20

Data Mining in Personalized Speech Disorder Therapy Optimisation

Danubianu Mirela, Tobolcea Iolanda and Stefan Gheorghe Pentiuc
"Stefan cel Mare" University of Suceava,
"A.I. Cuza" University of Iasi,
"Stefan cel Mare" University of Suceava,
Romania

1. Introduction

In the context of the Sustainable Development Strategy adopted by the European Council in 2006, one of the key challenges is related to public health, whose general objective envisages a good level of public health. In order to accomplish that, one of the specific targets includes better treatments of diseases. It is true that there are affections which by their nature do not endanger the life of a person however they may have a negative impact on the quality of life. Various language or speech disorders are part of this category, because language is the most common mean of interpersonal communication. It enables the transfer content from one person to another and also performs an important cognitive function of integration, design and conceptualization of thinking. Language enables an individual to live in community, it offers them the opportunity to demonstrate their qualities and to adapt to different situations. This kind of impairments affects a considerable percentage of people, most of them being children, but if they are discovered and treated in time, they can be often corrected.

The logopaedic intervention proposes some specific objectives such as detection, complex assessment and identification of language and communication disorders of pre-school and young school children and targeting logopaedic therapy to correct, recovery, compensation, adaptation and social integration. This last goal involves the application of a personalized therapy to each child or group of children with similar characteristics, therapy adjusted according to their disease severity and directed towards eliminating the causes that has generated the speech impairment.

Information technology is used by specialists in order to assist and supervise speech disorder therapy. Consequently they have collected a considerable volume of data about the personal or familial anamnesis, regarding various disorders or regarding the process of personalized therapies. These data can be used in data mining processes that aim to discover interesting patterns which can help the design and adaptation of different therapies in order to obtain the best results in conditions of maximum efficiency.

Data mining involves the application of analysis on large volumes of data using algorithms which, at acceptable efficiency of calculation, produce a particular enumeration of patterns from such data. As an exploration and analysis technique of large amounts of data in order to detect patterns or rules with a specific meaning, from apparently unrelated data, data

mining may help discover relationships that can anticipate future problems or might solve the studied problems.
According to the logopaedic activity the tasks performed by data mining can be grouped into the following categories: classification, clustering and association rules.
The aim of this chapter is to present some aspects regarding the possibility of applying data mining techniques in order to optimize the personalized therapy of speech disorders, in particular the therapy of dyslalia, and to present a data mining system designed for this purpose.

2. What are speech disorders?

2.1 Speech disorders and their implication in the individual's social life

Language is of outmost importance to each individual's mind and personality structuring, as it is a means of human communication, education and child development, of the understanding and creating of specific human relationships.
Most children have no speech difficulties; they prove fluency, expressivity and the communication is pleasant and attractive. Others, on the contrary, have difficulties when they wish to express their thoughts in words. Although they make great efforts, their speech is incorrect, altered, in more serious cases it becomes stammered or dyslalic, this creating an uncomfortable feeling, an overwhelming complex of inferiority. As a result, these children back away, "close the doors" to the ones around them, they isolate themselves from the group, as their deficient language doesn't allow them to establish good interpersonal relationships needed for good communication purposes.
So, speech disorders may be defined as a problem with fluency, voice, and/or how a person utters speech sounds. These may have different causes, from organic to functional, neurological or psycho-social causes.
Dyslalia is articulation disorder that consists in difficulties with the way sounds are formed and strung together. These are usually characterized by omitting, distorting a sound or substituting one sound for another. Dyslalia has the greatest frequency among handicaps of language for psychological normal subjects as well as for those with deficiencies of intellect and sensory.
The prevention and treatment of speech disorders is a complex issue, stirring the interest of speech therapists, as well as to those asked to contribute to the children's language education. Early treatment of speech impairments ensures improved efficiency, as psycholinguistic automatisms are not consolidated in young children, and through adequate educational interventions, they can be replaced by correct speech acts. Treating speech disorders prevents school or society failure for the child. This process also involves child therapy in his/her ordinary life style, involving family, parents and school.
Differential diagnosis will decide upon the therapy for correcting language as psycho diagnosis allows an adequate therapeutic program and the elaboration of a prognosis regarding the evolution of the child, along with the therapeutic process. The therapy has to be adapted to each language therapist, to each particular case, to the child's learning rhythm and style, as well as to the level of the impairment. Due to the complexity of the possible problems involved, the therapeutic methods will vary significantly, from analysis, synthesis to global therapeutic specific methods and techniques. This is why, a good knowledge of all these methods will allow therapists to pick the best ones for a specific case. Therapy stages are to be followed in a specific order, regardless of the methods, according to each child's

psychosomatic structure. The therapy starts with the involved psychological processes (cognitive, psychometric and affective-emotional) and it is build on stages, moments and concrete objectives, materialized in specific therapeutic techniques. The techniques materialize in exercises and procedures which the child has to perform in order to achieve the final aim: a correct speech act.

2.2 Information and communication technology in speech disorders therapy practices

Information and Communication Technology (ICT) can help persons whose physical conditions make communication difficult and can be used in speech therapy as a real clinical tool. The technical means can be used in combination with the speech therapist's clinical competences in order to help their patients, orienting the evaluation, the treatment and the constant feedback in a more flexible and modern way.

Computers can help in diagnosing the speech disorders, can produce audio-visual feedback during the treatment and monitor and evaluate the therapeutic progress. Also, they provide some sets of practical exercises for the patients who are not under the direct supervision of a speech therapist.

