Chapter 37

Data Mining in Psychiatric Research

Diego Tovar, Eduardo Cornejo, Petros Xanthopoulos, Mario R. Guarracino, and Panos M. Pardalos

Abstract

Mathematical sciences and computational methods have found new applications in fields like medicine over the last few decades. Modern data acquisition and data analysis protocols have been of great assistance to medical researchers and clinical scientists. Especially in psychiatry, technology and science have made new computational methods available to assist the development of predictive modeling and to identify diseases more accurately. Data mining (or knowledge discovery) aims to extract information from large datasets and solve challenging tasks, like patient assessment, early mental disease diagnosis, and drug efficacy assessment. Accurate and fast data analysis methods are very important, especially when dealing with severe psychiatric diseases like schizophrenia. In this paper, we focus on computational methods related to data analysis and more specifically to data mining. Then, we discuss some related research in the field of psychiatry.

Key words: Data mining, Machine learning, Psychiatry, Drug efficacy, Schizophrenia

1. Introduction

Data mining has been applied increasingly in the field of medicine over the last few years. Generally speaking, data mining can be described as the area of applied mathematics that tries to extract information from large datasets, often stored in huge computer databases. The methodological protocols (i.e., algorithms) are "born" from the interplay of different disciplines, such as statistics, artificial intelligence, and optimization. In the recent years, although both the number of challenges and algorithms that are specifically related to biomedical/clinical problems (1–4) have been increasing, data mining is still used for a broad range of applications that cover a large spectrum of science and engineering areas, like economics, machine vision, agriculture (5), etc. However,

Firas H. Kobeissy (ed.), *Psychiatric Disorders: Methods and Protocols*, Methods in Molecular Biology, vol. 829,
DOI 10.1007/978-1-61779-458-2_37,

novel data acquisition methods and new technological advances generate new computational challenges and problems.

Psychiatry is a field of medicine that specializes in studying and curing mental disorders. As every other branch of medicine, psychiatry includes patient monitoring, animal studies, and in vivo and in vitro research studies, which always generate large amounts of data. The data are usually collected in the form of questionnaires, biometric data, or even microarray data matrices. After acquisition, datasets are stored in computer databases also known as *data warehouses, where data is extracted, transformed, loaded, and aggregated*. Most data acquisition methods (especially these that deal with microarray and sequencing technologies) result in the accumulation of vast amounts of data. In these cases, by just making empirical observations, even the most experienced clinical scientist fails to analyze them properly and draw safe conclusions. At this point, mathematical modeling and data mining should be employed in order to assist with these complicated tasks. In this paper, we explain the most representative data mining algorithms, and give examples of successful data mining projects applied in psychiatry research.

2. Methods

Data mining is a very general term, and covers a large number of methods that originate from different branches of statistics and computer science. Some of them have empirical and some other solid mathematical foundations. In any case, these algorithms are usually judged by the accuracy of the results they produce, as well as their ability to assist the experienced clinical scientists. Next, we discuss some of the best-studied and applied data mining methodologies grouped in categories.

1. Data preprocessing includes all algorithms responsible for data preparation. Time-series filtering, outlier detection, data cleaning algorithms, and data normalization algorithms fall in this category. Proper data preprocessing is essential for a more efficient performance of learning algorithms.
2. Machine learning (ML): Learning from data is the most important part of data mining. Machine learning is a set of algorithms that have a dataset as input and also may be some information about it. The output of ML is a set of rules that let us make inference about any new data point.
3. Unsupervised learning (UL), sometimes also known as clustering, aims to find associations between data points (clusters). Clustering is usually performed when no information is given about the structure of the dataset. It can be used for

Fig. 1. General supervised learning process. The data are used in order to train an algorithm, for which some parameters have been selected first (through a model selection algorithm). Then, the trained model is used in data with unknown labels.

exploratory purposes (e.g., identify specific data structure that can be used for more efficient supervised algorithm design). For a more extensive tour to data clustering, we refer the reader to (6).

