Knowledge Discovery for Geriatric Diseases

Case Study: Dementia in Egypt

Nevine M. Labib, Edward W. Morcos
Sadat Academy for Management Sciences
Arvid C. Johnson, Mohamed O. Askar
Brennan School of Business, Dominican University, River Forest, IL, USA
Ahmed K. Mortagui, Sarah A. Hamza
Ain Shams University

Author Note

Nevine M. Labib, Department of Computer and Information Systems, SAMS.

Edward W. Morcos, Department of Computer and Information Systems, SAMS.

Arvid C. Johnson, Brennan School of Business, Dominican University
Mohamed O. Askar, Brennan School of Business, Dominican University
Ahmed Kamel Mortagui, Faculty of Medicine, Ain Shams University.

Sarah Ahmed Hamza, Faculty of Medicine, Ain Shams University.

The researchers express their profound sincere gratitude to the medical staff of the Geriatrics Department at El-Demerdash Hospital in Cairo for their cooperation.

Correspondence concerning this article should be addressed to Nevine Labib, Department of Computer and Information Systems, Sadat Academy for Management Sciences, Maadi, Cairo, Egypt.

ABSTRACT

A major challenge in the medical domain is the extraction of comprehensible knowledge from patient data. In this study, a knowledge discovery approach is proposed to extract meaningful patterns that can be used in clinical practice for early detection and better management of Elderly diseases with a special focus on Dementia. Initially, 65 years old or older patients were interviewed, underwent medical and cognitive examination. A data mining tool, Microsoft SQL Server 2008 Business Intelligence Tool was applied on these patients' medical data, whereby 13 factors were investigated. This tool made use of two techniques, Decision Trees and Naïve Bayes. The results showed that this approach to data analysis provides valuable knowledge concerning the predisposing factors that may lead to Dementia.

Keywords: Knowledge Discovery, Data Mining, Medical Informatics, Elderly Diseases, Dementia, Alzheimer Disease

Knowledge Discovery for Geriatric Diseases

Case Study: Dementia in Egypt

1 Introduction

The automatic discovery of knowledge in databases (KDD) has attracted substantial attention over the last several years especially in the medical field. It may be defined as a process of identifying novel, valid and understandable patterns of data in order to make predictions or classifications about new data and summarize the contents of large databases so as to support decision-making (Babic, 1999) and is considered a bottom-up approach, since it starts with the data and tries to uncover new patterns either in directed or undirected way.
Directed knowledge discovery attempts to explain or categorize some particular data fields while undirected knowledge discovery attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. All of these activities fall under the field of data mining.

Medical data mining has a great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for the classification of various diseases. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data.

One of the recent applications is the domain of Geriatric diseases since it still needs to be explored so as to uncover the predisposing as well as the associated factors related to aging. The study will focus on Dementia, which is a common Disease for the Elderly people. The most common form is the Alzheimer’s disease.

The 2009 World Alzheimer Report, released by Alzheimer's Disease International, a nonprofit federation of 71 national Alzheimer organizations, estimates that the global prevalence of dementia, predicted to be more than 35 million in 2010, will almost double every 20 years to 65.7 million in 2030 and 115.4 million in 2050.

1.1 Scope and Objectives of the Study

The study focuses on the domain of Elderly Diseases namely Dementia, which is characterized by the permanent and progressive deterioration of intellectual function caused by changes in the brain that results in impaired thinking and altered behavioral impulses (Cowart, 2004). It has different subtypes, Alzheimer's Disease (AD), Vascular Dementia and others.

The main objective of the research is to make use of Data Mining techniques in order to identify patterns of interest in the Geriatric diseases data, especially Dementia. As for the detailed objectives, they are the following:

1. Investigating different Data Mining techniques to be applied in the Geriatric Diseases.
2. Comparison of various classification techniques and finding the best classification technique for the given data.
3. Finding the most influencing risk factors causing Dementia.

1.2 Significance of the Study

Dementia affects about 1% of people aged 60-64 years and as many as 30-50% of people older than 85 years. It is a very serious condition that results in significant financial and human costs. Therefore, it is considered a serious personal, medical and social problem. According to the results of the first population-based survey study conducted in Egypt about prevalence of Dementia, it was found that the prevalence ratios for Dementia subtypes were 2.2 for Alzheimer's disease, 0.95 for multi-infarcts dementia, 0.55 for mixed Dementia and 0.45 for secondary Dementia (extracted from the Egyptian Recommendations in the Early Diagnosis and Management of Alzheimer's Disease, 1998). Recent research indicates early and accurate diagnoses as the key to effectively cope with it. No definitive cure is available but in some cases when the impairment is still mild the diseases can be contained.

