CRISP-DM Model Applied for Knowledge Discovery in Speech Disorders Therapy Area

Mirela DANUBIANU, Ştefan Gheorghe PENTIUC, Iolanda TOBOLCEA

***Abstract*—Technological development has led to the opportunity to accumulate large volumes of data. Extracting new knowledge from these data is a task which requires the use of data mining techniques. During evolution to maturity, were developed several models of the process of knowledge discovery in databases. The CRISP-DM model provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks and relationships between these tasks. It is a standard process model, non-proprietary and freely available. As shown in various research data mining technologies can be successfully applied in the field of speech therapy. Consequently is a natural attempt to use the CRISP-DM phases, to model the process of data mining. The aim of this paper is to point briefly, the contents of these phases related to data collected in the databases of computer based speech therapy systems.**

***Index Terms*—data mining, CRISP-DM model, speech disorders, personalized therapy**

I. INTRODUCTION

THE last years, the children with speech disorder have more and more become object of specialists' attention and investment in speech disorder therapy are increasing. The development and use of information technology in order to assist and follow speech disorder therapy allowed researchers to collect a considerable volume of data. This was possible due to the development of database technology and due to the development of media storage, which have the capacity to store an impressive amount of data. Increased volume of data available did not lead immediately to a similar volume of information to support the decisions of effective therapy, because the classical methods of data processing are not applicable in such cases. Answers to questions such as: how is the estimated duration of therapy for a particular case, what is the predicted final state for a child or what will be its state at the end of various stages of therapy or what are the best exercises for each case and how can dose their effort for effectively solve these exercises may be the subject of predictions obtained by applying data mining techniques on collected data.

M DANUBIANU is with the "Stefan cel Mare" University of Suceava (phone: +40744.547164; e-mail: mdanub@ eed.usv.ro).

St. Gh. PENTIUC is with the "Stefan cel Mare" University of Suceava (e-mail: pentiuc@ eed.usv.ro).

I. TOBOLCEA is with the "A.I.Cuza" University of Iasi (e-mail: itobolcea@ yahoo.com)

II. DATA MINING, KNOWLEDGE DISCOVERY IN DATABASES AND CRISP DM MODEL

A fortunate confluence of a variety of factors such as: the explosive growth in data collection, the storing of the data in data warehouses, the availability of increased access to data from Web navigation and intranets, the tremendous growth in computing power and storage capacity has lead to remarkable growth in the field of knowledge discovery. Clearly, a lot of data is collected, but what can we learn from all these data? John Naisbitt has observed that "we are drowning in information but starved for Knowledge" [1]. This was the reason for the appearance of data mining.

According to the Gartner Group [2], "data mining (DM) is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."

There are other voices, such as Fayyad which considers data mining as one of the phases of the KDD process [3]. The DM phase concerns, mainly, to the means by which the patterns are extracted and enumerated from data.

KDD process is defined as the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. There are considered five stages, presented in figure 1:

- Selection - this stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed;
- Pre-processing - consists on the target data cleaning and pre processing in order to obtain consistent data;

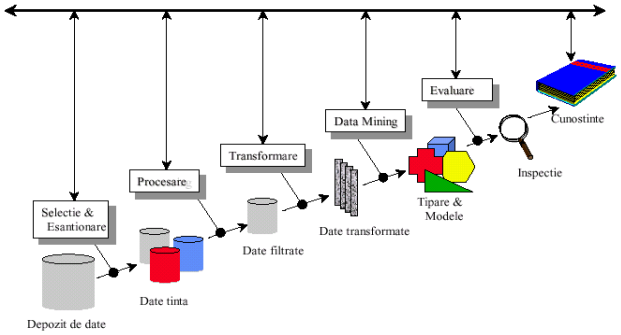

Fig. 1. KDD process.

- Transformation - consists on the transformation of the data using dimensionality reduction or transformation methods;
- Data Mining - consists on the searching for patterns of interest in a particular representational form, depending on the DM objective (usually, prediction);
- Interpretation/Evaluation - this stage consists on the interpretation and evaluation of the mined patterns.

