Determining the Profiles of Young People from Buenos Aires with a Tendency to Pursue Computer Science Studies

María Emilia Charnelli[1], Laura Lanzarini[2], Guillermo Baldino[3], Javier Díaz[1]

[1]LINTI - Laboratorio de Investigación en Nuevas Tecnologías Informáticas
[2]III LIDI - Instituto de Investigación en Informática LIDI
Facultad de Informática, Universidad Nacional de La Plata

[3]LINSI - Laboratorio de Innovaciones en Sistemas de Información
Dpto de Sistemas - UTN Facultad Regional La Plata

mcharnelli@linti.unlp.edu.ar, laural@lidi.info.unlp.edu.ar,
gbaldino@linsi.edu.ar, jdiaz@unlp.edu.ar

Abstract. Data Mining is a discipline related to the development and application of techniques for the extraction of new and useful information from large amounts of available data. The goal of this paper is to use these techniques on data extracted from a poll in order to identify the profiles of the young people showing interest in pursuing undergraduate studies in the field of computer science. The paper describes the process of identification of the most salient features for high school students in a wide age range. It also includes the data preprocessing stage, fundamental in the process, as it strongly influences the development of the model obtained. Finally, results and conclusions are presented, as well as future lines of work.

Keywords: Data Mining, Feature Selection, Classification and Prediction Techniques, Computer Science.

1 Introduction

Nowadays, owing to the advancement of technology, most processes have digital historical information large enough to make manual processing difficult.

Data Mining, one of the most important stages of the Knowledge Discovery in Databases process, gathers a set of techniques capable of modeling and summarizing these historical data, facilitating their understanding and aiding in future decision making processes.

This paper sets out to identify and select the most relevant features for establishing and defining the profile of young people showing interest in pursuing studies in the field of computer science. This study will allow every Academic Unit to recognize the aspects that attract the attention of their prospective students, which can be used to target the right audience, thus solving the two problems implied by wrong choice of studies: the frustration of discovering the course of studies does not fulfill expectations and the financial losses caused by wrongly allocating resources incurred in by students, institutions and the state.

2 Knowledge Discovery in Databases

This paper is framed in what is known as Knowledge Discovery in Databases or KDD, which consists of a series of phases that define the methodology used. The order of these phases is not strict and there may be variation, depending on the result of each phase, which may result in a cyclic process.

- **Phase 1. Domain understanding**: Before starting to work, it is necessary to understand the situation, determine the goals and design a work plan that will help us solve the problem.
- **Phase 2. Data collection and integration**: This phase starts with the obtention of the data, which are later studied and their origins identified. This stage features data collection, description, exploration and quality verification.
- **Phase 3. Data preparation**: It is necessary to select and prepare the subset of data to be used. This phase covers all the activities to build the final data set that will be used by the modeling techniques.
- **Phase 4. Modeling**: Also called Data Mining, because it is the most characteristic of KDD, it is the phase in which multiple modeling techniques are selected and applied, configuring their parameters for result obtention. Here is where new knowledge is produced, building models from the collected data.
- **Phase 5. Interpretation and evaluation**: The models obtained in the previous phase are interpreted and evaluated in order to check whether they fulfill the goals set in preliminary phases. Here, it is critical to determine whether important parts of reality have been considered sufficiently and to decide on the reuse of the DM process results
- **Phase 6. Result dissemination, use and measurement**: Knowledge acquired through model creation must be organized and presented in such a way as to be understood and used by the end user. The applicability of the model depends on this phase.

3 Description of the Problem

The information to be used comes from a survey done by the Sadosky Foundation in multiple high schools of the Province of Buenos Aires. The questions seek to collect the impressions students have in relation to computer science and to reveal why there are few female participants in software production processes; more specifically in the Software and IT sector.

The information gathered involves 627 young people between 13 and 22 years of age. The answers obtained have generated 236 different attributes or variables. Therefore, although the number of surveys was small, the amount of attributes composing each of them makes it difficult to identify patterns or relations existing in the opinions of different subjects. In relation to this, it is a good option to use objective techniques that allow for the identification of the most relevant attributes.