Therefore, the research has a great significance in the health care field as research findings should be useful to the society for the identification, prediction and treatment of patients since it will deal with the predisposing and precipitating factors of Dementia. Hence, it may provide cost savings and delay institutionalization. It will also enable the clinician the tailoring of specific treatment plan.

2 Related Work

Several studies were done in the domain of applying Data Mining (DM) for Geriatrics. Some of the recent and most relevant ones are the following:

1. A study dealt with the screening tools with KDD methods and extended these techniques to the differential diagnosis of Alzheimer's Disease, Vascular Dementia and other causes (Mani et al., 1997).
2. A research paper described a diagnostic tool that jointly uses the naïve credal classifier and the most widely used computerized system of cognitive tests in dementia research, the Cognitive Drug Research system, in order to deal with small and incomplete datasets. This diagnostic tool proved to be very effective in discriminating between Alzheimer's disease and dementia with Lewy bodies (Zaffalon et al., 2003).
3. A research study dealt with the identification of the different types of dementia using data mining techniques. Scalar, joint, histogram and voxel-level features were used. The results showed that cerebral metabolic rate of glucose consumption was efficient in the classification of dementia and data mining using voxel-level features with Principal Component Analysis and the logistic regression model method achieving the best classification (Wen et al., 2008).
4. Another study dealt with the relationships of 35 genetic and/or phenotypic factors, with incident cognitive decline and dementia, extracted from a large data base named 'the Conselice Study'. A new mathematical approach, called the Auto Contractive Map (AutoCM), was used and showed the differential importance of each variable (Licastro et al., 2010).
5. A novel data mining framework, in combination with three different classifiers including support vector machine, Bayes statitstics and voting feature intervals, was developed in order to derive a quantitative index of pattern matching for the prediction of the conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease using

MRI (Plant et al., 2010). The results showed that this framework reached a clinically relevant accuracy for the a priori prediction of the progression from MCI to AD.

6. A new model for the classification of AD, VD and Parkinson's disease is proposed. It deals with the selection of most influencing factors for both AD and PD using various attribute evaluation scheme with ranker search method (Joshi et al., 2010). Different models with the classification using various techniques such as Neural Networks and Machine Learning were developed. It was found that some specific genetic factors, diabetes, age and smoking were the strongest risk factors for Alzheimer's disease.
7. A study, we recorded fingertip pulse waves of elderly subjects, carried out chaos analysis on the plethysmogram data thus obtained, and examined their relationship with dementia (Oyama-Higa et al., 2005). It was found out that the Lyapunov exponent of the time series had a clear relationship with the severity of dementia and the communication skill of the elderly subjects.

3 Proposed Framework

In this study, the directed approach is applied. It starts by identifying the sources of pre-classified data, preparing these data for analysis then building and training a computer model, evaluating it and finally applying it to new data. This approach may be used in many applications and can accomplish many tasks such as *Classification*, *Clustering, Estimation and Prediction*. The proposed knowledge Discovery framework consists of the following stages:

3.1 Data Collection

The medical dataset was designed by Professors and specialists of Geriatrics. As for the patients' data, they were collected from patients' files extracted from Geriatrics section of El-Demerdash Teaching Hospital in Egypt. In addition, Structured Interviews with the domain Experts were conducted in order to assess the results of the cases.

3.2 Data Purifying

Some of the data fields, such as those related to the other diseases the patients are suffering from, were incomplete and inconsistent so we made some changes so as to turn them into binary fields with yes or no values instead of including the details. This stage was made on the fields related to the diseases the patient is suffering from such as cardiovascular, respiratory, gastro-intestinal tract (GIT), musculo-skeletal and neurological diseases.

3.3 Data Selection

Among an attributes space, each attribute may have certain relevance with another attribute, therefore to take the attributes relevant correlation into consideration in data pre-processing is a vital factor for the ideal pre-processing methods (Giannopoulou, 2008). Therefore, we chose only 45 attributes from the patients' files and mining techniques were be applied to these specific data items in order to reach the ones that represent an interest for the domain.

3.4 Data Mining

The tool used for this study was the *Microsoft SQL Server2008 Business Intelligence Tool* since it provides a full set of easy to use, graphical administration tools and wizards for creating, configuring and maintaining databases, data warehouses and data marts in addition to providing a scalable business Intelligence platform optimized for data integration, reporting, and analysis, enabling organizations to deliver intelligence where users want it.

3.5 Data Evaluation

Specific measures were used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre-specified interestingness constraints.