The use of Computer Based Speech Therapy (CBST) systems for speech therapy creates a new psychological and pedagogic situation, a special learning environment, by facilitating a superior method for information recovery. The children's interest in the therapeutic activity is cultivated by the enrichment of the material resources as a support for an efficient and rapid learning by using new tools of information and communication technology. The computer supports the children's curiosity for knowledge acquisition by giving new and rich information that can be provided in their natural dynamics. This fact, also increases the children's motivation for learning.

Consequently, there are some advantages of using the CBST , such as: the possibility to detect aspects impossible or difficult to realize by other means and to separate or recompose phenomena that are imperceptible by other means; the capacity to accurately playback the content and allow immediate playback of the information or as often as necessary and last but not least the appeal it has for children due to the original aspects involved, as well as to the aesthetic way of presenting the information.

To conclude, there are more arguments supporting the use of ICT for increasing the efficiency of speech therapy. Firstly, there is the possibility to record the verbal material, in order to provide that "language immersion" necessary for acquiring correct pronunciation. These records constitute tools of self-control of the errors made and the progress achieved, enabling the child to hear and to judge himself/herself from the outside.

For the speech therapist, a CBST constitutes a tool for controlling and evaluating the efficiency of the proposed strategy, helping to adapt the therapeutic schema. It is also an important aid for the therapist in the analysis of the speech therapy steps, the responses received from the patients, managing to appreciate the efficiency of the methods, the means used in reaching the established goal.

A CBST can develop the logical thinking of the children and their affectivity, it can create a pleasant, relaxed, and attractive climate, increasing the efficiency of the therapy.

Finally, a CBST constitutes the most complex method that comprises the audio-visual techniques. Its great advantage consists in realizing some educational and instructive software programs; it helps and increases the efficiency of the didactic activity.

Recently, a therapeutic software useful for correcting various speech disorders has been elaborated. Some programs are simple and produce a single type of visual and auditory

feedback, while others are extremely complex, allowing a sustained training, realized for several aspects of speech.`

There are some international projects whose priorities are represented by developing information systems that will allow the elaboration of personalized therapeutically paths.

The OLP (Ortho-Logo-Paedia) project (OLP 2002) for speech therapy started in 2002; the project is financed by the EU and it is a complex project, involving the Institute for Language and Speech Processing in Athens and seven other partners from the academia and medical domains. It aims to accomplish a three – module system (OPTACIA, GRIFOS and TELEMACHOS) capable of interactively instructing the children suffering from dysarthria (difficulty in articulating words due to disease of the central nervous system). The proposed interactive environment is a visual one and is adapted to the subjects' age (games, animations). The audio and video interface with the human subject will be the OPTACIA module, the GRIFOS module will make pronunciation recognition and the computer aided instruction will be integrated in the third module – TELEMACHOS.

An interesting project is STAR – Speech Training, Assessment, and Remediation (STAR 2002), started in 2002, a project which is still in the development phase. The members (AI. duPont Hospital for Children and The University of Delaware) aim to build a system that would initially recognize phonemes and then sentences. This research group offers a voice generation system (ModelTalker) and other open source applications for audio processing.

Speechviewer III developped by IBM (Speechviewer III) creates an interactive visual model of speech while users practice several speech aspects (e.g. the sound voice or special aspects from current speech).

The ICATIANI device developed by TLATOA Speech Processing Group, CENTIA Universidad de las Américas, Puebla Cholula, Pue. México uses sounds and graphics in order to ensure the practice of Spanish Mexican pronunciation. Each lesson explains sounds pronunciation using the facial expression with a particular accent on specifying articulation points and the position of the lips. The system includes several animated faces, each of them showing the correct method of vocal pronunciation and providing feedback to the child answers. In this case, if the child's pronunciation matches the system one, the child is rewarded by a smile or otherwise warned by a sad face.

The information systems with real time feedback that address pathological speech impairments are relatively recent due firstly to the amount of processing power they require. The progress in computer science allows at the moment for the development of such a system with low risk factors. Children pronunciation is also used to enrich the existing audio database and to improve the current diagnosis system's performances.

The personalized therapy system of dyslalia for Romanian language – TERAPERS was developed within the Center for Computer Research in the University "Stefan cel Mare" of Suceava (Danubianu & al, 2009, a). This project aims to develop a system which is able to assist teachers in their speech therapy of dyslalia and to follow how the patients respond to various personalized therapy programs.

It has reached some specific objectives (Danubianu & al 2009, a) such as: initial and ongoing therapy evaluation of children and identification of a modality for standardizing their progresses and regresses (at the level of the physiological and behavioral parameters); the development of an expert system for the personalized therapy of speech impairments that allows designing a training path for pronunciation, individualized according to the speech disorder category, the previous experience and evolution of the child's therapy; the

development of a therapeutically guide that allows mixing classical methods with the adjuvant procedures of the audio-visual system and the design and the achievement of a database that contains the child's data, the set of exercises and the results obtained by the child.

The system contains two main components as is presented in Figure 1: an intelligent system installed on each office computer of the speech therapists and a mobile system used as a friend for the child therapy. The two systems are connected (Danubianu &al, 2009, a).