4. Supervised learning (SL) is one of the most well-known data mining algorithms. SL algorithms are given a set of data points (data samples) with known properties (features) and the classes they belong (labels). Then, the SL algorithm trains a model which at the end of the training phase is capable of deciding on the identity of new data points with unknown labels (test dataset). In this category, one can include the artificial neural networks, Bayesian classifiers, k-nearest neighbor classification, genetic algorithms, and others (7). If the samples contain qualitative feature values, then the rule-based classification can be employed (8, 9). Especially for the two-class general case binary classification, one of the most commonly used approaches is Support Vector Machine (SVM) (see Note 1). Originally proposed by Vapnik (10), SVM aims to determine a separation hyperplane from the training dataset. SVM possesses a solid mathematical foundation in optimization theory. A scheme summarizing the general supervised learning idea is shown in Fig. 1. If the two classes of data cannot be discriminated with a linear hyperplane, then the problem can be addressed as a nonlinear classification problem. Such problems can be attacked using the so-called *kernel trick*. In this case, original datasets are embedded in higher dimension spaces, where perfect linear separation can be achieved (11). The combined use of supervised classification methods with kernels is the most common way to address data mining problems. Finally, in order to use these packages, the user must possess some software programming skills, i.e., MATLAB (see Note 2).
5. Semi-supervised learning lies in between supervised and unsupervised learning. In this case, class labels are known only for a portion of available data and some partial information is given to the algorithm usually in the form of pairwise constraints (e.g., points *a* and *b* belong/do not belong to the same class). The goal in this case is to achieve optimal utilization of this information in order to obtain the highest predictive accuracy.

6. Biclustering tries to find associate groups of features with corresponding groups of samples. In this way, one can decide on the most important features that are responsible for a group of samples with specific characteristics. It has been extensively used in microarray data analysis for associating genes with specific phenotypes. There are numerous algorithmic approaches to biclustering, ranging from greedy algorithms, spectral biclustering, column reordering, and 0–1 fractional programming. For a more mathematically rigorous review about biclustering and their applications, we refer the reader to these references (12, 13). It is worth mentioning that biclustering can have either a supervised or an unsupervised version.
7. Feature selection consists of determining the most important properties (features) of each data point (sample) that are used for training. For example, given a set of people (i.e., sample) and some of their features like weight, height, eye color, and hair color, we wish to distinguish the infants from the adults. A feature selection algorithm tries to select a small subset of features that have the largest combined discriminatory power (in this case, may be weight and height). In case of problems described by hundreds or thousands of characteristics, feature selection permits to reduce the number of variables, with a great advantage in terms of computational time needed to obtain the solution. Examples of algorithms for feature selection include the "RFE" and "Relief" described in reference 14. Here, we need to point out the difference between feature selection and feature extraction. Feature extraction consists of the construction of some features that do not necessarily belong to the set of the original features. The significant reduction of the original number of features, due to a feature selection or feature extraction algorithm is called *dimensionality reduction*, which is essential in order to improve the processing time. A standard method for feature extraction is the principal component analysis as shown in Fig. 2.

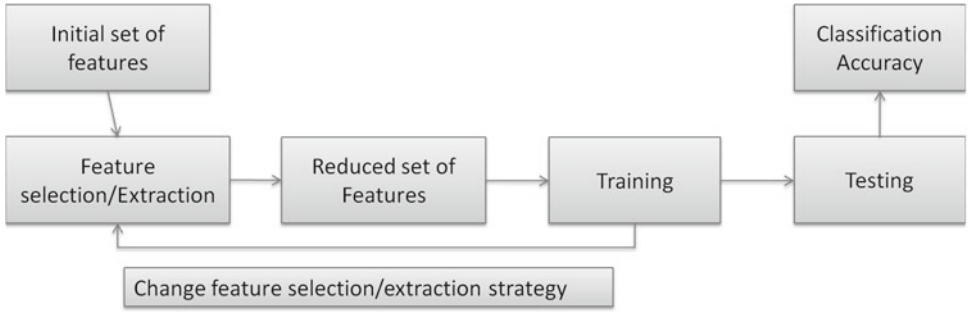

Fig. 2. A model of change feature selection/extraction strategy. The initial set of features is fed to a feature selection/extraction algorithm. The reduced set of features is used for learning, and based on the performance of the training phase one may reexamine the feature selection/extraction process.