3.6 Knowledge Representation

In this step, visualization and knowledge representation techniques were used to present the mined knowledge to the user. All the operations applied on the records and fields, as well as the mining process itself, are represented in the form of visualizations and graphics as illustrated in the next section.

Based on the previous research detailed description, the proposed research was found to be both **quantitative** and **qualitative**.

4 Case Study

The proposed framework was applied on 120 representative cases of 65 years old or older patients. Following is a detailed description of the case studies used for framework testing.

4.1 Selected Data Attributes

The initial cases comprised a total of 45 attributes, 30 attributes were related to the Personal history of the patient while 15 attributes were related to cognitive mental tests results. When revised with the Geriatricians, it was found that some of the attributes could be combined together to give better indication. This filtering has resulted in using the following 13 attributes:

1. Factors related to personal history of the patient such as the Age of the patient, his gender and his level of education. The last attribute is divided into 4 categories 0-4, 5-8, 9-12 and >12 years.
2. Factors related to the other diseases the patient may be suffering from, such as Diabetes, Hypertension, cardiovascular disease, respiratory, GIT, musculo-skeletal and neurological diseases. All of these attributes are classified into two categories. It is either the patient is suffering from this disease or not
3. The score of the Geriatric Depression Scale (GDS), which ranges from 1 to 15. This scale was developed as a basic screening measure for depression in older adults (Shehta, 1983). This is directly related to the diagnosis of Dementia as it includes indirect symptoms of depression.
4. Mini-mental state examination (MMSE) score whose ranges depend on age and educational level (Crum et al., 1993).

As for the last attribute, it presents the diagnosis which is either Dementia or not. For all the input cases, 34% were diagnosed as Dementia, 66% as none Dementia.

The prediction model was done using two Data Mining models in this study. They are Naïve Bayes (NB) and Decision Trees (DT).

4.2 Extracted Decision Tree

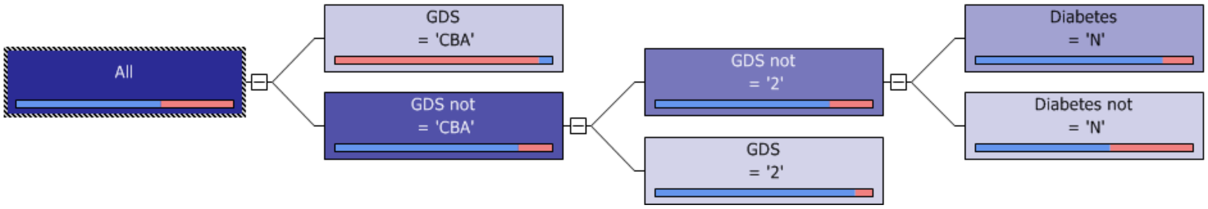

Fig. 1 Decision Tree Representing the Extracted Rules

Based on the extracted decision tree, the score of GDS was the most significant factor. When it is 'NTA' or '<2' it is associated with the presence of Dementia.

Since the Decision Trees technique did not provide significant results, we used Naïve Bayes.

