We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

16

Data Mining in Neurology

Antonio Candelieri[1,2], Giuliano Dolce[1],
Francesco Riganello[1] and Walter G Sannita[3,4]
[1]*Research in Advanced Neurorehabilitation, Intensive Care Unit S. Anna Institute,*
[2]*Laboratory of Decision Engineering for Health Care Delivery, University of Cosenza,*
[3]*Department of Motor Science and Rehabilitation, University of Genova,*
[4]*Department of Phychiatry, State University of New York,*
[1,2,3]*Italy*
[4]*USA*

1. Introduction

Data Mining intersects database technology, modelling techniques, statistical analysis, pattern recognition, and machine learning. It makes use of advanced tools for large databases management and automatic/semiautomatic analyses in order to identify significant trends and associations deemed informative because novel, implicit to the data, and of potential support in prediction and decision making.
Methodological relevance and application in healthcare and biomedicine are increasing, with implications in fields as different as information management in healthcare organisation, public health, epidemiology, patient monitoring and management, signals and images analyses. It essentially represents an effective and efficient solution providing new predictive criteria for early diagnosis and prognosis, or supporting medical staffs in patient management such as in therapy planning and personalization. Knowledge extracted from pertinent clinical databases through data mining techniques may be new or suitable of integration with consolidated knowledge and improve reliability while reducing subjectivity in decision making processes.
In this chapter we discuss about the general rationale underlying Data Mining and its peculiarities of application in the medical field, notably in the neurological domain. Relevant decision making problems, proposed solutions and open issues are summarized and the state-of-the-art of Data Mining in medicine and neurology is discussed in perspective. This review cannot and is not meant to be exhaustive, but should outline the potential use of Data Mining for supporting clinicians in their decision making.

2. Rationale and background

Data Mining was introduced in 1989 by Fayaad as a non-trivial process to identify reliable, novel, and potentially useful patterns in large data sets (Fayaad, 1996) though an iterative and multidisciplinary approach based on interaction with the application domain expert, data pre-processing, acquisition of consolidated knowledge, selection and use of the most suitable Data Mining methods, and evaluation and post-processing of the results. In this

regard, Data Mining is regarded as a step in a wider process known as Knowledge Discovery in Databases (KDD), or simply Knowledge Discovery.

Novel knowledge implicit to the dataset and of potential use can be extracted through several approaches following five different tasks or learning processes: Classification and Regression (Supervised Learning), Clustering (Unsupervised Learning), Association Rule Learning and Feature Selection. In Classification and Regression, a set of cases (instances) is available, where each case is represented by a set of variables (attributes) of varying size. One of these variables is the "target" attribute of the learning process: in classification tasks (*i.e.* diagnosis, good or poor prognosis, any rating at the outcome scales, etc.) it is a nominal variable and represents the "class" (group) to which each instance belongs. In Regression tasks the target attribute is a numeric variable (*i.e.* systolic/diastolic blood pressure, heart rate, glucose concentration, etc.). If a target (either nominal or numeric) variable exists, the learning task is "supervised" because the learning strategies try to find a reliable relationship of other attributes with the target. In this regard, supervised learning techniques may be used *e.g.* to find diagnostic/prognostic criteria or predict trends in clinical or vital factors depending on the subjects' profile (*e.g.* glucose concentrations to be expected based on genetic information, familiarity, life style, etc.).

Unlike Classification and Regression, Clustering is known as an "unsupervised" learning task, in which no target variable is identified: instances are essentially clustered at different levels, according to a predefined similarity or distance measure. Instances which are "near" may be considered similar and belonging to the same "class" or cluster according to the distance measures. Clustering algorithms may be also adopted in Classification tasks: the target attribute is in this case excluded from the analysis and instances are clustered according to a predetermined distance measure; if instances belonging to the same cluster are also attributable to the same class (target variable preliminary excluded), the adopted distance measure may be considered a reliable relationship among all the other attributes value and the target one.

Association Rule Learning is meant to identify relationship among attributes, with no reference to any particular target variable: in this procedure, the relevant relationships are ranked and presented to the analysts for their evaluation (*e.g.*, correlations among numeric variables are ranked according to their significance level).

Feature Selection approaches are adopted to identify the attributes that are more relevant for the learning goals. The selection of the most relevant attributes allows to 1) provide the analyst with a first "return of knowledge" about the main involved factors, and 2) reduce the computation time of learning tasks while improving reliability.

Several methodologies and implementations are today available for each Data Mining task; each one presents at least one parameter to be set to modify the "structure" and reliability of the extracted pattern (knowledge representation). The *a priori* identification of the most useful methodology and parameter(s) configuration is usually difficult; to test different methodologies/implementations and different parameter(s) values is a weaker, yet practicable approach.

A crucial issue in the use of any Data Mining task is the reliability evaluation of the extracted knowledge on data not used for analyses. This is more relevant in the medical field, when reliability is a predictive criterion for new individual patients.

The exploratory and confirmatory process in science provides a useful perspective on the problem of circular analysis. Hypotheses generated by exploring the dataset require

confirmation by means of independent data, because any relationship observed in a dataset will be consistent with it irrespective of a true relationship. An independent dataset for selective analyses would serve to ensure independence of the results under the null hypothesis and thus prevent circularity.
Data Mining methodologies may suffer from circularity and need independent datasets for validation. This use of the same dataset for selection and selective analysis, also known as "double-dipping" (Kriegeskorte et al., 2009) is crucial. Independent data may be unavailable, but suitable validation techniques to estimate the reliability of the predictive model "mined" from data are at hand. Cross-validation essentially works by repeatedly splitting the dataset into K smaller, independent subsets and to take advantage of a split-data analysis. A single subsample is adopted as test set and the remaining K-1 subsamples are used for training. The selection process (training) and test must be performed independently for each cross validation fold, and the procedure increases the computational demands as the independence each split-off subset needs to be guaranteed when implementing a correct cross-validation scheme.

