Data Mining for Breast Cancer Screening

Souad Demigha
CRI (Centre de Recherche en Informatique)
University of Paris1-Sorbonne
Paris, France

Abstract— **Data Mining refers to extracting or "mining" knowledge from large amounts of data. Data Mining concepts and techniques have been applied in many fields such as Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis and other Scientific Applications, in medicine but a little in radiology and imagery. We propose in this paper the application of Data Mining techniques in breast imaging for breast cancer screening.**

Keywords—data mining; knowledge; breast cancer; screening

I. INTRODUCTION

Data Mining refers to extracting or "mining" knowledge from large amounts of data [1]. It is defined as "the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules" [2].

Data Mining was also defined as the process of discovery of interesting, meaningful and actionable patterns hidden in large amounts of data.

In medicine and other fields of industry and science data mining techniques have been applied but a little in radiology and imagery.

In medicine Data Mining distinct from other fields due to nature of data: heterogeneous, with ethical, legal and social constraints.

The application of information mining techniques to the medical domain are helpful in extracting medical knowledge for diagnosis, decision-making, screening, monitoring, therapy support and patient management.

Data Mining provides many advantages for its application in medical domain (detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods) [3]. It also helps researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals,…[4].

In radiology and imagery, Data Mining techniques have been used just a little in the last decades but currently many research efforts have emerged [5], [6], [7], [8], [9], [10].

Image mining deals with the extraction of image patterns from a large collection of images [6]. Image mining deals with the extraction of implicit knowledge, that is, image data relationship or other patterns not explicitly stored in the images. The focus of image mining is on the extraction of patterns from a large collection of images [8].

Clinical data are stored in Hospital Information Systems (HIS), Radiological Information Systems (RIS) and Picture Archiving and Communication Systems (PACS) aimed to archive and retrieve patient records. These systems have strict operation performance requirements and provide a little support for ad hoc queries or data analysis [11].

For multimedia Data Mining, storage and retrieval techniques need to be integrated with standard Data Mining methods. Promising approaches include the construction of multimedia data cubes, the extraction of multiple features from multimedia data and similarity-based pattern matching.

The application of Data Mining in imagery allows to obtain additional knowledge about specific features of different classes and the way in which they are expressed in the image (can help to find some inherent non-evident links between classes and their imaging in the picture). It can help to get some nontrivial conclusions and predictions can be made on the base of image analysis [12].

We propose in this paper to develop a methodology for mining radiological knowledge (breast cancer knowledge). We consider both text and image data. Data Mining will convert large patient image and text datasets into useful information which improve patient diagnosis and follow-up and provides informative reports.

We develop a model based on data mining techniques for managing breast knowledge (text and images). This model will be able to facilitate diagnosis and structure patient records. It includes data information from history and physical examination, images (primary and logical features), context (conditions in which data are created: type of acquisition, imaging modality, processing, etc.), settings and connections to other hospital systems.

Section 2 describes different objectives of Data Mining in medicine [13].

II. OBJECTIVES OF DATA MINING IN MEDECINE

We can define the Data Mining objectives in medical field as follows:

978-1-4799-6600-4/15/$31.00 ©2015 IEEE

- Question based answers.
- Anomaly based discovery.
- New Knowledge discovery.
- Informed decisions.
- Probability measures.
- Predictive modeling.
- Decision support.
- Improved health.
- Personalized medicine.

Section 3 describes Data Mining functionalities.

III. Data Mining Functionnalities

Data Mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

A. Tasks

Data Mining tasks are used to specify the kind of patterns to be found in data mining tasks. Data Mining tasks are classified into 2 categories: *descriptive* and *predictive*. *Descriptive mining* tasks characterize the general properties of the data in the database. *Predictive mining* tasks perform inference on the current data in order to make predictions.

Fig.1. illustrates the different tasks of Data Mining.

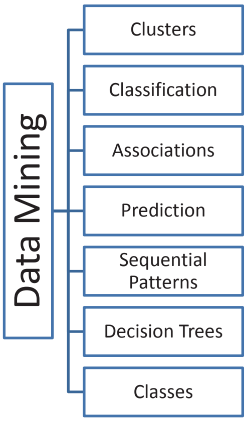

Fig.1. Types of Data Mining

1) Clusters: They consist of discovering groups and structures, without using known structures in the data. Clustering is the grouping of data into categories.
For breast cancer, data items may be grouped according to logical relationships or senologist (expert-radiologist in breast cancer) affinities or preferences.

2) Classification: Is concerned with the construction of classifiers that can be applied to “unseen" data so as to categorize that data into groups (classes). It is a supervised learning approach having known class categories.
For breast cancer, data of patients are classified according to the BI-RADS (Breast Imaging and Data Reporting System) standard into six categories. From BI-RADS0 (X-ray assessment is incomplete) to BI-RADS5 (Known biopsy-proven malignancy– Appropriate action should be taken).