8. Data visualization/representation: This last branch of data mining deals with methods, where the extracted information can be represented to/visualized by the end user. Massive datasets need special types of algorithms for analysis and representation (15). One common way to represent information and potential relationship between entities is through graphs (also referred to as networks). Network representation can be very useful in understanding the dynamics that govern a system. Software packages, like "Cytoscape software" (16), deal with representation of complex biological data. Two other topics related to data mining and more specifically to the supervised machine learning component of the process are the *model selection* and the *cross validation* of the learning scheme (see Note 3). Model selection is related to the "tuning" of the model itself in terms of the parameter selection. The most common method employed in this case is a uniform search over a grid spanned by the parameters. Lately, there have been proposed some more advanced model selection methods based on uniform design theory (17).

Taken together, cross validation of a model is very important, especially when one wants to statistically assure the independence of the output model accuracy from the specific properties of training and testing datasets. For this, the given data are partitioned many times in testing and training datasets using some resampling approach. The reported accuracy of the method is the average accuracy of all of the cross-validation repetitions. Cross-validation methods usually vary by their strategy for partitioning the original data. Most of the well-studied cross-validation techniques include k-fold, leave-one-out, and hold-out cross validation.

All the described methods are implemented in many open-source problem-solving environments, like *R* (http://www.r-project.org), (http://www.cs.waikato.ac.nz/ml/weka), *Octave* (http://www.gnu.org/software/octave/), and *Weka* (http://www.cs.waikato.ac.nz/ml/weka) (see Note 4). They provide simple graphical user interfaces that can also be used by nonexpert users.

3. Methods and Applications of Data Mining

3.1. Applications in Psychiatry

The aforementioned technology of data mining and its application are applied in psychiatric research. One of the most important problems in psychiatry is to diagnose and/or assess the disease accurately. Another important aspect is the running time of the computational algorithm. Ultimately, clinical society desires noninvasive data acquisition methods in combination with

accurate analysis protocols that can run in a timely manner in a not-so-demanding hardware configuration. Thus, we need to emphasize that the purpose of this paper is to give an overview with some illustrative applications of the field rather than trying to cover, and include, the full spectrum.

3.2. Data Mining in MRI Medical Imaging

One of the primary sources of data used by expert clinician is medical imaging. Magnetic resonance imaging (MRI) is one of the most widely used medical imaging techniques. In one study, Arnborg et al. used MRI scans taken from control subjects and patients with schizophrenia (18). The MRI scans are turned into 3D images and processed by the *BRAINS* software. The sample size included 144 subjects, 63 suffering from schizophrenia and 81 controls. Analysis of the different scans is performed through Bayesian modeling. The outputs of this analysis are networks whose nodes are connected by a variable covariate. The long-term goal of the study is to provide a modeling network to identify the underlying mechanism, which causes a mental disease. The networks produced from the physiological data measurements in the form of MRI correlate different parts of the brain. Bayesian modeling has been employed even in more recent studies in order to model, understand, and evaluate the nature of schizophrenia (19, 20).

3.3. Voice Recognition-Assisted Diagnosis

Data mining can be utilized for mental disease diagnosis using noninvasive data. For example, Diederich et al. used an SVM approach in order to distinguish between schizophrenia patients and controls (21). The data used for this classification task are voice samples from patients and control. Subjects are given a story and a group of semantic words that they had to use, and then they were asked to tell a story about it. This was done in order to constrain the number of different words subjects used. SVM achieved 77% accuracy and the systems performance was increased when only the 100 most frequently used words were used. Authors suggest fusion of the proposed speech recognition method with other datasets that would incorporate facial movement, medical imaging (e.g., MRI), and biological data (e.g., microarrays).

3.4. Psychoactive Drug Efficacy Assessment

Data mining has been also applied in the area of drug effect assessment. In one study, Kafkafi et al. uses a data mining algorithm in order to assess the efficacy of three kinds of psychopharmacological drugs: psychomotor stimulant, opioid, and psychotomimetic (22, 23). Animals were dosed with these drugs and then their movements were recorded. Based on the path they follow, several features were extracted which were then given into the supervised learning algorithm. The data mining algorithm ("Pattern Arrays"—PA) is used in order to determine the optimal predictors for each type of drug. The hypothesis is that the different types of psychoactive drugs produce different behavioral profiles that can be captured and

analyzed by movement. In general, drug efficacy assessment is a crucial and complicated task. The discovery of a therapeutic agent and the understanding of the whole underlying mechanism is a challenging optimization problem. For a more comprehensive review on this topic, we refer the reader to reference 24.