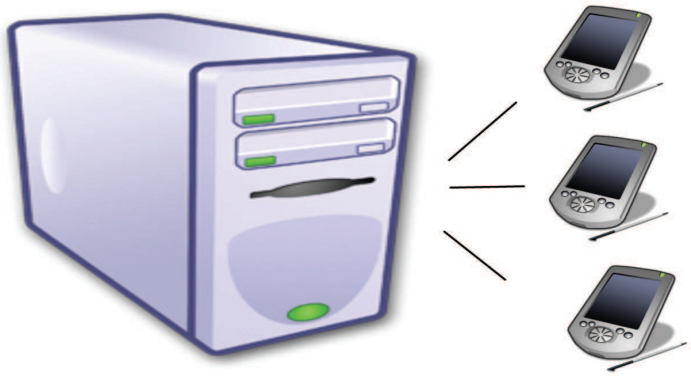

Fig. 1. TERAPERS Architecture

The intelligent system is the fix component of the system installed on each speech therapist's office computer. This system includes the following parts:

- an evaluation module of the children's progress;
- an expert system that will produce inferences based on the data presented by the evaluation module;
- a virtual module of the mouth, that would allow the presentation of every hidden move that occur in speaking,

The main stream activities of the intelligent system from TERAPERS are presented in Figure 2. All these activities are materialized in a consistent volume of data stored in a relational database (Danubianu &al, 2009, a).

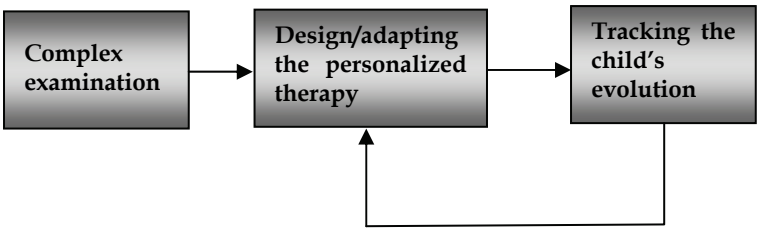

Fig. 2. The functional schema of the intelligent system of TERAPERS

Starting from a complex examination of the child, mainly related to the way he/she articulates phonemes in different construction, speech therapists can establish a diagnosis. Based on this diagnosis a personalized therapy program is designed. During the therapy the initial program can be modified and adapted to the child's current needs and evolution.

The mobile device has two main objectives. It is used by the child in order to resolve the homework prescribed by the speech therapist and delivers to the intelligent system a personalized activity report of the child.

3. Knowledge discovery in databases and data mining

3.1 Knowledge discovery in databases

Knowledge Discovery in Databases (KDD) has emerged as a consequence of the huge volumes of data, resulted from the technological progress, that we witnessed in the last years. These data collections have lead to a paradox generated by the fact that, although there was a lot of data, the information extracted from these data was poor.

The Knowledge Discovery in Databases was defined as the process of identifying valid, novel, potentially useful, and understandable patterns in data.

The process may be generalized to nondatabase sources of data, although it emphasizes databases as a primary source of data. (Cios & al. 2007)

Knowledge Discovery in Databases involves an interactive and iterative sequence of steps. An important objective, for which one spent considerable efforts was related to the establishment of a process model.

The first model for Knowledge Discovery in Databases has been developed in academia by Fayyad et al. but it was used in various domains, including engineering, medicine, e-business, or software development. It consists of nine steps, which are listed as follows: developing and understanding the application domain, creating a target data set, data cleaning and pre-processing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm, data mining, interpreting mined patterns and consolidating discovered knowledge.

The academic models were quickly followed by some industrial models. The most representative is the six-step CRISP-DM model, developed by a large consortium of European companies, which has become the leading industrial model.

The CRISP-DM model, presented in Figure 3, consists of the following six steps: business understanding, data understanding, data preparation, modeling, evaluation of the model and deployment (Chapman & al, 2000).

Business understanding is the first phase of the KDD process that focuses on understanding the project objectives and requirements from the business perspective. There are some tasks to do. First the data analyst must understand, from the business perspective what the client want to realize. Even if there are many objective and constraints, they must be properly balanced. It is also important to describe the criteria for a useful outcome of the project from the business point of view. Secondly, the analyst must assess the situation regarding the resources, constraints or other facts that should be taken into consideration to make a good project plan. After that, the data mining goal must be determined. A such data mining goal might be ”Find the optimal personalized therapy for the patient with the following characteristics….”. The result of this phase is a project plan.

Data understanding consist of initial data collection, data describing and assessing and verifying data quality. Initial collection may include data loading where necessary, into a specific tool for data understanding. This action might require some initial data preparation operation. If the data come from multiple sources the integration must be done. In conjunction with data collection this phase offer a data description that refers to the format

of the data and its volume measured in number of records and fields for each table. Finally the data quality must be examined. To achieve this it should answer some questions such as:

- Is the data complete?
- Is the data correct or does it contain errors?
- Does the data contain any missing values?

Data preparation aims to build the final data set from the initial raw data. Tasks specific to this phase contain table, record and attribute selection and transformation and cleaning data for modelling tools. In the selection step we decide which data to use for the analysis, based on criteria such as: relevance to the data mining goals, quality and limits on data volumes or data type. The cleaning data step involves the selection of clean subsets of the data or the insertion of proper values where they are missing, and its aim is to improve the data quality to the of analysts' requirements. Other data preparation operation refers to the production of derived attributes, the transformation of values for existing attributes and data integration. Sometimes it is necessary to format data as the modeling tool requires.

The next phase, when the model is build by various modeling techniques is called modeling. This phase is the core of the knowledge discovery process. The first step in this phase select the technique to be used. Then it is necessary to generate a procedure to test the model's quality. The model is created by running the modeling tool on the prepared data set. The last step of this phase is related to the assessment of the model. Now, the data mining specialist interprets the model according to the domain of knowledge.

Evaluation of the model consists of three steps: evaluation of results, reviewing of the process and determining the next steps. In this phase, evaluation of results aims to asses if the model meets the business objectives and what are the business reasons that may explain the model's efficiency. Even if the model is satisfactory it is advisable to review the whole process in order to see if there is a task that has been overlooked. Then, according to the assessment results and process review, the expert must decide whether the project is proper for deployment or it is necessary to initiate further iterations.