3) Associations: Data can be mined to identify associations. It is used in finding patterns of association among the attributes or variables and observations.
For breast cancer, variables may represent demographics and observations may represent observations.

4) Prediction: Used in combination with the other data mining techniques, it involves analyzing trends, classification, pattern matching, and relation. By analyzing past events or instances.
For breast cancer, current health status, and previous clinical examination of the patient will predict her diagnosis and follow-up.

5) Sequential patterns: Data is mined to anticipate behavior patterns and trends.
For breast cancer, the phases of the radiologic process of the patient’s diagnosis (clinical examination, image reading, radiological intepretation,…).

6) Decision tree: Related to most of the other techniques (primarily classification and prediction), the decision tree can be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure, the decision tree enables to predict procedures and diagnoses in the domain of medicine.
For breast cancer, we organize the disease of the patient by age criterion. Patient who is older than 55 is more succptible to develop a cancer contrarily to a young patient who has a minor probability to develop a cancer.

7) Classes: Sored data is used to retrieve data in preplanned fixed groups.
For breast cancer, for example to ensure the quality of data providing from different modalities of image’s examination such as, mammographies and echographies. These data may be mined from a single group modality using ionizing radiation for diagnostic objectives.

“Classification” is the most frequently used data mining function with a predominance of the implementation of Bayesian classifiers, neural networks, and SVMs (Support Vector Machines).

B. Data Mining Techniques

Data Mining techniques result in four categories [14]:

1) Artificial Neural Networks: A Neural network is defined as "a model of reasoning based on the human brain" [15]. Artificial neural networks are non-linear predictive models that learn through training and resemble biological neural networks in structure. It learns from a training set, generalizing patterns inside it for classification and prediction.

2) Genetic Algorithms: They are defined as optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

3) Decision Trees: They are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.

Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that we can apply to a new (unclassified) dataset to predict which records will have a given outcome.

4) Nearest Neighbor Method: The nearest neighbor method is a technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1).

5) Rule Induction: Rule induction is the extraction of useful if-then rules from data based on statistical significance.

6) Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Section 4 describes the Data Mining process illustrated by real examples of their use in breast imaging.

IV. Data Mining Process

Data Mining is the application of specific algorithms for extracting patterns from massive data and is only a step in the knowledge discovery in databases (KDD) process [16].

A. Acquisition of Data

We have developed a data warehouse system for breast cancer screening, details are in [17]. We complete here the process of Data Mining that we have not treated yet. We have used some data from this article because it is the same field of our study. The acquisition process in breast imaging includes four main phases which are complementary and complete the diagnosis of the patient.

- Clinical data, data about health patient history: screening history, current health status, and previous clinical examination.
- Image reading data, searching and extracting relevant information.
- Radiological interpretation data, clinical data (patient's history screening, current health status, information on previous clinical examination) and radiological data (information such as defined by the BI-RADS standard) [18].
- Anatomo-pathological data, which depends on the result of radiological interpretation. It grasps information about anatomo-pathological examinations such as type of procedure, reporting source, laterality, anatomo-pathology, staging and therapy.

Mammogram films selected and archived during screening are scanned and stored with the DICOM (Digital and Communication in Medicine) format [19].

B. The process

Data Mining consists of five elements:

- Extract, transform and load data.
- Store and manage data.
- Provide data access.
- Analyze data by a specific software.
- Present data in a useful format.

C. Functionning Principles in Breast Imaging

1) Extract, transform and load transaction data onto the data warehouse system

a) Extract Data: It is the process of capturing useful data from different sources, such as (RIS, HIS, PACS,...). We have to consider information system evolution. Data formats to be captured are heterogeneous and are being able to change in the time by evolutions of production systems as to integrate necessary supplementary data for taking into account new indicators, and to obtain a best granularity of information. At the phase of data extraction, tool usefulness is its capacity to capture data described under formats which are be to change in time and stored on systems which are be able to change too in time.

b) Transform Data: It consists of transforming captured data in such a way that to obtain an homogenous set of data which became comparables, additional, etc. We start by purifying captured data before transforming them. For that, we analyse data in order to identify abnormal ones which can not be introduced into the data warehouse. Data are filtered and duplications are eliminated.

c) Load Data: After completing the process of data extraction and transformation, data will be loaded: to transfer transformed and consolidated data into the decision database support, checking their consistency and building necessary index.

2) Store and manage data in a multidimensional database system: Data is stored in the database and managed in the multidimensional database system (data warehouse system).

3) Provide data access: Make data accessible and available for different users, radiologists, students and researchers.