3.5. Data Mining and Network Analysis in EEG

Another common noninvasive and low-cost technique for acquiring useful data for diagnosis is the electroencephalogram (EEG). EEG recordings are scalp recordings of the brain electric potential that is produced as a result of brain neural oscillations. Quantitative EEG analysis is a field with many open computational problems (2, 25, 26). One thoughtful way for representing EEG data is by using a network (or graph) representation. Usually, nodes of such networks are electrode sites of the brain, whereas the edges correspond to some kind of generalized similarity measure. Theoretical foundations of such similarity across measures can be found in linear or nonlinear time-series analysis literature. The dynamic alterations of such networks that are induced due to some pathology or as a result of some treatment effect can be indicative of the complicated underlying mechanism. The changes between networks of different classes can be measured by graphing theoretic quantitative measures. Network modeling and analysis have also been used for other pathological conditions that affect brain function, like epilepsy or Alzheimer's disease.

In another study, Sakkalis et al. used wavelet coherence as a measure of synchrony and compared the networks between schizophrenia and control subjects. They define the quantitative measures used for capturing the network's structure as the average degree of the graph, the clustering coefficient, and the average shortest path length (27). Preliminary results indicate that all three measures take lower values for patients with schizophrenia. The software tool developed for this study is described in 28.

Another area of quantitative EEG analysis, where data mining and mathematical theory of optimization have been successfully applied, is epilepsy research. Traditionally, experienced, board-certified electroencephalographers manually examine long-term EEG recordings and asses the severity of the recorded epileptogenic activity. This is a slow, expensive, and tedious process. Manual scoring is always subject to human errors that are related to experience and fatigue of the clinician. Sometimes, quantitative patterns might be very well hidden, making it impossible even to the most experienced observer to mine them. Automated analysis and quantification of patterns and trends that are "hidden" in EEG recordings is a very important problem in quantitative EEG analysis. Epilepsy researchers, especially, want to find features that precede the occurrence of seizures and thus propose efficient predictive algorithmic schemes. The ultimate goal is to provide a full system able to predict and anticipate seizure activity.

Several researchers have employed nonlinear time-series analysis and deterministic chaos-based approaches in order to establish biomarkers to predict the evolution of an epileptic seizure. Some of the most notable approaches toward epileptic seizure prediction algorithm have been demonstrated by Iasemidis et al. who employed short-term Lyapunov exponents and optimization process in order to propose the optimal EEG-based biomarker for this problem (29). Other approaches include correlation dimension (30), phase synchronization (31), similarity index (32), and other methods. For more extensive review of epileptic seizure prediction literature, we refer the reader to references 33 and 34.

3.6. Information Extraction Through Tree Mining

Medical record data mining represents another area for research. Medical record data is represented in the form of a decision tree. The trees that belong to subjects of the same category are mined to determine the common patterns among them.

In their work, Hadzic et al. employed XML modeling and tree mining in order to extract patterns that are able to provide useful information about mentally ill patients (35). In another study, the IMB3-Miner algorithm has been used (36); this analysis includes subtrees (smaller trees that are contained in the original trees) characterizing each class specifically. The XML database modeling in conjunction with data mining can provide future online tools for identifying the cause and factors that are mostly related to disease progression, which can yield better early-stage diagnosis and better treatment planning. The same authors in 37, 38 introduce an online system called *thinking PubMed*, whose goal is to properly extract and represent information stored in large databases like *PubMed* related to mental diseases.

4. Conclusion

In this chapter, we give a brief overview of data mining with discussion of some of its most prominent applications in psychiatric research. It is really very interesting to observe the development of techniques that combine traditional medical practice and state-of-the-art machine learning techniques. This leads to the development of automatic or semiautomatic software and systems able to assist and support clinician's profession, by extracting and representing knowledge in a user-friendly way, for an accurate and efficient diagnosis. Nevertheless, despite the progress and the success stories of many researchers applying computational tools in the medical field, there is still a gap between mathematics and medicine. The most important burden, which arises in every multidisciplinary collaborative research, is to overcome the technical language barrier between researchers and to be able to communicate

mathematical ideas to medical professionals easier and vice versa. It is our belief that computational and clinical science should go hand in hand for the mutual benefit of both.