However, before applying techniques specific to Data Mining, it is necessary to verify information in order to avoid discrepancies. Modifications and transformations operated on the original data are described following:

- **Attributes with inconsistent data**
  Attributes with an excessive amount of missing data were eliminated. Anomalous values resulting from loading errors and constant values such as language (all were done in Spanish) were cleaned. Redundant attributes were eliminated because they were the same value as another one or because they had the same value in all subjects.
- **Attributes with non generalizable data**
  Non generalizable attributes such as student names were eliminated. The cardinality of some attributes was reduced by using more generic categories. For example, school names were replaced by the corresponding geographic area: zona Norte, Sur, Oeste y Matanza.
- **Transformations**
  Some attributes were numerized and their range normalized, according to the requirements of the Data Mining techniques used. This type of transformations was applied to ordinal attributes such as "hability to perform different activities" or "academic level of the head of the family". In the case of questions with tabulated answers, a binary representation was used, composed by as many attributes as possible answers to the question. For example, this transformation was applied to questions such as "Use you give to the PC", "Activity you would like to work in", "What do you do in your free time?" and "How do you learn to program?".
- **Data Mining applied to open questions**
  Open questions in the survey required special treatment as they cannot be processed directly using any Data Mining technique. Some examples of these questions were "What is the first word that comes to mind to define a computer?" or "What is the first word that comes to mind to define a computer program?", where answers included lists of very different words.

  When it comes to operating with textual information, it is necessary to use Text Mining techniques in order to identify the terms most frequently used by the subjects. This was done using a process composed of multiple stages. One stage comprised the application of a stopword filter, which filters the words that match any stopword included in a file provided for this purpose. Then each word in the text was reduced to its root using a stemming algorithm [1]. The importance of this process lies in that it eliminates the syntactic variations related to gender, number and tense. Once the root of each word was obtained, the frequency of incidence was calculated and the three most frequent words were chosen.

This resulted in the three most representative terms for each of the open questions, which summarized the answers of each subject as a sequence of words in a specific and representative category.

4 Feature Selection

Data Mining techniques applied to structured information composed by a large amount of features results in complex models. Depending on the technique used, data with a high dimension produce either very large trees or sets of rules with high cardinality and backgrounds formed by a great amount of conjunctions [2] or discriminating functions that are difficult to interpret.

In order to solve this problem, it is necessary to identify the most representative attributes of the information available before the model is constructed. Thus, the technique used will be simplified in its task and result in a simpler, easier to interpret model [3].

In the particular case of the problem in this article, selecting the features is fundamental, as the goal is to identify the most important matters to young people with a tendency to pursue computer science related studies. The answer sought is, no doubt, the result of a feature selection process [4].

The two main feature selection techniques are: filter methods and wrapper methods [5]. Filter methods are based on general features of the training set to select some without using any learning algorithm. Wrapper methods require a predefined learning algorithm to select, and use its performance to evaluate and determine which features will be selected [6].

Due to the high dimensionality of the data set provided by the Sadosky Foundation, a filtering method was used to select the features, called Chi2. This method, proposed by Liu et al. in [7], is one of the most used feature selection methods and is based on a statistical method for comparing proportions. The χ^2 metric is used to measure attribute performance, as it determines a value proportional to the relation existing between a class c and a feature f which can take r possible values.

Given a set of data D with n examples, the χ^2 metric is calculated using the following formula:

$$\chi^2(D,c,f) = \sum_{i=1}^{r} \frac{(n_{i_{pos}} - \mu_{i_{pos}})^2}{\mu_{i_{pos}}} + \frac{(n_{i_{neg}} - \mu_{i_{neg}})^2}{\mu_{i_{neg}}}$$

where $n_{i_{pos}}$ and $n_{i_{neg}}$ represent the amount of positive and negative examples for value i of feature f, respectively, and $\mu_{i_{pos}}$ and $\mu_{i_{neg}}$ are the expected values if the data had a uniform distribution.

The score obtained when evaluating $\chi^2(D,c,f)$ follows the distribution χ^2 and the goal of the selection algorithm is to simply choose a subset of features among those with the highest scores, as they will be the most relevant when discriminating the classes.

The following criteria were used in order to determine how many attributes were to be selected:

- **Criterion 1**: Selecting the attributes whose score was higher than the value of the mean plus one and a half standard deviations.
- **Criterion 2**: Selecting attributes evaluating the performance of a classifier as the features are incorporated, one at a time, in order of decreasing score. A classifier is built for this purpose from the first feature and its performance is measured. This process is repeated for the remaining two characteristics with the highest score. The process continues this way, incorporating features one by one and evaluating the performance reached until no changes occur for a certain number of iterations [8].
  For the classifier performance criterion, the amount of correct answers as to who chooses computer science was the focus, i.e., the performance on true positives. This is due to the marked imbalance among the classes, as only 10% of the subjects showed interest in computer science; therefore, the general accuracy of the classifier, considering true predictions in both classes, may be high even if it does not perform well on the class of interest.