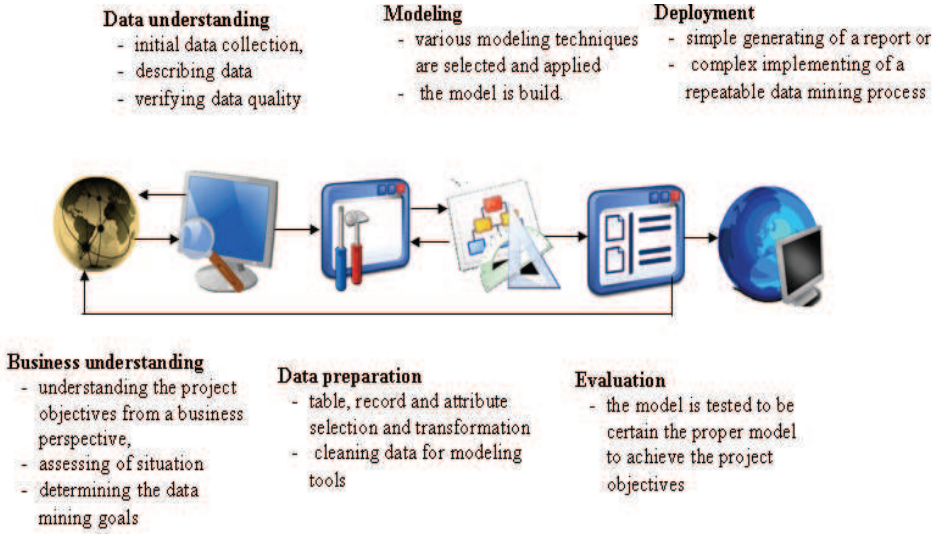

Fig. 3. CRISP-DM model for KDD process (Danubianu & al, 2010)

The last phase of the KDD process is deployment. This task considers the evaluation results in order to deploy the data mining results into the business and to establish a strategy for deployment. The first outcome of this step is a deployment plan. Then, if the model obtained from data mining phase become a part of business, it is important to design a monitoring and maintenance plan. At the end of the project, depending on the deployment plan, a final report is elaborated. This final report may be a summary of the project or a presentation of the data mining results. As part of this step may consist in assessing what was done well in the project and what needs to be improved.
Looking at the models presented above one can observe that both of them contains a step (called data mining in the Fayyad model and modeling in the CRISP-DM model) that applies different methods and algorithms in order to discover new patterns.

3.2 Data mining

Data mining involves the application of analysis on large volumes of data using algorithms which produce a particular enumeration of patterns from such data.
Data mining may facilitate the discovery from apparently unrelated data, relationships that can anticipate future problems or might solve the studied problems. So, data mining is defined as the operation of extracting the interesting and previously unknown information and represents one phase in the complex process of Knowledge Discovery in Databases.
Data mining solve problems which can be divided into two general categories: prediction and knowledge discovery (or description). Even prediction is the main goal of data mining, it is often preceeded by description. For example, in a health care application for a disease recognition, which belongs to predictive data mining, we must mine the database for a set of rules that describes the diagnosis knowledge, and this knowledge is further used for the prediction of the disease when a new patient comes in.
Each of these two problems has some associated methods. For prediction we can use classification or regression while for knowledge discovery we can use deviation detection, clustering, association rules, database segmentation or visualization.
Data classification is a supervised learning method which consists in a two-step process. First, by analyzing database tuples described by attributes a model is built. It describes a predetermined set of data classes or concepts. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. The data set analyzed to build the model form the training data set. In the second step, the model is used for classification. Before that, it is necessary to estimate the predictive accuracy. The simplest technique for this use a test set of class labelled samples, independent of the training set, and randomly selected from the whole data set. The accuracy of the model on a data test set is calculated as the percentage of test set sample that are correctly classified by the model previously build. To find this percentage, for each test sample, the known class is compared with the model's class prediction for that sample.
If the accuracy is acceptable the model can be used for classifying future data tuples for which the class label is not known. The various classification methods can be compared and evaluated according more criteria, such as: predictive accuracy, speed, robustness, scalability and interpretability. (Han & Kamber, 2000)
Whereas classification determines the set membership of the samples, the prediction of continuous values can be modeled by *regression.* In this case model design consists of finding a structure for it, on computing an optimal value for its parameters and assessing