5. Notes

1. All the methods described in this chapter are usually implemented in practice through software programming languages. For SVM classification software, there are several open-source implementations like libSVM or SVMlite. In order to use these packages, the user must possess some software programming skills, i.e., MATLAB.
2. One of the most complete data mining software suite for Matlab is the open-source toolbox Matlab Arsenal. This toolbox includes a large number of functions related to data clustering; feature extraction, and feature selection. The installation and use are straightforward, requiring basic Matlab programming knowledge.
3. Graph drawing and visualization tools are becoming more and more popular every day as there are needs for massive graph representation. The mentioned open-source package Cytoscape (that is mentioned in the chapter) is one of them. It is worth noting that this package is not only used for graph drawing, but also allows for some pretty sophisticated data analysis methods. Thanks to the developer's platform that comes with it, programmers can develop and add data analysis protocols written in Java and add them as plug-ins to the main program.
4. Another data analysis mentioned in the text is Weka. This is another open-source data mining toolbox developed from University of Waikato. It is written in Java and it offers a number of functions related to data mining. Thus, it can be used by programmers in order to add Weka functionality in for their application. For users less experienced with programming, Weka offers a user interface named "Weka Knowledge Explorer" that offers a large variety of windowed user-friendly options for data analysis and plotting.

References

1. Alves, C. J. S., Pardalos, P. M., and Vicente, L. N. (2008) *Optimization in medicine*, 1st ed., Springer, New York.
2. Chaovalitwongse, W. A., Pardalos, P. M., and Xanthopoulos, P., (Eds.) (2010) *Computational Neuroscience*, Vol. 38, Springer, New York.
3. Pardalos, P. M., and Romeijn, H. E. (2009) *Handbook of Optimization in Medicine*, Springer, New York.
4. Seref, O., Kundakcioglu, O. E., and Pardalos, P. M. (2007) *Data mining, systems analysis, and optimization in biomedicine : Gainesville,*

Florida, U.S.A., 28–30 March 2007, American Institute of Physics, Melville, N.Y.