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user [4]. The KDD process is preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It must be continued by the knowledge consolidation, incorporating this knowledge into the system [3].

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. According to CRISP–DM, a given data mining project has a life cycle consisting of six phases, as illustrated in Figure 2. The phase sequence is adaptive. That means the next phase in the sequence often depends on the outcomes associated with the preceding phase.

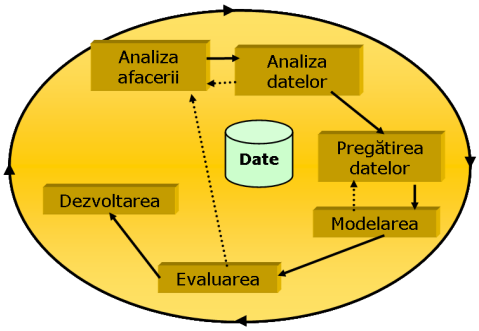

Fig. 2. CRISP-DM Model for Data Mining.

- Business understanding-this initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives;
- Data understanding- starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information;
- Data preparation- is the phase which covers all activities to construct the final dataset from the initial raw data;
- Modeling-in this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values;
- Evaluation-at this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives;
- Deployment-creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it [5].

CRISP-DM is extremely complete and documented. All his stages are duly organized, structured and defined, allowing that a project could be easily understood or revised.

III. Data Mining in Speech Disorder Therapy Area

The development and use of information technology in order to assist and follow speech disorder therapy allowed researchers to collect a considerable volume of data. Increased volume of data available did not lead immediately to a similar volume of information to support the decisions of effective therapy, because the classical methods of data processing are not applicable in such cases.

The logaoedic intervention is a complex process that consists of the following actions:

- Finding the persons with speech impairments, introducing their data into a database and proposing a first diagnosis. Having finished the discovery process, a report including the synthetically and complete results about these persons will be generated.
- selecting the persons that will follow the speech therapy
- Programming the therapy sessions and looking after the therapy progress. To asses the way in which the therapy evolves means having the give-away of person attendance at therapy sessions and the individually activity report. This activity report gives out information regarding the exercises that have been done, how many times each exercise has been repeated, the time needed for each exercise and the results.
- the periodical evaluation of the persons, during therapy and the establishment of final results.

In the area of speech therapy we can use the following data mining tasks:

- classification which aims to put the persons with speech disorders in some predefined segments. This method allows estimating the dimensions and the structures of the different groups. Classification uses the information contained in the set of predictor variables (e.g. those relating to the personal or family anamnesis, or to the lifestyle), for relating the persons with the different segments.
- clustering which is able to forms groups of persons with speech disorders based on the similarity of some characteristics, but it does not use the apriori defined groups. It is an important task since it helps the speech therapists to understand who their patients are. For instance by clustering they can discover a subgroup of a predetermined segment who has an homogeneous behavior related to different therapy methods who can be efficient targeted by a specific therapy.
- association rules which find the connections between data. An important task of this method should be to determine why a certain therapy program has had success for a segment of patients while for another segment it was inefficient.

IV. USING CRISP-DM MODEL FOR SPEECH DISORDER THERAPY AREA

In present, because the needs of efficient use of time or due to the economic needs, have become of interest information such as:

- how is the estimated duration of therapy for a particular case,
- what is the predicted final state for a child or what will be its state at the end of various stages of therapy,
- what are the best exercises for each case and how can dose their effort for effectively solve these exercises,
- how is associated the family receptivity - which is an important factor in success of the therapy - with other aspects of family and personal anamnesis.

All this may be the subject of predictions obtained by applying data mining techniques on data collected by using various Computer Based Speech Therapy systems [6].

Adapting the therapy programs involves a complex examination of children and recording of relevant data relating to personal and family anamnesis.

Complex examination of how the children articulate the phonemes in various constructions allows a diagnosis and classification in a given category of severity.