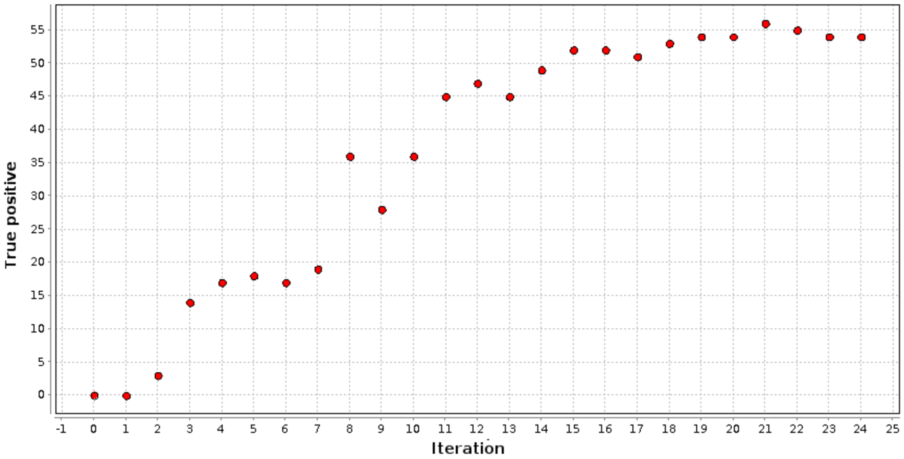

Fig. 1. Amount of true predictions in the True class of each iteration of Criterion 2

5 Results

Table 1 shows the list of attributes selected by each of the criteria described in the previous section. Regarding Criterion 2, figure 1 shows a slow improvement in the classifier until it reaches a maximum with 22 features. The classifier used in Criterion 2 was a tree generated with the popular method C4.5 [9].

The aforementioned is summarized in the two columns of table 1, where Criterion 2 can be seen to include more attributes than Criterion 1. The difference between the two criteria is less than 3% of the original amount of attributes. The 22 attributes selected constitute 9.3% of the original attributes.

It is worth mentioning that the fact that an attribute appears in table 1 does not mean that the answer to the corresponding question was positive, but rather that the answer given by the subject helps determine whether they will choose to pursue computer science related studies.

Table 2 shows the confusion matrix of the tree obtained with the C4.5 method using the 22 features in table 1. Observing the structure, it is proved that the tree correctly predicts the 80% of the cases in which subjects have answered that they would choose to study computer science. This result shows that feature selection has been successful, since the tree built from the 236 attributes shows an equivalent true prediction rate.

Finally, with the goal of identifying relevant attributes on different groups of subjects, the sample set was divided using two different criteria: by gender and by whether the subjects were in their junior years or in their senior years. Each subset was applied the Chi2 feature selection method with criterion 2. Results are represented in table 3 where only features with at least 3 matches among the different data sets were included.

For example, the first row of table 3 shows women in their senior high school year do not show a strong tendency when answering the question related to the gender of a domestic worker.

Moreover, the attributes selected show the relevance of computer use patterns in discriminating the classes. Using the computer outside of social networks, as well as for games (soccer games or other, more complex ones) seems to be an indicator of a preference for computer science related studies. Also salient are an aptitude for assembling and disassembling things and installing and configuring programs. Some less evident yet relevant attributes for the classification are an aptitude to show and receive affection.

6 Conclusions and future lines of work

This paper identifies the 22 most relevant attributes of a survey by the Sadosky Foundation that help determine whether a person will choose to pursue computer science related studies. This reduces the total of questions by 90% as the original number of attributes was 236.

The preliminary results evidence the importance of showing young people the multiple potential functions of a computer, i.e., high schools should encourage the creation of spaces and activities that take students closer to computers in unconventional ways in order to broaden their range of application, e.g, through music, image processing, robotics, etc.

Table 1. Characteristics selected by criteria 1 and 2 as relevant to determine whether a person will choose computer science related studies using the full data set.

