Prediction Using Inferential Statistics Programs

Dr. Luis Santiago Garcia Merino[1],
Dr . Segundo Cesar Tapia Cabrera[2],
Mgtr. Felicitas Eumelia Tapia Cabrera[3],
Dra. Blanca Yannet Avila Valdiviezo[4],
Dr. Alex Miguel Hernandez Torres[5],
Dra. Cecilia Eugenia Mendoza Alva[6]

[1]ORCID: https://www.orcid.org/0000-0001-9392-2474
Universidad Catolica Los Angeles de Chimbote
[2]Instituto de Investigación, innovación ciencia y tecnologia
ORCID https://orcid.org/0000-0003-1798-2437
Universidad Nacional de Tumbes
[3]ORCID https://orcid.org/0000-0003-0483-446X
Universidad Nacional de Tumbes
[4]ORCID: https://orcid.org/0000-0001-9090-5070
yanneavila06@gmail.com
Universidad Cesar Vallejo
[5]ORCID https://orcid.org/0000-0002-5682-2500
alex.hernandez@upn.pe
Universidad Privada del Norte
[6]ORCID https://orcid.org/0000-0002- 3640-2779
Universidad Cesar Vallejo
ceciliae@ucvvirtual.edu.pe

ABSTRACT

The purpose of this Final Degree Project is to delve into a very interesting concept related to obtaining and extracting relevant information that we can find in a data collection. We are talking about Data Mining. Under this name all those techniques that help us extract relevant knowledge and information that are implicit in the databases are grouped. Raw information can be much more useful and easier to work with if it is ordered, classified and divided or grouped into common concepts. These two tasks are addressed by data mining, it provides us with tools that classify and group this "raw" data and thus be able to get the most out of it. However, it is not the only thing we can achieve by applying different data mining methods and techniques. Through these mechanisms of calculation, association and segmentation, through the search for common patterns in the data, situations that are always repeated

or implicit "rules" in the data itself, we are able to predict different situations or data that we are going to receive in a future. The classic example for this is the shopping cart. Through simple methods of analyzing purchases made in any supermarket, looking for patterns of behavior and, as we have mentioned, situations that are repeated on several occasions, we can predict that, for example, when someone buys hamburgers, there is a high probability that they will also buy hamburger bun. At first glance it may seem like a somewhat obvious prediction, but it is not so obvious when patterns of behavior of people when buying are discovered and, always by analyzing the data obtained, we reach the conclusion that placing supermarket products in one distribution or another can become very relevant when registering more purchases of some products or others.

Keywords: data mining, opinion mining, prediction with weka, behavior, patterns, rules for generalization.

INTRODUCTION

Information is always considered to be able to do and have; The human being has always tried to know and investigate in depth everything that surrounded him to make the most of his possibilities of progress and success, and for this, having exclusive and relevant information has always been of great help. From the first mathematicians with their statistics and probability tables, anticipating the events that could happen was key for the human being. In this way, they could model the world around them, try to adjust it to a series of patterns that often repeated over time, and in this way, take advantage of this "knowledge". Today, we live in a world saturated with information. We have technological tools that put vast and huge amounts of information and data at our fingertips. The expansion of the Internet and information systems has considerably revolutionized our ability to obtain information easily and quickly. However, "with the unprecedented degree of growth with which information is collected and stored electronically today in virtually every field of human behavior/development, extracting useful information from all available data is becoming a growing scientific challenge and a massive economic necessity." [Zaki and Ho 2000]. It is estimated that the amount of information in the world doubles every 20 months [AI Magazine]. This is where technological development at the computational level comes into play, better computers with which to develop exhaustive analysis of the data in search of relevant information, relationships between the data, etc. Thanks to this development and the growing need to filter and organize these

amounts of data, a concept called: KDD for its acronym in English: Knowledge Discovery in Databases. In this chapter we will introduce you to the most important concepts of one of the steps that make up the KDD process, data mining. The rest of the steps of the process will also be developed, however it will be a very superficial analysis because it escapes the objectives of this project.

