data mining, (2006). Pearson Addison Wesley.

(Perez 08) Pérez López C., Santín González D., Minería de datos Técnicas y Herramientas, (2008). International Thomson Ediciones.

(Pérez 12) Pérez S., Puldón J. J., Espín Andrade A., Modelo clustering para el análisis en la ejecución de procesos de negocio, (2012). Revista investigación operacional, Vol 33, No. 3. Instituto Superior Politécnico José Antonio Echeverría.

(Pinzón 11) Pinzón L. L., Aplicándo minería de datos al márquetin educativo, (2011). Notas D Marketing, Escuela de márquetin y publicidad USA.

(Shu 2012) Shu-Hsien L., Pei-Hui Ch., Pei-Yuan H. Data mining techniques and applications – A decade review from 2000 to 2011,(2012). Expert Systems with Applications. Recuperado: 22/04/2015. URL: www.elsevier.com/locate/eswa

(Torsten 09) Torsten H., Brian S. E., A Handbook of Statistical Analyses Using R, Second Edition, (2009). Taylor & Francis Group, a Chapman & Hall Book, Boca Ration London New York: CRC Press.

(Fayyad 96) Fayyad U., Piatetsky-Shapiro G., and Smyth P. From Data Mining to Knowledge Discovery in Databases, (1996). AI Magazine Volume 17 Number 3.

(Valenga 07) F. Valenga, E. Fernández, Et al. Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina. (2007) Instituto Tecnológico de Buenos Aires, Argentina.

(Yakushev 14) Yakushev A., & Mityagin S., Social networks mining for analysis and modeling drugs usage, (2014). Procedia Computer Science Volume 29, ICCS 14th International Conference on Computational Science. Elsevier.

ISSN: 2410-3438