the model quality. The model structure relates the type of mathematical formula that describes the system behavior. Depending on the model structure, regression models may be categorized as follows: simple linear regression, multiple linear regression, polynomial regression, logistic regression or nonlinear regression.
We can also distinguish between static and dynamic models. Static models produce outcomes based only on the current input, whereas dynamic models produce outcomes based on the current input and the past history of the model behavior.
Clustering, often referred to as unsupervised learning, involve a process that discovers structures in data without any supervision. As the name clustering implies, unsupervised algorithm is capable of discovering structures on its own by exploiting similarities or differences between individual data points on a data set.
There are a lot of strategies for clustering formation, and many approaches try to determine what similarities between data mean.
Clustering techniques can be divided into three main categories: partition, hierarchical clustering and model based clustering. For each of these categories, the clustering principles are different because they use different ways of processing and formatting the results.
Partition based clustering methods use objective functions whose minimisation is supposed to lead to the discovery of the structure existing in the data set. This category of methods works with a predefined number of clusters and proceeds to the optimisation of the objective function. There are some variants in which successive splits of the clusters are allowed. In this case we have a dynamically adjusted number of clusters.
The successive development of clusters is the idea of hierarchical clustering. We can start with a single cluster that is the entire data set successively divided or we can start with individual points treated as initial clusters which are merged and form new clusters. The last way of forming final clusters leads to the concept of agglomeration clustering.
In model-based clustering we assume a probabilistic model of the data and then we estimate its parameters.
Association rules mining is also an important data mining method that aims to find interesting dependencies in large sets of data items. These items are often stored in transactional databases that must have a specific format. This format can be generated by an external process or can be extracted from relational databases or data warehouses. Interesting associations between data items can often lead to information used for decision making.
The algorithms used in data mining are often well-known mathematical algorithms, but in this case they are applied to large volumes of data and to general business problems. The mostly used are: statistical algorithms, neural networks, decision trees, genetic algorithms, nearest neighbor methods, rule induction and data visualization.
Statistical algorithms have been used by analysts to detect unusual patterns and explain patterns using statistical models as linear models. Such systems cannot determine the form of dependencies hidden in data and require that the user provides his own hypotheses that will be tested by the system.
One of the main statistical concepts which can be used for data mining techniques is Bayes theorem. This can be also used for implementing of more complex Knowledge Discovery in Database techniques, such Bayesian networks.
Neural networks simulate the human brain capacity to find patterns. In our acceptation a neural network is a set of connected inputs/outputs where each connection has an

associated weight. For this reason neural network learning is also referred to as connection learning.
One of the most widespread architectures for neural network, multilayered perceptron with back propagation of errors, emulates the work of neurons incorporated in a hierarchical network. In this case the input of each neuron of the current layer is connected with the outputs of all neurons belonging to the previous layer. The data to be analyzed is treated as neuron excitation parameters and is fed to inputs of the first layer. These excitations of a layer neurons are propagated to the next layer neurons, being weakened or amplified with the weights assigned to the corresponding intraneural connection. At the end of this process a single neuron, situated at the topmost neuron layer, acquires some values considered to be a prediction.
Neural networks involve long training periods and require a number of parameters which are determined empirically. One of these parameters may be the network topology. Some of the neural network advantages include their tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained, but their critics underline their poor interpretability, since it is difficult to interpret the symbolic meaning behind the learned weights. So, a major disadvantage of neural networks consists in their Knowledge representation.
Decision trees can be applied for classification or clustering tasks. If we have a heterogeneous data collection and a set of attributes that describe the data, decision trees aim to divide the data set into smaller, more homogeneous subsets, using the values of the attributes selected.
There are several techniques for constructing or modeling the trees, referred to as decision-tree based algorithms which aim to minimize the size of the tree while maximizing the accuracy of the classification.
Genetic algorithms are based on the principles of natural evolution. They use terminology and concepts analogous to those used in biology. For example, genetic algorithms encode each point in a solution space into a string called chromosome. The features in this string are genes and their position in the string is called locus. During the execution of genetic algorithms samples evolve similar to the natural evolution process to provide optimal solutions. In data mining this kind of algorithms can be used for prediction, clustering and association rules inference.

4. Application of data mining techniques in speech disorders therapy area

The tasks performed by data mining and related to speech therapy activity can be grouped into the following categories:

- classification which aims to place the people with different speech impairments in predefined classes. It is a possibility of tracking the size and structure of various groups of patients. Classification is based on the information contained in many predictor variables, such as personal or familial anamnesis data or related to lifestyle, to join the patients with different classes.
- if there are no predefined classes we can group people with speech disorders on the basis of similarity of different features by clustering, It is an important task which helps the therapists understand who they patients are. Clustering aims to find subsets of a predetermined segment, with homogeneous behavior towards various methods of therapy that can be effectively targeted by a specific therapy.

- association rules which find out associations between different data which seems to have no semantic dependence. An important task of the association is to determine why a specific therapy program has been successful on a segment of patients with speech disorders while on the other it was ineffective.

From the above, we may conclude that data mining can be a useful tool. Nevertheless, there is a limitation. Data mining applications generate information by analyzing patterns of data obtained from the systems which assist and supervise the speech therapy. Such patterns can help to predict the evolution of the individuals that are currently in the process of therapy, or to design a scheme of appropriate therapy for them. However data mining technology can not provide information about impairments, people or behaviors that are not found in the databases that provide data for analysis.

5. Logo-DM system

5.1 System objectives

A considerable volume of data was collected by the researchers due to the development and use of information technology in order to assist speech disorder therapy. 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.

We think these data can be the foundation of data mining processes that show interesting information for the design and adaptation of different therapies in order to obtain the best results in conditions of maximum efficiency.

The idea of trying to improve the quality of logopaedic therapy by applying some data mining techniques started from the TERAPERS project developed within the Center for Computer Research of "Stefan cel Mare" University of Suceava (Danubianu & al, 2009,a). The data collected in this system together with data from other sources (e.g. demographic data, medical or psychological research) may be the set of raw data that will be the subject of data mining. This is why we have proposed the development of the Logo-DM system.

Currently, economic needs require correct answers to questions such as the following

- 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 they dose their effort to effectively solve these exercises?,
- how is family receptivity associated - which is an important factor for a successful 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 TERAPERS. It is also interesting, as part of the knowledge discovered by data mining algorithms, to use it to enrich the knowledge base of the expert system embedded in TERAPERS.