3. Data mining in medicine

In recent years, computer technology is increasingly implemented in healthcare to meet the needs of solutions supporting clinicians in their daily decision making activities. In this regard, Data Mining tools may be useful to control for human limitations such as subjectivity or errors due to the fatigue and to provide ready indications for the decision processes (early diagnosis and prognosis, improvement or worsening, etc.).
Predictive models provide the best support to the clinicians' knowledge and experience. In order to reduce subjectivity, several (expert) systems have been proposed to codify and provide consolidated medical knowledge. Data Mining can be integrated into these systems to reduce subjectivity while providing potentially useful new medical knowledge (evidence-based medicine). For instance, Bratko and co-workers have developed a system to interpret ECGs through models extracted by Data Mining techniques (KARDIO; Bratko et al, 1989). Several techniques to analyze biomedical data from tissues or body fluids have been developed to obtain predictive models and identify small sets of relevant variables (biomarkers) to be used for validation. Schummer and colleagues, have applied cluster analysis to compare breast cancer and healthy tissues and identify markers for early diagnosis and prognosis (Schummer et al., 2010). In particular, authors identified 43 differentially expressed genes and compared them on two sets of data from breast cancer patients with good or poor outcome and from healthy women undergoing reduction surgery, respectively. The study identified three genes with high expression only in cancer with poor outcome and further research on their reliability as markers in the early detection of cancer with poor outcome is in progress. These authors also found that some histologically normal breast tissues removed from distant site in a breast with cancer displayed a cancer-like expression profile, suggesting that these regions may be predisposed to malignancy despite apparent histological normality. These findings might help in the early diagnosis and treatment.
Another relevant application is in the processing of biomedical signals expressive of internal regulation and responses to stimulus conditions, whenever detailed knowledge about interactions among different subsystems is lacking and standard analysis techniques may be ineffective, as it is often the case with non-linear associations. In this regard, Data Mining

allows identify relationships explaining continuous data, such as biomedical signals acquired on patients in the Intensive Care Units, and develop intelligent monitoring systems also sending reminders, alerts and alarms for preselected critical conditions.

A paradigmatic medical field benefitting from this approach is cardiology, where the analysis of monitored vital parameters signals the patient's worsening or alarming events. Candelieri and colleagues have proposed Data Mining in the early detection of criticalities in chronic heart failure patients monitored in remote at home, to be compensated for by prompt action avoiding hospitalization (Candelieri et al., 2008 and 2009; Candelieri & Conforti, 2010); in this project, a Classification task was performed on a limited number of vital parameters (systolic blood pressure, heart rate, body temperature, body weight) easy to be acquired in semi-automatic/automatic way. Approach proved able to reliably predict a patient's risk of heart criticality in two weeks, with reduced healthcare costs and improved quality of life. Some extracted criteria also provided the medical staff with a return of knowledge "easy-to-understand" because obtained through methodologies using understandable codification patterns such as Decision Trees and Rule Learners (Candelieri et al., 2008).

Data Mining techniques also assist clinicians in the diagnosis through computer-based systems for the interpretation of images with medical relevance (endoscopy, ecography, radiology, tomography, ultrasonography, magnetic resonance, etc.). These systems aim is usually to reproduce the specialists' expertise in the pre-identification of the affected regions (Innocent et al., 1997; Zhu and Yan, 1997; Phee et al., 1998; Veropoulos et al., 1998; Karkanis et al., 1999).

Data Mining also offers a support to identify reliable relationship between the patients' profiling or therapy and outcome. Madigan and Curet used Data Mining to predict the length of hospitalization and destination after discharge of patients with obstructive pulmonary disease, heart failure and hip replacement (Madigan & Curet, 2006) at variance with the limits and poor suitability of traditional statistical approaches (Iezzoni, 2004). Data from 580 patients living in the US were obtained though the 2000 National Home and Hospice Care Survey (NHHCS) and the survey which was conducted by the National Center for Health statistics (NCHS). CART (Classification and Regression Trees) were applied with two purposes, namely to identify the parameters predictive of the destination at discharge and length of hospitalization (home healthcare service outcome), and investigate the applicability of Data Mining in the analysis of home healthcare data. The patient's age (especially when 85 or older) was a relevant factor both in destination and length of stay, irrespectively of the disorder. Other contributions were from type of agency and payment, and ethnicity; in particular, hospitalization was shorter for hospital-based agencies.

Some caution was expressed about the relations identified by CART, which were suggested to explain the dataset without any cause-and-effect relationship been implicated; in particular, the type of agency was associated with outcome, but ranking the agencies standard on this basis would be wrong. Despite the appropriate cautions, however, the study confirmed the potentialities of Data Mining in home healthcare monitoring.

Lu and colleagues (Lu et al., 2006) ran a Data Mining study to detect impaired motility in elderly subjects and provide information about risk factors to be used when planning interventions for outcome improvement. Authors started from the prevision that by the 2030 more than 70 million of people in the US will be elderly (65 or older) and mobility will play

a key role in health management (Center and Disease Control and Prevention, and Merck Institute of Aging & Health in American, 2004). The study was performed on a dataset of 8259 patients with eight demographic and patients' care attributes (age, gender, race, service, primary insurance, marital status, religion, and disease code) by means of a Decision Tree algorithm (J48) able to predict impaired mobility and a ten-fold cross validation. Feature Selection methods (Wrapper Subset Evaluator and Naïve Bayes classifier) were also applied to pre-identify relevant attributes and therefore ameliorate the Decision Tree performance. J48 provided an initial accuracy of 69.5% (specificity: 70%; sensitivity: 69%) when using all variables and a 68,5% accuracy after a Feature Selection reduction to five variables (specificity: 72%; sensitivity: 65%). (The three attributes proving useless were race, primary insurance and religion).

4. Data mining in neurology

Herskovitz and Gerring (Herskovitz & Gerring, 2003) suggested that Data Mining techniques may be applied for a better understanding of the existing relationships among variables obtained from lesion-deficit analysis (LDA). Bayesian methods proved computationally tractable, effective in representing non-linear associations among LDA variables and more sensitive and specific than methods based on Chi-square and Fisher exact statistics. LDA provides extensive information about associations between the brain structure and function, but usually generates large cohorts of variables, thus making the modelling of data relations by traditional statistical approaches difficult.