Attribute	**Criterion 1**	**Criterion 2**
01.Most adequate gender for a domestic worker	X	X
02.Aptitude for showing affection	X	X
03.Aptitude for assembling and disassembling	X	X
04.Would you like to do informational work (not software)	X	X
05.Using the PC for social networks	X	X
06.Using the PC for soccer games and other complex games	X	X
07. Using the PC to configure, investigate or update programs	X	X
08.Use of the PC to play online games	X	X
09.¿Do you know what a computer program is?	X	X
10.¿Is programming about creating or inventing?	X	X
11.Computer Science courses by gender	X	X
12.¿Does a new car need a program to operate?	X	X
13.¿Are computer scientists like me?	X	X
14.Would you like to work assisting people	X	X
15.¿Do you like assembling and disassembling things?	X	X
16.¿Do you think computer scientists make money?	X	X
17.¿Do you like to google?		X
18.¿Do you like to show affection?		X
19.Would you like to work in the arts or show business		X
20.Would you like to work in a professional activity		X
21.¿Is the salary of a scientist high?		X
22.¿Does a lamp need a program to operate?		X

Table 2. Confusion matrix of the tree obtained from the C4.5 method using the 22 features in the table 1

	Chooses computer science	Does not choose computer science	Class precision
Predicts Chooses computer science	56	6	90.32%
Predicts Does not choose computer science	14	552	97.53%
class recall	80.00%	98.92%	

Table 3. Common features selected by at least three data subsets

Attribute Selected	Full	Only women	Only men	First age	Last age
Domestic worker gender	X		X	X	
Aptitude for showing affection	X		X	X	X
Aptitude for assembling and disassembling	X		X	X	X
Activity you would like to work in: Informational work (not software)	X		X	X	X
Free time to use the computer not for social networks	X	X	X	X	X
Using the computer for soccer games and other complex games	X	X		X	X
Using the computer to configure, investigate or update programs	X	X	X		X
¿Is programming about creating or inventing?	X		X		X
¿Does a new car need a program to operate?	X		X	X	
¿Like me?	X	X		X	
¿Does a lamp need a program to operate?	X	X		X	

Using games and configuring and managing software applications seem to be strong indicators of a tendency in students to choose computer science careers. In this sense, proposing workshops on these topics could increase an interest in the field.

This paper represents a first step in defining the features of computer science students. The obtained results will allow for a definition of the direction of future surveys that are similar to the one performed by the Sadosky Foundation. As a future line of work, other techniques will be used to generate models from the set of attributes selected and similar surveys will take place within the National University of La Plata.

References

1. Gupta, V., Lehal, G.S.: A survey of common stemming techniques and existing stemmers for indian languages. Journal of Emerging Technologies in Web Intelligence **5** (2013) 157–161
2. Sebban, M., Nock, R., Chauchat, J.H., Rakotomalala, R.: Impact of learning set quality and size on decision tree performances. Int. Journal of Computers, Systems and Signals **1** (2000) 85–105
3. Thrun, S.B., Bala, J., Bloedorn, E., Bratko, I., Cestnik, B., Cheng, J., Jong, K.D., Dzeroski, S., Fahlman, S.E., Fisher, D., Hamann, R., Kaufman, K., Keller, S., Kononenko, I., Kreuziger, J., Michalski, R., Mitchell, T., Pachowicz, P., Reich, Y., Vafaie, H., Welde, W.V.D., Wenzel, W., Wnek, J., Zhang, J.: The monk's problems a performance comparison of different learning algorithms. Technical report (1991)
4. Alelyani, S., Tang, J., Liu, H.: Feature selection for clustering: A review. (2013)

5. Das, S.: Filters, wrappers and a boosting-based hybrid for feature selection. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML 2001, Williamstown, MA, USA, Williams College (2001)
6. Song, Q., Ni, J., Wang, G.: A fast clustering-based feature subset selection algorithm for high-dimensional data. Knowledge and Data Engineering, IEEE Transactions on **25** (2013) 1–14
7. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: Proceedings of the Seventh International Conference on Tools with Artificial Intelligence. TAI '95, Washington, DC, USA, IEEE Computer Society (1995) 88–
8. O'Mahony, M.P., Cunningham, P., Smyth, B.: An assessment of machine learning techniques for review recommendation. In: Artificial Intelligence and Cognitive Science. Springer (2010) 241–250
9. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)