5. Mucherino, A., Papajorgji, P., and Pardalos, P. M. (2009) *Data Mining in Agriculture*, Springer.
6. Jain, A. K., and Dubes, R. C. (1988) *Algorithms for clustering data*, Prentice Hall.
7. Bishop, C. M. (2006) *Pattern recognition and machine learning*, Springer, New York.
8. Quinlan, J. R. (1992) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.
9. Quinlan, J. R. (1996) Improved use of continuous attributes in C4.5, *J Artif Intell Res 4*, 77–90.
10. Vapnik, V. N. (1995) *The nature of statistical learning theory*, Springer, New York.
11. Shawe-Taylor, J., and Cristianini, N. (2004) *Kernel methods for pattern analysis*, Cambridge University Press, Cambridge, UK ; New York.
12. Busygin, S., Prokopyev, O., and Pardalos, P. M. (2008) Biclustering in data mining, *Comput Oper Res 35*, 2964–2987.
13. Xanthopoulos, P., Boyko, N., Fan, N., and Pardalos, P. M. (2010) Biclustering: algorithms and applications in data mining and forecasting, in *Encyclopedia of Operations Research and Management Science* (Wiley, Ed.), p to apprear.
14. Guyon, I., and Elisseeff, A. (2003) An introduction to variable and feature selection, *Journal of Machine Learning Research 3*, 1157–1182.
15. Abello, J., Pardalos, P. M., and Resende, M. (2002) *Handbook of massive datasets*, Kluwer Academic Publisher, Dordrecht, The Netherlands.
16. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome Res 13*, 2498–2504.
17. Huang, C. M., Lee, Y. J., Lin, D. K. J., and Huang, S. Y. (2007) Model selection for support vector machines via uniform design, *Comput Stat Data An 52*, 335–346.
18. Arnborg, S., Agartz, I., Hall, H., Jönsson, E., Sillén, A., and Sedvall, G. (2002) Data mining in schizophrenia research – preliminary analysis., in *Principles of Data Mining and Knowledge Discovery* (Elomaa, T., Mannila, H., and Toivonen, H., Eds.), pp 27–38, Springer.
19. Hall, H., Lawyer, G., Sillen, A., Jonsson, E. G., Agartz, I., Terenius, L., and Arnborg, S. (2007) Potential genetic variants in schizophrenia: A Bayesian analysis, *World J Biol Psychia 8*, 12–22.
20. Lawyer, G., Nyman, H., Agartz, I., Arnborg, S., Jönsson, E. G., Sedvall, G. C., and Hall, H. (2006) Morphological correlates to cognitive dysfunction in schizophrenia as studied with Bayesian regression, *BMC psychiatry 6:31*.
21. Diederich, J., Al-Ajmi, A., and Yellowlees, P. (2007) E-x-ray: Data mining and mental health, *Appl Soft Comput 7*, 923–928.
22. Elmer, G. I., and Kafkafi, N. (2009) Drug Discovery in Psychiatric Illness: Mining for Gold, *Schizophrenia Bull 35*, 287–292.
23. Kafkafi, N., Yekutieli, D., and Elmer, G. I. (2009) A Data Mining Approach to In Vivo Classification of Psychopharmacological Drugs, *Neuropsychopharmacol 34*, 607–623.
24. Enna, S. J., and Williams, M. (2009) Challenges in the Search for Drugs to Treat Central Nervous System Disorders, *J Pharmacol Exp Ther 329*, 404–411.
25. Pardalos, P. M. (2004) *Quantitative neuroscience : models, algorithms, diagnostics, and therapeutic applications*, Kluwer Academic, Boston.
26. Pardalos, P. M., and Príncipe, J. C. (2002) *Biocomputing*, Kluwer Academic, Dordrecht ; Boston, Mass.
27. Sakkalis, V., Oikonomou, T., Pachou, E., Tollis, I., Micheloyannis, S., and Zervakis, M. (2006) Time-significant wavelet coherence for the evaluation of schizophrenic brain activity using a graph theory approach, in *Proceedings of 28th Annual International Conference of IEEE EMBS, New York, NY.*, pp 4265–4268.
28. Oikonomou, T., Sakkalis, V., Tollis, I., and Micheloyannis, S. (2006) Searching and visualizing brain networks in schizophrenia, in *Biological and Medical Data Analysis* (Maglaveras, N. a. C., Ioanna and Koutkias, Vassilis and Brause, Rüdiger, Ed.), pp 172–182, Springer.
29. Iasemidis, L. D., Shiau, D. S., Pardalos, P. M., Chaovalitwongse, W., Narayanan, K., Prasad, A., Tsakalis, K., Carney, P. R., and Sackellares, J. C. (2005) Long-term prospective on-line real-time seizure prediction, *Clin Neurophysiol 116*, 532–544.
30. Lehnertz, K., and Elger, C. E. (1998) Can epileptic seizures be predicted? Evidence from nonlinear time series analysis of brain electrical activity, *Phys Rev Lett 80*, 5019–5023.
31. Mormann, F., Lehnertz, K., David, P., and Elger, C. E. (2000) Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients, *Physica D 144*, 358–369.
32. Le Van Quyen, M., Martinerie, J., Baulac, M., and Varela, F. (1999) Anticipating epileptic seizure in real time by a nonlinear analysis of

similarity between EEG recordings, *Neuroreport 10*, 2149–2155.

33. Mormann, F., Andrzejak, R. G., Elger, C. E., and Lehnertz, K. (2007) Seizure prediction: the long and winding road, *Brain 130*, 314–333.
34. Sackellares, J. C. (2008) Seizure prediction, *Epilepsy Currents 8*, 55–59.
35. Hadzic, M., Hadzic, F., and Dillon, T. (2008) Tree Mining in mental health domain, in *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pp 230–230.
36. Tan, H., Dillon, T. S., Hadzic, F., Feng, L., and Chang, E. (2005) MB3-Miner: mining eMBedded subTREEs using tree model guided candidate generation, in *Proceedings of the 1st International Workshop on Mining Complex Data 2005 in conjunction with ICDM 2005* pp 103– 110, Houston, TX.
37. Hadzic, M., D'Souza, R., Hadzic, F., and Dillon, T. (2008) Synergy of Ontology and Data Mining: Increasing Value of the Mental Health Information within PubMed database, in *Proceedings of the Second IEEE International Digital Ecosystems and Technology Conference*, pp 600–603.
38. Hadzic , M., D'Souza , R., Hadzic , F., and Dillon, T. (2008) Thinking PubMed: an Innovative System for Mental Health Domain, in *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems.*