4.1 Neurological diagnosis and prognosis

Decision in the management of traumatic brain injury patients may be crucial. In particular, neurologists and neurosurgeons usually have to make decisions in a short time and on the basis of several patient's data. Several studies analyzed genomic data, clinical parameters at admission, and laboratory tests while comparing different Data Mining techniques.

Ji and colleagues proposed a Data Mining procedure to provide the clinician with useful guidelines supporting the decision making processes in the management of traumatic brain injury patients (Ji et al, 2009). They proposed a multi-level system able to give suggestions congruent to the condition in which data were acquired: on-site (data acquired at the side of accident), off-site (information acquired at admission to the hospital, such as co-morbidities and complications), and helicopter (data acquired during transportation to the hospital). The on-site and off-site dataset were used to obtain predictive models about the patients' outcomes (survival, clinical outcome with rehabilitation or at home), while the helicopter dataset was used to work out a model able to predict the length of hospitalization in intensive care unit (ICU). The days in ICU ranged between 0 and 49, but data was clustered in two groups of non-severe and severe patients' with cut-off stay in ICU at two days. The decision problems (survival and outcome predictions and estimation of ICU length of stay) were defined as classification tasks and were approached by AdaBoost, C4.5 (Decision Tree algorithm), CART, Artificial Neural Network with Radial Basis Functions (RBF-ANN), and Support Vector Machine (SVM). Authors also adopted the Feature Selection methods (specifically the Logistic Regression identifying the most significant variables prior to the training process) in order to improve the classifiers performance. All classifiers were evaluated through ten-folds cross validation techniques. A combined C4.5-CART approach

using the significant variables proved the best solution for the three decision problems, with accuracy of 84% and 89.7% for survival and clinical outcome prediction, respectively. The C4.5-CART combination attained an accuracy of 93.1% in predicting ICU permanence of patients transported to hospital by helicopter.

The system proposed by Ji and colleagues can be regarded as an effective support tool improving the clinician diagnostic and prognostic accuracy of traumatic brain injury; it will be tested in all the 17 hospitals of the Carolina Healthcare system (CHS), improved and made available to the research community (as a web-based or stand-alone application) in order to receive useful feedbacks.

Important previous studies aimed at optimizing management of traumatic brain injuries by using data mining methodologies were performed by Choi and colleagues (Choi et al., 1991) used decision trees procedures to predict outcome after severe brain injuries and achieved 77.7% correct predictions. Nissen and colleagues (Nissen et al., 1999) used Bayesian networks to predict poor or good outcome (death, survival in a vegetative state or with disabilities) with 75.8% overall accuracy.

More recently, Yin and colleagues (Yin et al., 2006) carried on a pilot study on the effectiveness of different analysis strategies in the prediction of outcome after severe brain injury; they used Bayesian Networks, Decision Trees, Logistic Regression, Support Vector Machines and Artificial Neural Networks on a dataset of over seven hundred patients with severe brain injury and estimated the model performance through ten-folds cross validation. The rating at the Glasgow Outcome Scale (GOS) (Jennet and Bond, 1975) was assessed for each patient and represented the target attribute of the Data Mining analysis. Class labels were defined in six different ways and all the related classification definitions were investigated: a- five different classes corresponding to the five GOS classes (1=death; 2=vegetative state; 3=severe disability; 4=moderate disability; 5=full recovery or recovery with minor disabilities); b- three classes (obtained by clustering GOS classes 2, 3 and 4); c- two classes (obtained by clustering GOS classes 1 with 2 and classes 3, 4 and 5); d- two classes (clusering GOS classes 2, 3, 4 and 5; e- two classes clustering GOS classes 1, 2, 3 and 4; f- two classes, clustering GOS classes 5 with 4 and GOS classes 1, 2 and 3.

A reliable model predicting the outcome proved not practicable, but several aspects to be taken into account for this kind of studies were outlined. In particular, the validation techniques for evaluating the realistic prediction reliability of extracted models and then the significant influence of outcome classes aggregation on prediction performance proved crucial. No individual algorithm outperformed the others, and authors suggested to apply multiple algorithms in parallel to reduce errors.

Grzymala-Busse and colleagues confirmed these findings in a study (Grzymala-Busse et al., 2008) in which they compared the classification methods LEM2 (a rule-based learner) and BeliefSEEKER (a Bayesian network-based approach) on 42 clinical variables. Nets obtained through BeliefSEEKER were successively converted into set of rules to be compared with those obtained through LEM2 in order to discover a set of rules predicting the most probable outcome after severe brain injury. BeliefSEEKER produced simpler rules than LEM2 and proved most performing at the ten-folds cross validation. Weak rules could be removed from the LEM2 set therefore improving its performance to the same level of BeliefSEEKER. It was concluded that these Data Mining approaches are comparable, without any indication as to clinical usefulness being given.

Early prognosis is necessary for subjects in a vegetative state (a condition of severe impairment of consciousness requiring continuous care); the issue was approached via

Decision Tree (Dolce et al., 2008a) and Artificial Neural Networks (Pignolo et al., 2009) in studies analyzing the appearance/disappearance of twenty-two relevant clinical signs with respect to outcome in three hundred and thirty-three subjects in a vegetative state, whose outcome was rated according to the Glasgow Outcome Scale. Aim of studies was to identify the clinical signs observed by the medical staff at the admission and after 50, 100, and 180 days to be used as markers of good or poor outcome.

A model (Figure 1) based on CART algorithm proved reliable in predicting the outcome after identifying a limited set of significant clinical signs and the timing of observation (Dolce et al., 2008a). Performance as evaluated through cross-validation techniques ranged from 74% to 83% depending on the follow-up time point. Outcome was good (GOS classes 4 and 5; accuracy: 89-91%) when visual pursuit (or eye tracking) and spontaneous motility re-appeared and oral automatisms disappeared early during the follow-up. Absence of eye tracking and spontaneous motility at any time point and appearance of oral automatisms at 100 days after the admission indicated poor outcome (GOS classes 1 and 2) with 80-100% accuracy. Aetiology proved a relevant variable particularly at the initial phases of follow-up, with better outcome for traumatic brain injuries.