In this sense, according to the phrase that refers to "Ignorance is the curse of God, knowledge the wing wherewith we fly to heaven" by William Shakespeare; It takes into account that data mining consists of "the application of techniques in large volumes of data to discover useful, applicable and non-trivial information". This definition, applied to a more business environment, could be reconstructed as "the set of methods, which together with a deep knowledge of the business, are aimed at identifying, in large volumes of data, relationships and trends hidden so far" (Carlos Creus, 2006). We can say that data mining is a process within a process that encompasses everything, the KDD. In this step, data mining is responsible for looking for relationships and patterns between all the amount of information available.

On the other hand, it is important to consider that a pattern is something that is repeated, a trend, as a representation of the data and information obtained from a source of information, such as a database. A pattern must meet a series of characteristics so that it is useful when working with it and obtaining useful information.

And it has the features, which allow us to meet our expectations of searching for information. It would be useless to know that when it rains we get wet, if what we are looking for is to know when it is going to rain. • It must be applicable, that is, it must be able to adapt to a large amount of the data we have, in order to be relevant, the more data that meets that pattern the better. • It does not have to be trivial, it must provide us with some kind of useful knowledge for what we are analyzing. • It must be new and unknown before applying the methods to discover it. • It should be understandable, twisted patterns that relate data to each other based on complex and "far-fetched" "interrelationships" are not useful to us. To obtain these patterns and be able to obtain relevant and useful information, data mining has several methods and algorithms, which applied to large amounts of data are able to discover these new patterns and hidden trends. These methods can be classified into two large groups, according to the information we obtain by applying them conveniently. Thus, we can divide them into predictive methods and descriptive methods. Predictive methods include the use of some variables or fields of the database to predict future or unknown values, or even other variables of interest. Descriptive methods focus on finding patterns understandable to humans that describe the

information we have. Although the boundaries between some methods and others are not clearly defined, since some predictive methods can be descriptive and vice versa, the distinction is useful to understand the overall objective of the discovery process. There are many methods, we will go on to make a brief introduction and classification of the methods that we will talk about in more depth throughout the document.

Among the predictive methods known in the world, there are decision trees and methods based on mathematical regression. Decision trees can be used to know if, for example, one day we can go out to play tennis using a history of weather data of the days we have been able to go out to play and those in which the weather has not allowed it as a base. 14 Regressive methods can be used to predict customer purchases by age groups, given a history of purchases by age for a range of ages, or even the price of a second-hand vehicle if we are based on a relationship of data on second-hand cars of similar characteristics with their corresponding prices, characteristics and attributes; Clustering is a method by which we discover groups and structures in the data and that to some extent are similar or fulfill similar characteristics without using known structures in the data.

So it is important to specify that the data mining process depends closely on the method or technique that we are going to use to solve the problem or the requirement of information that has been presented to us. Predictive methods usually require "training" to be able to model the rules that must be applied to new data for prediction, as well as some other verification steps to check the accuracy of the model obtained. However, there are other methods that only need to be executed on a collection of data to obtain results.

As I mentioned, I consider data mining as part of a larger business and data processing task called Knowledge Discovery in Databases (KDD). Many experts in the field agree that the way to gain "insights" from raw information can only be achieved through process-modelled techniques. Placing data mining methods strategically at the center. However, for mining processes and methods to provide us with conclusive and useful results, the preliminary steps of preparation of the information and the post-processes that verify the information obtained are essential. These additional tasks make up the KDD process.

KDD is solely the concept of a multi-step process that identifies patterns in data to find new information. Data mining is just one of those steps in the process of applying computational techniques to find such patterns in data. This step consists of the use of algorithms that provide patterns in an acceptable response time, always obtained from data collections such as databases. Other steps in the KDD

process are comprehensibility and validation of discovered patterns. KDD is the concept and data mining is its tool.