4.3 Extracted Naïve Bayes

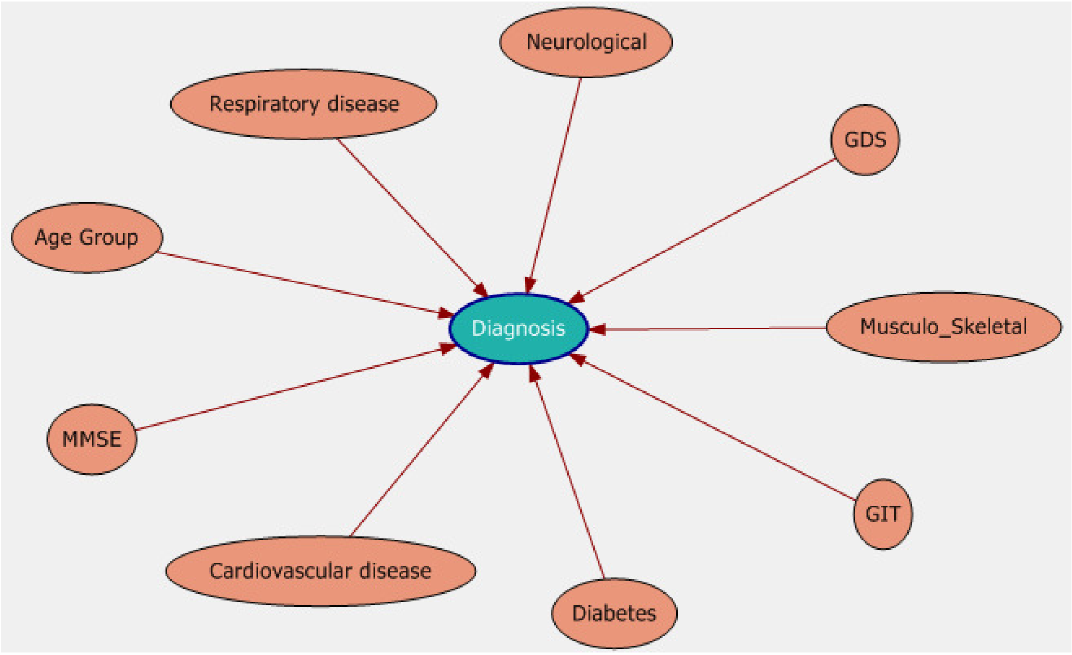

Figure 2 Dependency Network for Naive Bayes Model

This graph represents the relation between the predictable attribute (Diagnosis) and the other dependent attributes. It means that the Diagnosis is depending on all of the related factors (age group, GDS, MMSE, etc.).

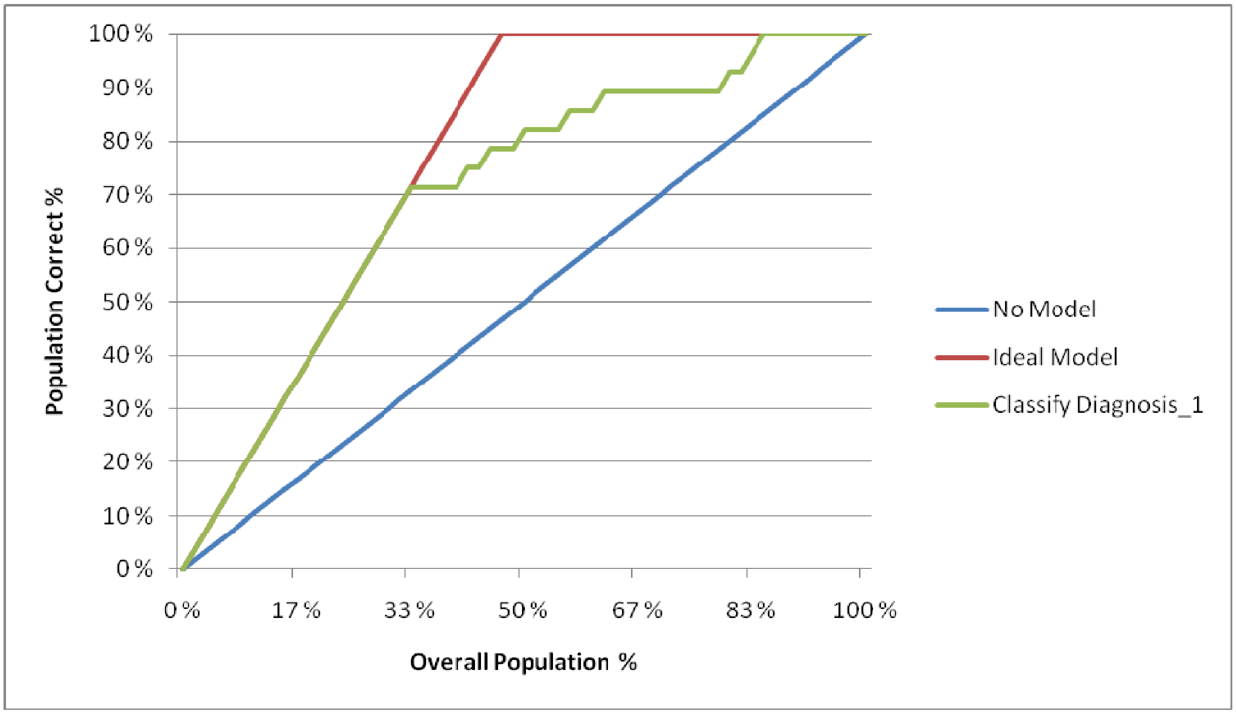

Figure 3 Accuracy Chart for Naïve Bayes Model

The previous figure represents the accuracy chart for Naïve Bayes model, the blue line represents the no model, the red line is the ideal model and the green line represents the Naïve Bayes model. Based on the graph it is conclude that the Naïve Bayes model is quite near the ideal model.

4.4 Medical Assessment of the Results

The previously mentioned results were revised by two Geriatricians. They found them acceptable although they had some comments related to the small number of cases.