The relationship between clinical signs appearance/disappearance and outcome was investigated by same research group with Artificial Neural Networks (Pignolo et al., 2009) in a model equating a neural network to a "black box" with no return of easy-to-understand knowledge. The results were comparable, but the decision tree-based model (Dolce et al., 2008a) performed better and proved more understandable.

4.2 Therapy planning and rehabilitation

Catalano and colleagues used Data Mining analysis to investigate the interaction of demographics and parameters such as work disincentives and vocational rehabilitation services patterns with the employment outcome of traumatic brain injury patients (Catalano et al, 2007). Traumatic brain injury patients could be clustered in 29 homogeneous subgroups with different employment rates ranging from 11% to 82%, where differences were essentially explained by work disincentives, race and rehabilitation service variables. In particular, European Americans showed a higher employment rate (53%) than others ethnic groups: Native, Asian, African, and Hispanic/Latino Americans with employment rate of 50%, 44%, 42%, and 41% respectively. Furthermore, subjects without psychiatric disabilities and work disincentives had a higher employment rate than those with such characteristics (51% versus 41% and 58% versus 45%, respectively). Vocational rehabilitation service features (notably, job search and placement assistance and on-the-job support services) were relevant in predicting employment outcomes for traumatic brain injury patients.

More recently, Gibert and colleagues proposed an approach integrating Data Mining techniques, traditional statistics, and tools for interpretation to predict the evolution of life quality among patients with spinal cord injury (Gibert et al., 2009) with a life expectancy comparable to the healthy, but persistent disability.

Differing psychological responses to similar physical impairments were observed, suggesting that factors generating negative psycho-emotional responses (*i.e.* depression) and worse quality of life should be promptly identified to provide the subjects with appropriate assistance.

All studied patients were in follow-up after discharge and were periodically evaluated by the Periodic Integral Evaluation (PIE), with procedures taking into account aspects of medical, functional, neuropsychological, social, health education and health risk prevention