4) Analyze data by a specific software: Data analysis is performed by using application software.

5) Present data in a useful format: Present the data in a useful format, such as a graph or table.

Section 5 describes material and methods selected for analyzing and modeling data derived from breast cancer screening knowledge.

V. Material and Methods

A. Material

For our experiment, we have used a database of mammographies of 20 patients with verified diagnoses.

B. Methods

Due to the complexity of medical data, it will be better in certain projects or diagnoses to adapt existing algorithms or optimize their use to obtain better results [20]. According to "Harper", the best performing algorithm depends on the features of the data at hand as well as any preference of the end-user [21].

On the other hand, the heterogeneity of the medical data such as: volume and complexity, physician's interpretation, poor mathematical categorization and canonical form motivates medical data miners to develop new approaches to analyze data [20].

To remediate to these deficiencies it will be advisable to create standard vocabularies, interfaces between different sources of data integrations, and design of electronic patient records.

In [22], the authors propose a strategy "decision fusion" for the classification of imaging data from multiple modalities, multiple sources and having various types of features [23].

A knowledge base represented in an ontology was introduced by [23] and [24] in order to improve the mining of temporal associations in clinical data.

C. Our Approach

We have applied all tasks of Data Mining, we illustrate one example for the Decision Tree.

1) Decision Tree

We have developed in a previous research work an ontology in the breast cancer domain to regroup and homogenize knowledge of experts according to BI-RADS. We have represented these data in a hierarchy [25]. Fig.2. illustrates a case of two categories of patients whom disease is classified with category 2 and 4 of BI-RADS.

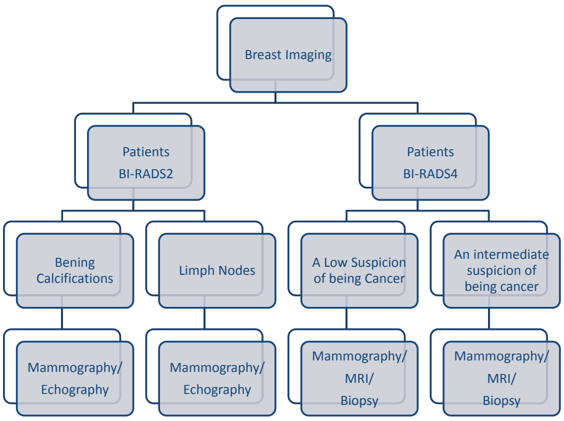

Fig.2. Illustration of the Decision Tree

2) Classification

Bayesian classifiers, neural networks, SVM, C4.5 and K-nearest neighbor (KNN) are the most frequently used algorithms for classification due to their high performance. The K-means algorithm is popular for clustering purposes. We adapt neural networks and the K-means algorithm for our study. We have classified all senological vocabulary according to BI-RADS. Because of the complexity of the domain, we don't represent examples of classification in this paper. Details for examples of classification are in [26].

Section 6 describes Data Mining tools available.

VI. Analytical Tools

Clinical analytics and business intelligence tools have emerged as a priority for hospital IT leaders. Clinical analytics or clinical decision support (CDS) is the rapidly developing field that harnesses real-time medical data to inference programs in order to generate fact-based diagnostic and therapeutic decisions, capture revenues and save costs [27].

Clinical decision support tools, generally defined as technologies that provide information to aid the diagnosis and treatment of patients, are set to fundamentally change the way medicine is practiced [28].

Data Mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved.

Commercial data-mining software and applications exist in medicine and other fields.

Section 7 concludes the work presented in this paper.

VII. Conclusion

We have presented concepts and techniques used to develop a Data Mining system particularly in medical field and imaging. We have shown that the different tasks of Data Mining are very benefit in a medical field. We can quote for example, the use of "Decision Tree" technique, which is very useful in the knowledge discovery process.

On the other hand, the utilization of this technique is very convenient because the Decision Tree is simple to understand. It combines data with various types, models non-linear functions, handles classification, and most of the available tools use it.

It is essential to properly collect and prepare the data, and to check the models against the real world.

The potential of Data Mining techniques in medical field allows to improve the quality and decrease the cost.

The application of information mining techniques to the medical domain are very helpful in extracting medical knowledge for diagnosis, decision-making, screening, monitoring, therapy support and patient management.

REFERENCES

[1] J. Han and M. Kambar, “Data Mining: concepts and techniques,” Second Edition, Editor, Morgan Kaufmann, 2006.

[2] N.M. Labib and M. N. Malek, “Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia,” in International Journal of Medical, Health, Biomedical and Pharmaceutical Engineering, 2007, vol.1, No:8, pp. 507-512.