Regarding the results of the DT, they indicate that the memory loss is either due to Depression or to Dementia. Since it is not Depression then these symptoms are due to Dementia. The results extracted based on the Naïve Bayes technique are more accurate as they indicate the significant factors related to the diagnosis of Dementia, which are age MMSE, cardiovascular disease, Diabetes, GIT, musculo-skeletal diseases, GDS, neurological and respiratory diseases. They added that the environmental and lifestyle factors are important as well since previous studies indicate they activities even at midlife are important for predicting later AD (Bendlin et al., 2010).

5 Conclusions and Future Work

5.1 Results and Conclusion

The first aim of this study was to provide a knowledge discovery framework for identifying predisposing factors related to Dementia. It is concluded that the proposed knowledge discovery framework was able to discover some hidden data relationships, patterns and attribute dependencies out of the medical datasets. It should provide valuable knowledge which supports the medical doctors in taking suitable decisions for the prediction and diagnosis of medical cases related to Dementia.

As for using SQL Server Business Intelligence tool, it is very convenient because of its compatibility with many DB formats and its ability to implement powerful decision trees.

Therefore, it is recommended to implement the proposed knowledge discovery framework in the field of Geriatrics. Moreover, using other data mining techniques or tools may lead to discover extra or complementary knowledge in the medical domain.

The proposed framework can also serve as a training tool to train medical students to diagnose Dementia.

As in any research, this work also has several limitations which are the following:

The main limitation concerns the data collected for developing the mining model.

The proposed framework is based on some selected attributes. This list may need to be expanded to provide a more comprehensive diagnosis tool.

Another limitation is that it only uses categorical data. For some disease diagnosis, the use of continuous data may be necessary.

The size of the dataset used in this research is still quite small. A large dataset would definitely give significant results.

A third limitation is that it only uses two data mining techniques. Additional data mining techniques, such as Artificial Neural Networks can be incorporated to provide better indicators.

5.2 Future Work

Based on previous conclusions and a number of issues that rose during the study, some topics can be considered as future opportunities to be explored by interested researchers. They are the following:

- Designing a comprehensive database related to Dementia with its three types that can be used in Hospitals in Egypt.
- Incorporating other medical attributes, than those used in the model, including image data extracted from EEG.
- Making use of other data mining techniques, e.g., Time Series, Clustering and Association Rules.
- Integrating techniques related to text mining and image mining with those related to data mining.
- Classifying the different subtypes of Dementia using data mining techniques.

The proposed model can be further enhanced and expanded to deal with other Elderly Diseases.

References

Babic, A. (1999). Knowledge discovery of advanced clinical data management and analysis. *Medical Informatics Europe*, 409-413.

Cowart M. E. (2004). Dementia in older adults. *Encyclopedia of Applied Psychology, (1),* 585-591.

The Egyptian recommendations in the early diagnosis and management of Alzheimer's Disease, Educational Grant by Pfizer, 1998.

Mani S. et al. (1997). Differential diagnosis of dementia: a knowledge discovery and data mining KDD) approach. Retrieved from http://Citeseex.ist.psu.edu.

Zaffalon M. et al. (2003). Reliable diagnoses of dementia by the naïve credal classifier inferred from incomplete cognitive data. Artificial *Intelligence in Medicine, 29(1-2)*, 61-79.

Wen L. et al. (2008). Classification of dementia from FDG-PET parametric images using data mining. *2008 IEEE International Symposium on Biomedical Imaging,* 412-415.

Licastro F. et al. (2010). Multivariable network associated with cognitive decline and dementia. *Neurobiology of Aging*, *31(2)*, 257-269.

Plant, C. et al. (2010). Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease., *Neuroimage, 50 (1),* 162-174.

Joshi S. et al. (2010). Classification of neurodegenerative disorders based on major risk factors employing machine learning techniques. *IACSIT International Journal Of Engineering and Technology, 2(4),* 350-355.

Oyama-Higa, M., Setogawa, M. and Miao, T. (2005). Distinction of patterns within time_series data using constellation graph. Retrieved from http://www.pubzone.org

Giannopoulou, E.G. (2008). Data mining in medical and biological research. *In-teh, Croatia.*

Development and validation of a geriatric depression screening scale: A preliminary report (1983). *Journal of Psychiatric Research (17),* 37-49. Arabic Version: Shehta, A.S. Prevalence of depression among egyptian geriatric community. *Ain Shams Univeristy: Geriatric Department Library,* 3-5.

Crum, R.M. et al. (1993). Population-based norms for the mini-mental state examination by age and educational level. *JAMA, (269)*, 2386-91.

Bendlin, B.B. et al. (2010). Midlife predictors of Alzheimer's disease. *Maturitas, (65)*, 131-137. Doi:10.1016/j.maturitas.2009.12.014