Anamnesis data collected may provide information relative to various causes that may negatively influence the normal development of the language. It contains historical data and data provided by the cognitive and personality examination.

On provide to the applied personalized therapy programs data such as number of sessions/week, exercises for each phase of therapy and the changes of the original program according to the patient evolution. In addition, the report downloaded from the mobile device collects data on the efforts of child self-employment. These data refer to the exercises done, the number of repetitious for each of these exercises and the results obtained.

The tracking of child's progress materializes data which indicate the moment of assessing the child and his status at that time. Often all these data are stored in relational databases.

Data collected by the information system, together with data from other sources (eg demographic data, medical or psychological research) is the set of raw data that will be the subject of data mining.

A first remark is related to the fact that we have a relational database, characterized by a high degree of normalization, making the various characteristics to be in different tables.

Creating target data set is accomplished through a join of tables containing useful features followed by a projection on a superset of appropriate attributes, as is shown in (1):

$$\prod_{I_i}\left(T_1 \triangleright\triangleleft T_2 \triangleright\triangleleft \ldots \triangleright\triangleleft T_k\right) \quad (1)$$

where: *Ii* is a superset of the attributes regarding the useful characteristics

T1 …Tk is the set of tables containing the attributes in the list of projection.

Another remark concerning the data contained in the database relates to the encoding characteristics by the values specified in the various fields. These codes were converted so that the data set on which apply data mining algorithms contain descriptive understandable values.

It is preferable to change the numerical values during the data preprocessing step, so that the final data set contains the descriptive values of characteristics.

As most data mining techniques were not designed to cope with large amounts of irrelevant features, combining them with feature selection techniques has become a necessity in many applications. The most important objectives of feature selection are: to avoid over fitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering; to provide faster and more cost-effective models and to gain a deeper insight into the underlying processes that generated the data. In the feature selection problem, we are given a fixed set of candidate features for use in a learning problem, and must select a subset that will be used to train a model that is "as good as possible" according to some criterion.

The next step consists in applying different data mining algorithms on the data previously obtained. Starting from the needs of information and from the analysis of available data in the database, we can make the following observations.

According with the nature of the predictor data is suitable to use, for classification, decision trees. To choose the best solution were tested implementations of algorithms CART, ID3/C4.5 and RAIN FOREST. An example of classification for the patients, according to their possible status at the end of therapy, is: corrected (C), improved (I) or stationary (S).

In order to mine the association rules the system uses an Apriori algorithm implementation, applied on the data processed so as to obtain a transactional structure.

V. CONCLUSION

The data mining technology may be a useful tool for the logopaedic therapy because it may conduct to reduction of duration of therapy, to increasing the possibilities of achieving superior results and finally to lower cost of the therapy. To avoid accidental procedural approach to achieve a DM project, we considered the CRISP-DM model meet the requirements and characteristics of speech therapy field. So we made a brief presentation of the sequence of activities related to discovering interesting patterns by DM techniques grouped by phases of this model.

REFERENCES

[1] J. Naisbitt, *Megatrends*, 6th ed., Warner Books, New York, 1986

[2] The Gartner Group, www.gartner.com

[3] Fayyad, U. M. et al. 1996. From data mining to knowledge discovery: an overview. In Fayyad, U. M.et al (Eds.), Advances in knowledge discovery and data mining. AAAI Press / The MIT Press

[4] Brachman, R. J. & Anand, T., 1996. The process of knowledge discovery in databases. In Fayyad, U. M. et al. (Eds.), Advances in knowledge discovery and data mining. AAAI Press / The MIT Press.

[5] Chapman, P. et al, 2000. CRISP-DM 1.0 - Step-by-step data mining guide.

[6] M. Danubianu, S.G. Pentiuc, T. Socaciu (2009), Towards the Optimized Personalized Therapy of Speech Disorders by Data Mining Techniques, Proceedings of ICCGI 2009, Vol: CD, 23-29 Aug., 2009, Cannes, France, ISSB/ISBN: 978-0-7695-3751-1