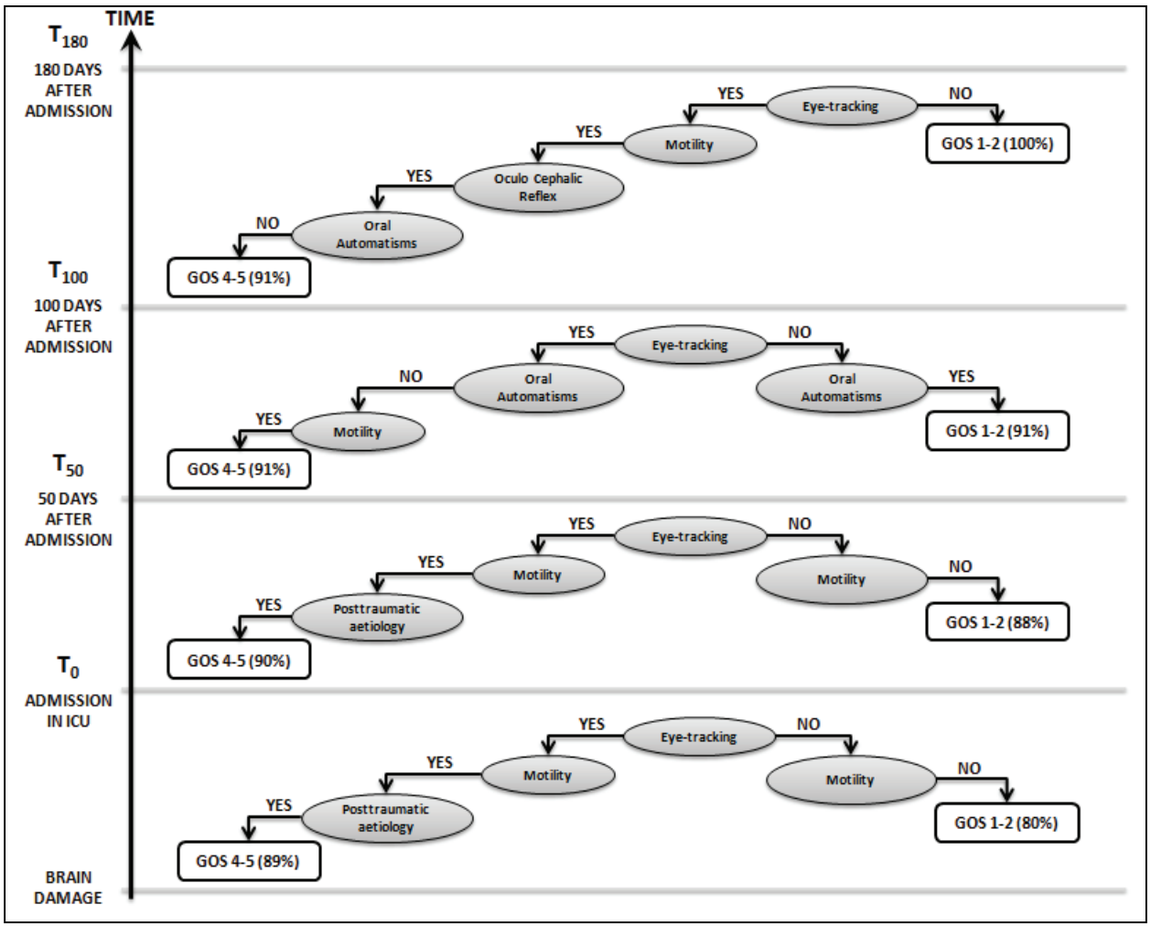

Fig. 1. Eexample of application of a CART model in the prediction of outcome of patients in vegetative state. Four major clinical signs and the timing of their observation provided the significant information accounting for the model. The overall cross-validated accuracy of prediction ranged 74-83% for each time point during the follow-up. Recovery of spontaneous motility, eye tracking and oculo-cephalic reflex not observed at admission or in early phase of clinical monitoring and the disappearance of oral automatisms correlated with a positive outcome (i.e. in classes 4 or 5 of the Glasgow Outcome Scale; correct prediction was in 89-91% of patients). Absence of eye tracking and mobility at any time point and the appearance of oral automatisms at T100 or later were indicative of poor prognosis (classes 1 or 2 of the Glasgow Outcome Scale; 80-100% accuracy depending on time of observation). Aetiology proved crucial at T0 and T50, when reappearing eye tracking and spontaneous motility allowed a favourable prediction of outcome in 89-90% of patients in VS due to traumatic head injury, to become irrelevant at T100. A retrospective cohort study in about 400 subjects in a vegetative state (Dolce et al., personal communication) confirmed the predictive power of eye tracking as indicated by the CART model. The patients' rating at the Glasgow Outcome scale after a 250-day follow-up was better in those subjects with recovered visual tracking and inversely correlated with the time of re-appearance (i.e. early recovery reliably predicted better outcome), although subjects with late recovery of eye tracking (after 230 days or more) had better outcome than those without it.

to predict asymptomatic pathologies and prevent complications, long hospitalization and survival risks. As result, PIE ratings were mainly characterized by the interaction among the patients' functionality and psychological variables, whereas demographic and social features appeared irrelevant with the exception of time from injury (that had negative effects on the quality of life), academic degree and living in couple. Psycho-emotional responses were not correlated to the severity of brain lesion, although patients with more severe impairment have a worse quality of life.

With the identification of targeted therapies remaining a critical goal, application of data mining techniques is increasing. Saatman and colleagues have outlined the needs of a new classification system for therapeutic interventions in traumatic brain injury, suggesting to adopt tools for intelligent data analysis such as Knowledge Discovery and Data Mining methods (Saatman et al., 2008).

The clinical classification systems in use can be subdivided into *etiological*, *symptom*, *prognostic* and *pathoanatomic* classification systems. Traumatic brain injuries are usually classified by one of three main systems: clinical indexes of severity, pathoanatomic classifiers, and physical mechanism evaluation schemes. Clinical indexes of severity belong to symptom classification systems and remain the major inclusion/exclusion criteria in clinical trials for traumatic brain injury. The Glasgow Coma Scale (GCS) severity scale is commonly used because of its high inter-observer and prognostic reliability (Teasdale and Bennet, 1974); it assesses the consciousness level after brain damage, is of help for early prognosis (*e.g.* at admission), has allowed develop three prognostic models of increasing complexity, and proved informative for clinical management and prognosis, but offers no information about physiopathology mechanisms of neurological deficits (Murray et al., 2007).

Several schemes were proposed and adopted to characterize the pathoanatomy of brain injury, including the Marshall score for Computerized-Tomography (CT) images (Marshall et al., 1992) and the Rotterdam score (Maas et al., 2005). The first one proved to be reliable in predicting both the risk of increased intracranial pressure and outcome in early severe and moderate traumatic brain injury adults, but presents many limitations in classifying patients with multiple brain damage types and standardization of certain CT features. On the other hand, the second score system is well standardized and able to predict outcome, but is too recent and not fully validated.

About traumatic brain injury classification by physical mechanism there is a considerable, but not perfect, correlation with pathoanatomic damage type. Under this respect, mechanistic classification may be really useful in modelling injuries and prevention, but not in clinical practice because the usually incomplete details of the traumatic event. In traumatic brain injury classification, physiopathologic mechanisms may be adopted to characterize targets for treatment. One widely accepted and used schema consists in differentiating "primary" versus "secondary" damage. The first refers to the unavoidable damage occurring at time of injury, and the second to secondary insults, such as hypoxia, hypertension, etc. However these systems are not commonly used in treatment trials because limited availability and usage of sophisticated monitoring parameters.

Saatman and colleagues suggested that improved procedures for classification would help understand pathological mechanisms in greater detail, while supporting the clinician's titration of treatment and improving outcome. Advances in diagnostic tools, technical solutions and intelligent methods for data analyses would promote the development of multidimensional classification systems for traumatic brain injury based on diagnostic,

prognostic, anatomic, and pathophysiological parameters and able to select patients potentially benefitting of medical interventions and the best treatment (Saatman et al., 2008). An *ad hoc* committee would develop the multidimensional database and provide solutions for data sharing and data mining in order to facilitate collaboration and knowledge discovery (Saatman et al., 2008).

4.3 Image and signal analysis

Computer-assisted systems for images and signals interpretation support prompt decisions and reduce subjectivity of the assessment. Liao and colleagues proposed a novel method based on a combination of machine vision with Data Mining to automatically detect intracranial hematomas through the analysis of CT brain scans (Liao et al., 2007).

CT is usually preferred in the emergency room when intracranial hematomas are suspected and type, location and shape (*e.g.* epidural, subdural or intracerebral) are to be defined. However, identification following the current guidelines remains essentially qualitative. The model proposed by Liao and colleagues was able to a- identify hematomas on digital CT slices by a machine vision technique; b- assess severity by automatic labelling of pixels by depth and affected regions; and c- apply a decision tree-based algorithm to provide hematomas diagnosis independent of, and to support clinical diagnosis. Data Mining techniques (C4.5 decision tree) identified a reliable relationship between the features measured by machine vision and the clinician's diagnosis. The approach was evaluated on 48 pathological images and provided two decision rules similar to those used by medical experts and able to make correct diagnosis. The method resulted faster, less expensive and safely applicable also to patients with unstable vital signs than the magnetic resonance imaging (RMI) and was congruent with the development of a good clinical decision support system.

Application of data mining to the functional magnetic resonance imaging (fMRI) is aimed at developing paradigms for functional investigation in the absence of *a priori* hypotheses. Several approaches, such as Support Vector Machines and Fisher discriminant analysis were applied, mostly for patter recognition (Haxby et al., 2001; Haynes and Rees, 2006; Ku et al., 2008; and Hardoon et al., 2007).