The process of knowledge discovery in databases is iteractive, as it consists of several steps that may have to be repeated to extract the optimal information, and interactive, as it includes several steps where the intervention of an expert user is essential. In 1996, Brachman and Anand proposed a practical view of the process, emphasizing the interactive nature of it. Below we will outline the basic steps of the process: Development and understanding of the work context. Identify the objective of the KDD process from the point of view of the information required. Group a set of data to serve as a process objective. Select a group of variables, a subset of data, and so on. Data cleansing and preprocessing. Eliminate useless data, decide strategies to handle fields with empty fields, collect the necessary information.

Data reduction and projection. In this way we can obtain a more appropriate way to represent our data collection. Improve the efficiency of data by removing or combining variables, or leaving out invariant data. Decide the appropriate data mining method for the data we want to obtain using the KDD process. Analysis. Here you decide which models and parameters may be suitable and decide which exact method agrees with the overall objective of the process. Data Mining. Search for patterns of interest. Interpretation of the patterns obtained, possibly returning to any of the previous steps (iteration). Use and implementation of the knowledge obtained. Verification of the data obtained and other technical checks. The process can involve a significant iteration, that is, we can find several loops between any of the steps or states of which the process is composed. In the figure below we can see a scheme where the basic steps of this process are detailed. Most of the research and published documents focus on step 7, data mining, however all the steps of the process are equally important for obtaining useful and quality information and data.

Tareas	Métodos
Predicción y descripción	Á de árboles de decisión, análisis cesta de la compra, análisis de series temporales, redes neuronales, tecnología de agente de red
Clasificación	Análisis cesta compra, árboles de decisión, redes neuronales, ordenamiento
Regresión	Regresión lineal, regresión logística, regresión multinominal
***Clustering* (Agrupamiento)**	Análisis de grupos, redes neuronales
Summarization	Algoritmos genéticos
Modelado de dependencias	Análisis de la varianza, análisis de enlace
Cambio y detección de desviación	Lógica difusa

Figure 1.- Forms of prediction using data mining or the so-called opinion mining.

A decision tree is used as a classifier to determine an appropriate action or decision (from a predetermined set of actions) for a given situation. A decision tree helps us correctly identify the factors to consider and how each of these factors has historically been associated with decision outcomes. [SAPDOCS]. The schematic vision of this method makes it one of the simplest methods of interpreting and assimilating the information they contain. It is called a decision tree because the result of the model is represented in the form of a tree. Decision trees are a method classified as supervised learning methods, since they must be trained with information that contains a history of the data itself and the results that have been a consequence of said data in order to be used to create predictions. To verify these predictions obtained as a result and check the accuracy, we can run the trained model against another known collection of data to evaluate the accuracy of the trained model. The steps it follows is given by: 1. Training. The tree is modeled to represent the patterns detected in the data history as best as possible. 2. Evaluation. In this step, totally optional however, we can test the validity of the trained model by confronting it with another different collection of data (same theme and same content, but different in itself). If the precision achieved is not as desired, we must redesign the model and repeat the process. 3. Prediction. Finally, we obtain the predicted result from the designed model, that is, the value or values, or the decision we seek to make, for a given case for our data set. With this we can generate the graphical representation of the tree. The tree is constructed with the following components: Root node • : As a single node, it forms the entry point of the tree. Usually the highest point. • Decision nodes: These act as routers to decide which branch we should take while we go through the tree from top to bottom. • Leaf nodes: These nodes are those that do not contain any node with "success", that is, nodes where the objective is met, or where the value we are trying to predict is made positive.

Clustering is used to group data into well-cohesive and defined sets. We can differentiate it from normal classification methods in the following fact: the classes in which the data is grouped are not predefined as in the normal classifications, but are determined from the data. It is an unsupervised learning method. The results that we can obtain by applying this method can be used to summarize and analyze the contents of a given data collection considering the characteristics of each set rather than the characteristics of each

record. This method can be used descriptively as well as predictive (to which group a new data will belong).