Consequently, the Logo-DM system aims to optimize the personalized therapy of dyslalia for pre-school and young school children using data mining techniques. By implementing classification, clustering and association rules algorithms we can:

- group the patients in clusters with similar characteristics regarding diagnosis and its severity and anamnesis data (e.g. family and personal history);

- associate groups with general therapy schema which will then be customized for subgroups or individual;
- prediction of intermediate states and final status of new patients by placing them in a class labeled S (stationary), A (improved) or C (corrected).

5.2 Overview of the logo-DM data mining process

As we have previously mentioned, the main source of data for the Logo-DM system is the data collected in TERAPERS database. We have also presented the main stream of TERAPERS activities in Figure 2. All these activities are materialized in the data stored in a relational database.

To build personalized therapy programs speech therapists need a complex examination for the children. They should also make record of relevant data related to personal and family anamnesis. Analysis of how children articulate phonemes in various constructions allows diagnosis and classification in a certain class of severity of speech disorder. The collection of anamnesis data may provide information related 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.

The applied personalized therapy programs request data such as number of sessions/week, exercises for each phase of therapy and changes of the original program according to the patient's 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 repetitions for each of these exercises and the results obtained.

The estimation of the child's progress materializes data which indicate the moment of assessing the child and his/her status at that time.

Figure 4 partially presents the database schema which contains data collected in TERAPERS. This database, together with data from other sources (e.g. demographic data, medical or psychological research) is the set of raw data that will be the subject of data mining.

In that form, the data is not appropriate for data mining algorithms, so Figure 5 presents the complete sequence of the operations applied, according to the CRISP-DM model, in order to transform them from raw data into useful knowledge.

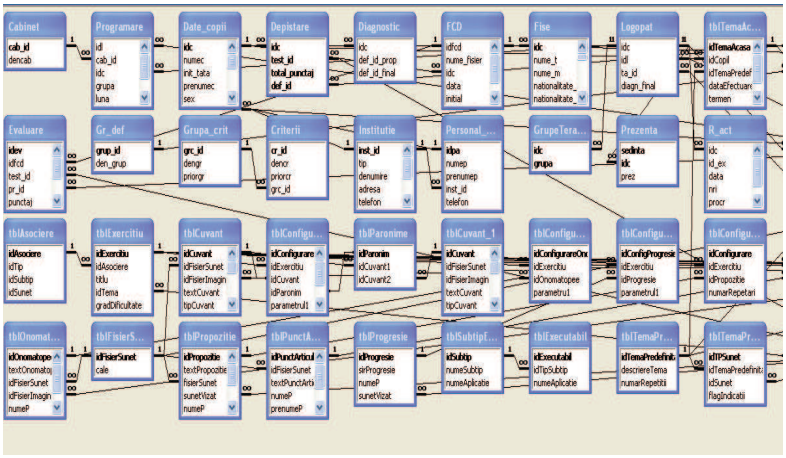

Fig. 4. The TERAPERS Database Schema

As illustrated above and according to Figure 5 the main data source for Logo-DM is a relational database. To avoid actualization anomalies, this database is characterized by a high degree of normalization so various features, potentially useful for data mining are placed in different tables

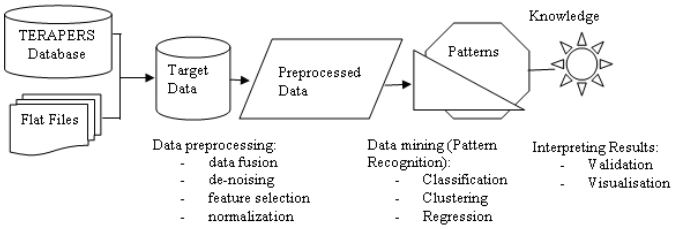

Fig. 5. LOGO-DM view of the end to end data mining process (Danubianu & al, 2009, b)

If we make a system final goals analysis combined with a data analyze we conclude that we need to work only with tables that contain data regarding the children anamnesys and complex logopaedic examination, data regarding different types of speech or language disorders, tests and assessment trials of each test, data regarding the personalized therapy content and management. So, as a first step in data preprocessing is to eliminate those tables which do not contains such data. Thus we achieve a significant decrease in the number of tales used, from 60 to 24. Here we have a superset of necessary data for data mining.

For building and managing of the TERAPERS database we have used Microsoft Access because is cheap and easy to use.

After analyzing the available technologies, it was concluded that the effective implementation of the Logo-DM system can be made with a database management system which entails multi-user, increased security and, last but not least, provide facility for analysis and implement data mining algorithms. We consider that Oracle meets these conditions.

In this context a problem which must be solved is the one concerning the migration of data from MS Access to Oracle. We have done that by using Oracle SQL Developer Migration Workbench. To migrate date it is necessary to cover four steps: capturing the database source, converting the database capture, generating the Oracle database and migrating data.

After data migration from MS Access a series of modifications were required on data types. For example, a data type subject to conversion, is "Date and Timestamp".

The analysis of the database content can reveal interesting issues related to data quality or the need for transformation. We have made a first assessment of data quality through the following measures: completeness, conformity, accuracy, consistency and redundancy. The mechanisms provided by the used database management system have imposed a minimum, controlled redundancy and have assured data consistency. The values were stored in the fields correspond to reality, but unfortunately useful data for analysis are missing from some records. Therefore it is necessary to supplement data gaps, and if this is not possible, the removal of the record for accurate results is suggested.

To obtain proper data for the analysis we should make the following types of transformation: transformations of the structure, and changes aimed at value.

Structural transformations are dictated by the fact that there are fields in the database containing data related to a complex of features to be addressed individually in the analysis.