Blaschko and colleagues proposed a semi-supervised regression analysis using data at rest (Blaschko et al., 2009). Resting state activity is defined as the background level of brain activation in the absence of functional tasks and is generally measured in the awake subjects by long fMRI scanning sessions where the only instructions given to subjects are to close the eyes and do nothing. The spontaneous fluctuations of neural activity in these conditions are thought to provide relevant information on brain structural and functional aspects. For example, some brain regions resulted more active at rest than while performing a task, therefore suggesting a sort of (homeostatic) default brain state (Biswal et al., 1997; Raichle et al., 2001; Raichle and Snyder, 2007), whereas spontaneous fluctuations were usually found directly correlated to metabolic activity and behavior (Biswal et al., 1997; Bianciardi et al., 2009).

However, fMRI analyses at rest are limited by the absence of time-locking to events, and traditional statistics may become useless due to noise. Blaschko and colleagues adopted semi-supervised learning techniques to improve the accuracy of a regression model based on fMRI changes in response to stimuli (viewing a movie), making use of instances with and without the related class labels in order to obtain a reliable relationship among the input

and output variables. They noted that brain activity at rest is similar to that induced by stimulus conditions, allowing to augment functional data by using resting state data acquired for completely different purposes (e.g., baseline recording).

The processing of biomedical signals related to internal regulation and response to stimuli may benefit of data mining techniques whenever knowledge about (non-linear) interactions among different subsystems is lacking, *a priori* hypotheses are difficult to formulate and traditional statistics are not practicable. Signals are usually less expensive to acquire than images and can be recorded over time with negligible discomfort (as in the case of monitoring). Recording procedures are non-invasive and Data Mining techniques are powerful solutions when useful knowledge of the regulation and response mechanisms are to be investigated.

Riganello and coworkers applied several classification algorithms to detect in healthy controls, posttraumatic patients and subjects in vegetative or minimally conscious states a reliable relationship between the emotional status induced by complex sensory stimuli (symphonic music). The emotional responses were independently classified by the controls and posttraumatic patients' report and by the heart rate variability (HRV) parameters in all subjects (Riganello et al., 2008). A model based on one HRV parameter (nominally the normalized unit of low frequency band power, nu_LF, with low frequency band power ranging from 0.04 to 0.15 Hz) could classify the emotional responses in all subjects (including those in vegetative or minimally conscious state) with accuracy (evaluated via suitable cross-validation techniques) of about 70% for both healthy subjects (training set) and posttraumatic patients (independent test set).

The applicability of the knowledge acquired on healthy and traumatic subjects to investigate brain processing in patients in a vegetative state was tested (Riganello et al., 2010a; Riganello & Candelieri, 2010). A comparative analysis and validation on different data mining methods is reported elsewhere (Riganello et al., 2009).

Following a comparable approach, Dolce and coworkers identified by HRV spectral analyses the emotional response of subjects in a vegetative state patients to the presence or voice of a relative (the *mom's effect*) (Dolce et al., 2008b). Although preliminary, these findings suggest that autonomic concomitants of emotional changes can be induced by complex stimuli also in vegetative state, with implications on the residual responsiveness of these subjects.

The model used to classify the emotional response to symphonic music (Riganello et al., 2010a) was applied retrospectively, without retraining, to analyze the emotional response of 12 subjects in a vegetative state to a relative (the "mom's effect") (Dolce et al., 2008b). The emotional condition was classified as being "positive" or "negative" in an experimental paradigm including baseline, the mother's presence or voice (test condition), and the presence/voice of persons unfamiliar to the subject (control or sham condition). Data mining classified the emotional response as being "positive" in 8 subjects in the test condition and as "negative" in 11 subjects in the control condition (Riganello et al., 2010b).

5. Conclusions and discussions

Data Mining - the non-trivial process of identifying valid, novel, and potentially useful patterns in data (Fayaad, 1996) - has been applied with success to different fields, such as engineering, banking, marketing and customer relationship management, and various areas of science.

To date, application in the analyses of datasets with medical/neurological relevance is still limited. However, several approaches are suitable of use in the investigation of practical problems in medicine and the expectations are that efficient and practicable solutions will be made available in increasing number and variety of application.

The potentialities of Data Mining in clinical medicine is mainly in the identification of relations, patterns and models supporting prediction and the clinician's decision making processes, *e.g.* for diagnosis, prognosis, and treatment planning. When validated, these predictive models could be embedded in the clinical information systems as clinical decision support modules, reducing both subjectivity and time in making decisions.

Application in the medical field differs from the Data Mining use in business, marketing and economy. For instance, medical datasets and decisions are usually biased to some extent by measurement errors, missing data or miscoding the information in textual reports. In addition, some major issues concern the processes of knowledge extraction and representation: decision making should be supported by conceptually user-friendly models (*e.g.* decision tree and rule sets) rather than by "black-box" models (*e.g.* artificial neural networks and support vector machines). The reliability and wide application in fields such as images and signals interpretation notwithstanding, research on Data Mining should focus in greater detail also on the extraction of understandable rules from trained black-boxes (such as neural networks); theoretical research should provide mathematical justifications for the properties of Data Mining algorithms (Magoulas & Prentza, 2001).

6. References

Bianciardi, M., Fukunaga, M., van Gelderen, P., Horovitz, S. G., de Zwart, J. A. & Duyin, J. H. (2009). Modulation of spontaneous fMRI activity in human visual cortex by behavioral state, *Neuroimage*, 45, 1, 160-168. 1053-8119.

Biswal, B. B., Van Kylen, J. & Hyde, J. S. (1997). Simultaneous assessment of flow and bold signals in resting state functional connectivity maps, *NMR Biomed*, 10, 4-5, 165-170. 0952-3480.

Blaschko, M., Shelton, J. & Bartels, A. (2009). Augmenting feature-driven fMRI analyses: semi-supervised learning and resting state activity, *Proceedings of the 2009 Conference on Neural Information Processing Systems (NIPS 2009),* 126-134. 01-2010.

Bratko, I., Mozetic, I. & Lavarac, N. (1989). KARDIO: a study in deep and qualitative knowledge for expert systems, *Cambridge, Massachusetts: MIT Press*. 0262022737.