CONCLUSIONS

It should be noted that data mining is a tool with incredible potential and applicable to many projects, circumstances, realities and purposes. Through this project, it has been possible to introduce a little into this world of obtaining relevant information from masses of compact data. We have introduced the concept of data mining, as well as analyzed several of its numerous types and families of calculation and results extraction methods. Obviously, you can go much deeper into data mining, but it escapes the objective of this project. However, we have been able to learn to use a very powerful tool with OpenGPL license, GNU, among others that serve for data mining called Weka, we have verified and reviewed its characteristics and we have sailed over its numerous possibilities and functionalities. Through this powerful application, it has been possible to analyze and apply in a practical and constructive way several of the data mining methods that are implemented in it. We have been able to analyze the results and discern between influential and non-influential attributes, we have filtered instances of the large mass of data obtained and we have created new mining models for our sample dataset. When analyzing the results given by these new models, we had to understand and interpret these results, to discern which model best suited our needs and obtained the best results. However, all this would not have been possible without the realization in Visual Basic for Applications and Access a small but powerful application that has been able to download information from more than 6000 vehicles and organize this information to store it correctly in a database in just a few minutes. In addition, it has provided us with an ideal environment to implement the mining model obtained through the M5P algorithm and thus implement the prediction simultaneously and effectively. I feel that I will repeat myself if I continue with all the conclusions I have drawn from the realization of this project. If I can say, however, that it has served me more than to achieve my personal goal, which was none other than to enter the world of data mining. Since I did not have the opportunity to delve into this series of concepts in any of the subjects of the degree, it has been very helpful to be able to carry out this project to solve my doubts and concerns about this matter. Likewise, these new skills and knowledge acquired will be very useful in my future professional life, because as has been commented on numerous occasions in this report, data mining is a tool of vital importance in the business world. World to which I am strongly linked, since I have been lucky enough to enter a multinational company as a computer specialist and where I am sure that all this knowledge

acquired will be of great help to progress professionally. I cannot forget, that my abilities and skills for programming and abstraction of algorithmic processes have improved, as well as my level in the Visual Basic programming language. On a personal level, I can say that removing the theoretical-practical knowledge that I have learned, I have been able to learn to value the effort and dedication involved in the development of an application from scratch, being in charge of taking needs, analyzing the situation, development, implementation, testing; It has been possible to understand in a better way, the entire cycle involved in the implementation of an application, that is, its life cycle, a cycle that we have heard so many times throughout the degree. I would like to say that my teamwork has improved, however, due to incompatibilities, I had to choose to carry out the project individually, however, my project manager has been of vital importance to me solving my doubts and guiding me along the way when I did not know what exit to take exactly.

REFERENCES

- J. Hernández, M. J. Ramírez, C. Ferri "Introduction to Data Mining" © Prentice Hall / Addison-Wesley, ISBN 84 205 4091 9
- BW380 Data Mining "SAP Business Intelligence: Analysis Processes and Data Mining" SAPDOCS
- BW310 "Data Warehousing" SAPDOCS
- BW360 "SAP BI Performance & Administration" SAPDOCS Master Thesis "Evaluation of Data Mining Methods to support data warehouse administration and monitoring in SAP business warehouse" Narasimha Raju Alluri (and consequently, all his bibliography added)
- A presentation on data mining with SAP
- BW 3.5. SAPNET Lesley 2004
- L. García Merino "Digital Marketing" 2018 – editorial saxo yo publico PERU. ISBN 9788771431735

INTERNET ADDRESSES

- http://www.google.es
- http://www.witnessminer.com
- http://www.appstate.edu/~whiteheadjc/service/logit/
- http://en.wikipedia.org/wiki/Main_Page
- http://www.cs.waikato.ac.nz/ml/weka/
- http://old.nabble.com
- http://comments.gmane.org/gmane.comp.ai.weka/20508
- http://wekadocs.com/
- http://www.opentox.org/dev/documentation/components/m5p
- www.canalvisualbasic.net/
- www.vb-mundo.com
- www.vbtutor.net/vbtutor.html

- www.lawebdelprogramador.com
- www.microsoft.com
- www.wordreference.com