[3] H. C. Koh and G. Tan, “Data Mining Application in Healthcare,” in Journal of Healthcare Information Management, 2005, vol. 19, No. 2.

[4] D. Tomar and S. Agarwal, “A survey on Data Mining approaches for Healthcare,” in International Journal of Bio-Science and Bio-Technology, 2013, vol.5, No.5 , pp. 241-266.

[5] M. R. Islam, M.U. Chowdhury and S. M. Khan, “Medical Image Classification Using an Efficient Data Mining Technique,” in Complexity International, 2005, vol.12, pp.1-9.

[6] J. Zhang W.Hsu, M.L. Lee, “Image Mining: Issues, Framesworks and Techniques,”.

[7] F. Gao, W. P. Kustas and M. C. Anderson, “A Data Mining Approach for Sharpening Thermal Satellite Imagery over Land,” in Remote Sensing Journal, 2012, vol. 4, pp. 3287-3319.

[8] N. Mishra and S. Silakari, “Image Mining in the Context of Content Based Image Retrieval: A Perspective,” in International Journal of Computer Science Issues, 2012, vol. 9, No 3.

[9] B. Thuraisingham”, “Managing and Mining Multimedia DataBases,” in International Journal of Artificial Intelligence Tools, 2004, vol. 13, No. 3, pp.739-759.

[10] H. M. Yousif, A. Al-Hussaini and M. A. Al-Hamami, “Using Image Mining to Discover Association Rules between Image Objects,” pp.1-17.

[11] J. W.H. Inmon, “Building the Data Warehouse,” Fourth Edition, Wiley Publishing, 2005, in Indianapolis, Indiana.

[12] P. Perner, “Mining Knowledge in Medical Image Databases,“ in Data Mining and Knowledge Discovery: Theory, Tools, and Technology, Belur V. Dasarathy (eds.), Proceedings of SPIE, 2000, vol. 4057, pp. 359-369.

[13] Timothy Hays, PhD, “Medical Data Mining,” Dynamics Research Corporation (DRC), 2012.

[14] B. Palace, “Data Mining Overview,” Highly cited summary of data mining describing technology, applications, and risks, [online], 1995, elsegundoca.ncr.com.

[15] M. Negnevitsky, "Artificial Intelligence, A Guide to Intelligent Systems," England: Pearson Education Limited, 2002.

[16] U.Fayyad, G.Piatetsky-Shapiro, P.Smyth, “KnowledgeDiscovery and Data Mining: Towards a Unifying Framework,” in Proc. 2nd International Conference. on Knowledge Discovery and Data Mining. AAAI Press; 1996, pp. 82-8.

[17] S. Demigha, “A Data Warehouse System to Help Assist Breast Cancer Screening, in Diagnosis, Education and Research,”, in CSA the Second International Conference on Computer Science and its Applications, IEEE, Jeju, Korea (South), 2009, pp. 1-6.

[18] BI-RADS, ACR, American College of Radiology, http://www.acr.org/.

[19] DICOM, “Digital Imaging and COmmunication in Medicine,” http://medical.nema.org/.

[20] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, A. Geissbuhler, “Clinical Data Mining: a Review,” Geissbuhler A, Kulikowski C, editors. IMIA Yearbook of Medical Informatics, 2009.

[21] P.R Harper, “A review and comparison of classification algorithms for medical decision making,” Health Policy, 2005, vol.71, pp. 315-31.

[22] J.L. Jesneck, L.W. Nolte, J.A. Baker, C.E Floyd, J.Y Lo, “Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis,” Med Phys, 2006; vol.33, pp.2945-54.

[23] G. Tusch, C.E Bretl, M. Connor, A Das, “SPOT Towards Temporal Data Mining in Medicine and Bioinformatics,” in AMIA Annu Symp Proc 2008, pp. 1157.

[24] R. Raj, M.J. O’Connor, A.K. Das, “An ontology-driven method for hierarchical mining of temporal patterns:application to HIV drug resistance research,” AMIA Annu Symp Proc 2007, 614-9.

[25] S. Demigha, “An Ontology Supporting the Daily Practice Requirements of Radiologists Senologists with the Standard BI-RADS,” ICEIS, Ninth International Conference on Enterprise Information Systems, Proceedings of Information Systems Analysis and Specification, 2007, Vo.l ISAS, pp. 243-249.

[26] S. Demigha, “A Knowledge-Based System in Radiology-Senology dedicated to the junior-senologists training,” PhD thesis in computer science, 2005, University of Paris 1 Panthéon-Sorbonne.

[27] E. McCann, “Clinical analytics 'next big thing',” Managing Editor *at* Healthcare IT News, 2012.

[28] R. Pizzi, “CDS tools can change medical practice,” Editorial Director at HIMSS Media, 2008.