Candelieri, A., Conforti, D., Perticone, F., Sciacqua, A., Kawecka-Jaszcz, K. & Styczkiewicz, K. (2008). Early detection of decompensation conditions in heart failure patients by knowledge discovery: The HEARTFAID approaches, *Procedings of Computers in Cardology 2008*, 893-896. 0276-6547.

Candelieri, A., Conforti, D., Sciacqua, A. & Perticone, F. (2009). Knowledge Discovery Approaches for Early Detection of Decompensation Conditions in Heart Failure Patients, *Proceedings of ISDA, 2009 Ninth International Conference on Intelligent Systems Design and Applications*, 357-362, 2009. 978-0-7695-3872-3.

Candelieri, A. & Conforti, D. (2010). A Hyper-Solution Framework for SVM Classification: Application for Predicting Destabilizations in Chronic Heart Failure Patients, *The Open Medical Informatics Journal*, 4, 135-139. 1874-4311.

Catalano, D., Pereira, A. P., Wu, M. Y., Ho, H. & Chan, F. (2006). Service patterns related to successful employment outcomes of persons with traumatic brain injury in vocational rehabilitation, *Neurorehabilitation*, 21, 4, 279-293. 1053-8135.

Center and Disease Control and Prevention, and Merck Institute of Aging & Health in American. (2004). The state of aging and health in american, *Washington DC: Merck Company Foundation.*

Choi, S., Muizelaar, J., Barnes, T., Marmarou, A., Brooks, D. & Young, H. (1991). Prediction tree for severely head injured patients, *Journal of Neurosurgery*, 75, 251-255. 0022-3085.

Dolce, G. & Sazbon, L. (2002). The posttraumatic vegetative state, Thiene, 1-58890-116-5.

Dolce, G., Quintieri, M., Serra, S., Lagani, V. & Pignolo, L. (2008a). Clinical signs and early prognosis in vegetative state: a decisional tree, data-mining study, *Brain Injury*. 22, 7-8, 617-623. 0269-9052.

Dolce, G., Riganello, F., Quintieri, M., Candelieri, A. & Conforti, D. (2008b). Personal interaction in vegetative state: a data-mining study, *Journal of Psychophysiology*, 22, 3, 150-156. 0269-8803.

Fayaad, U. (1996). From data mining to knowledge discovery: an overview. *American Association for Artificial Intelligence (AAAI) Press, Menlo Park*, 1-34, 0-262-56097-6.

Gibert, K., Garcia-Rudolph, A., Curcoll, L., Soler, D., Pla, L. & Tormos, J. M. (2009). Knowledge discovery about quality of life changes of spinal cord injury patients: clustering based on rules by states. *Studies in Health Technology and Informatics*, 150, 579-583. 0926-9630.

Grzymala-Busse, J. W., Hippe, Z. S., Mroczek, T., Bucinski, A., Strepikowska, A. & Tutaj, A. (2008). Prediction of severe brain damage outcome using two data mining methods, *Proceedings of Conference on Human System Interactions*, 585-590. 978-1-4244-1542-7.

Hardoon, D. R., Mourao-Miranda, J., Brammer, M. & Shawe-Taylor, J. (2007). Unsupervised analysis of fMRI data using kernel canonical correlation. *Neuroimage*, 37, 4, 1250-1259. 1053-8119.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L. & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex, *Science*, 293, 5539, 2425-2430. 0036-8075.

Haynes, J. & Rees, G. (2006). Decoding mental states from brain activity in humans, *Nature Reviews Neuroscience*, 7, 7, 523-534. 1471-0048.

Herskovits, E. H. & Gerring, J. P. (2003). Application of data-mining method based on Bayesian networks to lesion-deficit analysis, *Neuroimage*, 19, 4 ,1664-73. 1053-8119.

Iezzoni, L. I. (2004). Risk adjusting rehabilitation outcomes, an overview of methodologic issues, *Am J Phys Med Rehabil*, 83, 316-326. 0894-9115.

Innocent, P. R., Barnes, M. & John, R. (1997). Application of the fuzzy ART/MAP and MinMax/MAP neural network models to radiographic image classification, *Artificial Intelligence in Medicine*, 11, 241-263. 0933-3657.

Jennet, B. & Bond, B. (1975). Assessment of outcome after severe brain damage: a practical scale, *Lancet*, 1, 480-484. 0140-6736.

Ji, S., Smith, R., Huynh, T. & Najarian, K. (2009). A comparative analysis of multi-level computer-assisted decision making system for traumatic brain injuries, *BMC Medical Informatics and Decision Making*. 9, 2. 1472-6947.

Karkanis, S., Magoulas, G. D., Grigoriadou, M. & Schurr, M. (1999a). Detecting abnormalities in colonoscopic images by textural description and neural networks, *Proceedings of Workshop on Machine Learning in Medical Applications, Advanced Course in Artificial Intelligence-ACAI99*, Chania, Greece, 59-62.

Karkanis, S., Galoussi, K. & Maroulis, D. (1999b). Classification of endoscopic images based on texture spectrum, *Proceedings of Workshop on Machine Learning in Medical Applications, Advanced Course in Artificial Intelligence-ACAI99*, Chania, Greece, 63-69.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. (2009). Circular analysis in systems neuroscience – the dangers of double dipping, *Nature Neuroscience*, 12, 535-540. 1097-6256.

Ku, S., Gretton, A., Macke, J. & Logothesis, N. K. (2008). Comparison of pattern recognition methods in classifying high-resolution bold signals obtained at high magnetic field in monkeys. *Magnetic Resonance Imaging*, 26, 7, 1007-1014. 0730-725X.

Liao, C., Xiao, F., Wong, J. & Chiang, I. (2007). A knowledge discovery approach to diagnosing intracranial haematomas on brain CT: recognition, measurement and classification, *Proceedings of the 1st international conference on Medical biometrics*, 73-82. 0302-9743.

Lu, D., Street, W. N. & Delaney, C. (2006). Knowledge discovery: detecting elderly patients with impaired mobility, *Stud Health Technol Inform*, 122, 121-3. 0926-9630.