Value transformation refers to the replacement of coded data by the rules, enabling, for example, the effective storage with descriptive values of characteristics allowing rapid interpretation of results.
An important operation required in this stage is filling the data gaps. We found that, due to the fact that in the TERAPERS database schema some fields values are not restricted to *not null*, these values are partially filled. They should be filled automatically or if this is not possible manually. It is a problem that can be solved in time by setting those fields that have relevance for analysis and by configuring them as *not null* in data sources tables.
Creating target data set is accomplished through joined tables containing useful features followed by a projection on a superset of appropriate attributes, as shown in (1)

$$\Pi_{I_i}(T_1 \rhd\lhd T_2 \rhd\lhd \ldots \rhd\lhd T_k) \tag{1}$$

where: I_i is a superset of the attributes regarding the useful characteristics
$T_1 \ldots T_k$ is the set of tables containing the attributes in the list of projection.
For example, target data set necessary to establish the profile of children with speech disorders, can be obtained by joining tables which contain: general data about children, family and personal anamnesis, data on complex evaluation and associated diagnosis. The result is a set of 129 features. The statement that performs that is presented in (2)

```
create table caract_copii as
    select f.*, l.diagn_final
        from fise f, logopat l
        where f.idc=l.idc;
```
(2)

Data mining techniques were not designed to process large amounts of irrelevant features. Consequently before their application, a selection of the relevant features is required (Guyon & Elisseeff 2003) (Liu & Motoda, 1998). The most important objectives of feature selection are: to avoid over fitting and improve model performance.
Concretely, in the feature selection problem, we are given a fixed set of candidate features for use for building a model, and we must select a subset that will be used to train the model that is "as good as possible" according to some criterion.
We have used for feature selection a variant of the mRMR method (Peng & al, 2005) for categorical values. It is based on mutual information criteria, formally defined, for two discrete random variables X and Y, as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right) \tag{3}$$

where $p(x,y)$ is joint probability distribution function of X and Y, and $p_1(x)$ and $p_2(y)$ are the marginal probability distribution functions of X and Y respectively.
For discrete random variable, the joint probability mass function is:

$$p(x,y) = p(X = x, Y = y) = p(Y = y \mid X = x)^* p(X = x)$$

$$= p(X = x \mid Y = y)^* p(Y = y) \tag{4}$$

Since these are probabilities, we have

$$\sum_x \sum_y p(X = x, Y = y) = 1 \tag{5}$$

The marginal probability function, p ($X = x$) is:

$$p(X = x) = \sum_y p(X = x, Y = y) = \sum_y p(X = x \mid Y = y) p(Y = y) \tag{6}$$

The criterion used is related to minimizing redundancy and maximizing relevance for the chosen characteristics.
The result of the tests performed on data collected from TERAPERS and prepared as described in the above mentioned example revealed that for classification, the minimum error is obtained if we deal with a number ranging from 50 and 55 features selected out of 129.
The target data set, obtained after these steps, is subject to data mining algorithms. For an effective implementation of algorithms we have taken into account, and tested, two possibilities: the use of the Oracle Data Mining kernel (ODM) which offers the possibility to apply algorithms for classification, clustering and association rules and the use of some open source implementations of relevant algorithms adapted and integrated into our own system.
We took into account the types of data included in the set and we used implementations in Oracle of Adaptive Bayes Network, Seeker Model and decision trees build with CART and ID3/C4.5 for classification, in order to cluster the Oracle implementation of A-Clustering algorithm and for association rules Apriori algorithm.

5.3 System architecture

Data mining aims to derive knowledge from data. The architecture of a data mining system plays an important role in the efficiency of data mining.
Data mining systems must satisfy the following requirements:

- not limit the size of the dataset;
- performance optimization should be done for large data sets;
- architecture must serve as a flexible support for various data mining techniques and algorithms such as classification, clustering or association rules;
- they must support the specific priorities of the user or groups of users and they must have the ability to manage concurrent sessions of data mining;
- they should allow a total control of data access;
- they should provide remote administration and maintenance.

The basic components of a data mining system are: user interface, specific data mining services, access services and the data itself.
User interface allows the user to select and prepare data sets on which to apply the data mining techniques. Formatting and presentation of data mining results is also an important task for the user interface. Data mining services comprise all components of a system that processes a special algorithm for data mining, for example the discovery of association rules. These components access data through data access service and can be optimized for certain database management systems, or provide a standard interface such as ODBC. Data is the fourth component of the data mining system. These four components are present in all data

mining systems. In practice there are three different architectures, where these components are distributed on different levels. These are: the one level architecture, the two levels architecture and the most complex architecture is on three levels.
Considering the characteristic of the domain we have proposed for the system a two levels client server architecture. This architecture is presented in Figure 6.
On the client side there is the user interface (GUI) which allows the user to communicate with the system in order to select the task to perform, to select and submit the datasets on which data mining needs to be applied. Pattern evaluation and the post-processing step consisting in pattern visualization are also performed on the client. The knowledge base is the module where the background knowledge is stored.
The more difficult computational tasks of data mining operations are carried out on the server. Here, the data mining kernel contains modules able to perform classifications, clustering and association rule detection. Supplementary the pre-processing data module allows data to become suitable for applying data mining algorithms.

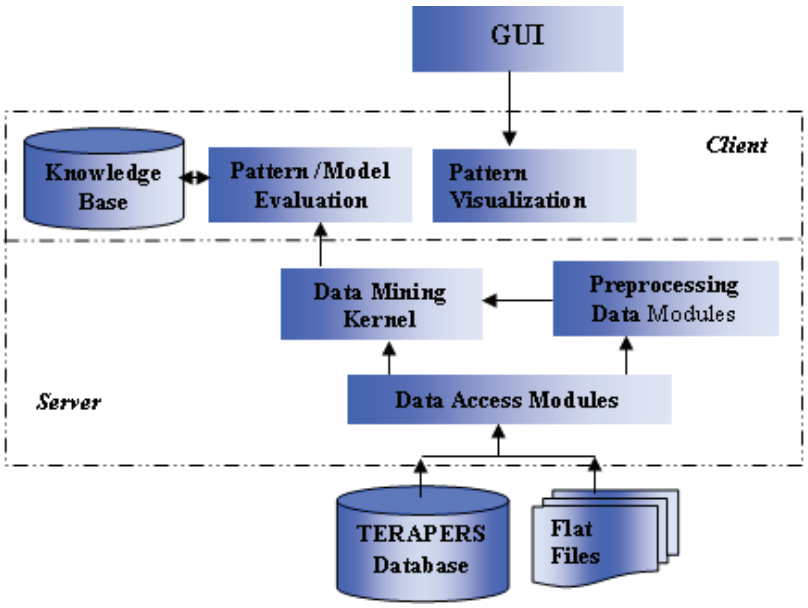

Fig. 6. Logo-DM Architecture (Danubianu &al, 2010)

6. Experimental results

Although the TERAPERS system, which is the main source of data for Logo-DM is used since September 2008, the volume of data which are available to data mining algorithms is still quite low. However we have obtained good results in predicting a patient's status, considering certain characteristics, in different phases of therapy.
Table 1 presents a comparison between the prediction obtained by using the Logo-DM system on the status of a group of 25 children with various forms of dyslalia (rhotacism, polymorphic dyslalia) and the real state of these children, assessed by the therapists at the end of the first two stages of speech therapy.

Patient's state	At the end of the first stage of therapy		At the end of the second stage of therapy	
	Predicted	Real	Predicted	Real
Stationary (S)	1	1	0	1
Improved (A)	24	23	22	23
Corrected (C)	0	1	3	1

Table 1. Results of prediction made by Logo-DM

7. Conclusion

Data mining technology can be a useful tool for the speech disorder therapy because it is able to provide information that enables the implementation of personalized therapy programs optimized and adapted to the characteristics of each child. This leads to a decreased duration of therapy, increasing the possibilities of achieving superior results and ultimately to lower cost of therapy.

Considering the opportunity of data mining techniques application on data collected in the process of speech therapy, we have concluded that methods such as classification, clustering or association rules can provide useful information for a more efficient therapy. Consequently, we have designed and we are currently implementing a data mining system that aims to use data provided by TERAPERS system, developed by the Research Center for Computer Science of "Stefan cel Mare" University of Suceava , in order to optimize the personalized therapy of dyslalia.

We have tested the modules for data pre-processing and on target data sets obtained from these modules, and we have applied more algorithms for detecting the most appropriate solutions for the data mining kernel.

We have obtained good results regarding the prediction of the future state of a new patient.

At present our efforts are directed towards the implementation of visualization modules and towards building a user friendly interface.

8. References

Chapman P., Clinton J., Kerber R., Khabaza T.,Reinartz T.,Shearer C., Wirth R.,(2000) *CRISP-DM 1.0. Step by step data mining guide*, 2000, Crisp-DM Consortium

Cios K., Pedrycz W., Swiniarski R., Kurgan L. (2007*) Data Mining. A Knowledge Discovery Approach*, Springer Science-Business Media, ISBN 978-0-387-33333-5, New York

Danubianu M., Pentiuc St.Gh., Schipor O., Nestor M., Ungurean I., Schipor D.M.,(2009), TERAPERS - Intelligent Solution for Personalized Therapy of Speech Disorders, *International Journal on Advances in Life Science*, Vol. 1, No. 1, 2009, p.26-35, ISSN: 1942-2660

Danubianu M., Pentiuc St. Gh., Socaciu T. (2009) Towards the Optimized Personalized Therapy of Speech Disorders by Data Mining Techniques, *Proceedings of The Fourth International Multi Conference on Computing in the Global Information Technology ICCGI* 2009, pp. 1-6, ISSN/ISBN 978-0-7695-3751-1, France, 23-29 August, Cannes - La Bocca, IEEE Computer Society Conference Publishing Services

Danubianu M., Pentiuc St.Gh., Tobolcea I, (2010) Advanced Information Technology-Support of Improved Personalized Therapy of Speech Disorders, *Proceedings of*

*International Conference on Computers, Communications & Control, ICCCC 2010,*Romania, May 12-16, 2008, Baile Felix

Guyon I., Elisseeff A.,(2003) *An introduction to variable and feature selection*. J. Mach Learn Res., 3, p.1157–1182, 2003

Han J, Kamber M., (2000), *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, 2000

Liu H., Motoda H. (1998) *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA, 1998

OLP Ortho-Logo-Paedia (2002) Project for Speech Therapy (http://www.xanthi.ilsp.gr/olp)

Peng H, Long F, Ding C (2005) *Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, p. 1226-1238, 2005

Speechviewer III (http://www.synapseadaptive.com/edmark/prod/sv3)

STAR (2002) Speech Training, Assessment, and Remediation (http://www.asel.udel.edu/speech)

Knowledge-Oriented Applications in Data Mining
Edited by Prof. Kimito Funatsu

ISBN 978-953-307-154-1
Hard cover, 442 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by 'Data Mining' address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Danubianu Mirela, Tobolcea Iolanda and Stefan Gheorghe Pentiuc (2011). Data Mining in Personalized Speech Disorders Therapy Optimization, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1, InTech, Available from: http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/data-mining-in-personalized-speech-disorders-therapy-optimization

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821