Maas, A. I., Hukkelhoven, C. W., Marshall, L. F. & Steyeberg, E. W. (2005). Prediction of outcome in traumatic brain injury with computed tomographic characteristics: a comparison between the computed tomographic classification and combinations of computed tomographic predictors. *Neurosurgey*, 57, 1173-1182. 0148-396X.

Madigan, E. A. & Curet, O. L. (2006). A data mining approach in home healthcare: outcomes and service use, *BMC Health Services Research*, 6, 18. 1472-6963.

Magoulas, G. D. & Prentza, A. (2001). Machine Learning in medical applications. *Machine Learning and Its Applications, Lecture Notes in Computer Science*. 2049, 300-307. 3-540-42490-3.

Marshall, L. F., Marshall, S. B., Klauber, M. R., Van Brukum, C. M., Eisemberg, H., Jane, J. A., Luerssen, T. G., Marmarou, A. & Foulkes, M. A. (1992). The diagnosis of head injury requires a classification based on computed axial tomography, *Journal of Neurotrauma*, 9, Suppl. 1, 287-292. 0897-7151.

Murray, G. D., Butcher, I., McHugh, G. S., Lu, J., Mushkudiani, N. A., Maas, A. I., Marmarou, A. & Steyeberg, E. W. (2007). Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study, *Journal of Neurotrauma*, 24, 329-337. 0897-7151.

Nissen, J. J., Jones, P. A., Signorini, D. F., Murray, L. S., Teasdale, G. M. & Miller, J. D. (1999). Glasgow head injury outcome prediction program: an independent assessment, *Journal of Neurology, Neurosurgery, and Psychiatry*, 67, 769-799. 0022-3050.

Phee, S. J., Ng, W. S., Chen, I. M., Seow-Choen, F. & Davis, B. L. (1998). Automation of colonoscopy part II: visual-control aspects, *IEEE Engineering in Medicine and Biology,* May-June, 81-88. 0739-5175.

Pignolo, L., Riganello, F., Candelieri, A. & Lagani, V. (2009). Vegetative State: Early Prediction of Clinical Outcome by Artificial Neural Network, *Proceedings of The Fifth International Workshop on Artificial Neural Networks and Intelligent Information Processing (ANNIIP 2009), In conjunction with Sixth International Conference on Informatics in Control, Automation and Robotics (ICINCO 2009),* 91-96. 978-989-674-002-3.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A. & Shulman, G. L. (2001). A default mode of brain function, *Proc Natl Acad Sci USA*, 98, 2, 676-682. 1091-6490.

Raichle, M. E. & Snyder, A. Z. (2007). A default mode of brain function: a brief history of an evolving idea, *Neuroimage*, 37, 4, 1083-1090. 1053-8119.

Riganello, F., Quintieri, M. Candelieri, A., Conforti, D. & Dolce, G. (2008). Heart rate responses to music: an artificial intelligence study on healthy and traumatic brain-injured subjects, *Journal of Psychophysiology*, 22, 4, 166-174. 0269-8803.

Riganello, F., Lagani, V., Pignolo, L. & Candelieri, A. (2009). Data-mining approaches for the study of emotional responses in healthy controls and traumatic brain injured patients: comparative analysis and validation, *Proceedings of The Fifth International Workshop on Artificial Neural Networks and Intelligent Information Processing (ANNIIP 2009), In conjunction with Sixth International Conference on Informatics in Control, Automation and Robotics (ICINCO 2009)*, 125-133. 978-989-674-002-3.

Riganello, F. & Candelieri, A. (2010). Data Mining and the functional relationship between heart rate variability and emotional processing: comparative analyses, validation and application, *Proceedings of The Third International Conference on Health Informatics (HEALTHINF 2010)*, 159-165. 978-989-674-016-0.

Riganello, F., Candelieri, A., Quintieri, M. & Dolce, G. (2010a). Heart rate variability: an index of brain processing in vegetative state? An artificial intelligence, data mining study, *Clin Neurophysiol*, (in press). 1388-2457.

Riganello, F., Candelieri, A., Dolce, G. & Sannita, W. G. (2010b). Residual emotional processing in the vegetative state: a scientific issue?, *Clin Neurophysiol,* (in press). 1388-2457. (doi 10.1016/j.clinph.2010.09.006).

Saatman, K. E., Duhaime, A., Bullock, R., Maas, A. I. R., Valadka, A., Manley, G. T. & Workshop Scientific Team and Advisory Panel Members. (2008). Classification of traumatic brain injury for targeted therapies, *Journal of Neurotrauma*, 25, 7, 719-738. 0897-7151.

Schummer, M., Green, A., Beatty, J. D., Karlan, B. Y., Karlan, S., Gross, J., Thornton, S., McIntosh, M. & Urban, N. (2010). Comparison of breast cancer to healthy control tissue discovers novel markers with potential for prognosis and early detection. *PLoS ONE*, 5, 2, e9122. 1932-6203.

Teasdale, G. & Bennet, B. (1974). Assessment of coma and impaired consciousness. A practical scale, *Lancet*, 2, 443-448. 0140-6736.

Veropoulos, K., Campbell, C. & Learmonth, G. (1998). Image processing and neural computing used in the diagnosis of tuberculosis. Colloquium on Intelligent Methods in Healthcare and Medical Applications, York, UK.

Yin, H., Li, G., Leong, T. Y., Kuralmani, V., Pang, H., Ang, B. T., Lee, K. K. & Ng, I. (2006). Experimental Analysis on Severe Head Injury Outcome Prediction—A Preliminary Study, Technical Report TRD9/06, School of Computing, National University of Singapore.

Zhu, Y. & Yan, H. (1997). Computerized tumor boundary detection using a Hopfield neural network, *IEEE Transactions on Medical Imaging*, 16, 55-67. 0278-0062.

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:

Antonio Candelieri, Giuliano Dolce, Francesco Riganello and Walter G Sannita (2011). Data Mining in Neurology, 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-neurology

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

IntechOpen