A KNOWLEDGE DISCOVERY APPROACH FOR BREAST CANCER MANAGEMENT IN THE KINGDOM OF SAUDI ARABIA

Asem Omari

Assistant Professor
College of Computer Science and Engineering
Hail University, Hail, Kingdom of Saudi Arabia
a.omari@uoh.edu.sa

ABSTRACT

In this paper, we introduce an approach to improve and support decision-making process for breast cancer management in the Kingdom of Saudi Arabia. This can be accomplished by applying different association rule mining algorithms on the cancer information system in Saudi Arabia. It also provides valuable information about predicted distribution and segmentation of cancer in Saudi Arabia, which may be linked to possible risk factors. From the extracted patterns, the information need to be considered in the decision-making process can be identified and recognized as well, which yields to knowledge based decisions. Consequently, identifying health risk behaviors among target group of patients and adopting interventional and preventive measures can be initiated in order to decrease breast cancer incidence and prevalence and ultimately the health care costs.

KEYWORDS

KDD, Association Rule Mining, Breast Cancer, Cancer in Saudi Arabia

1. INTRODUCTION

In addition to heart diseases and car accidents, cancer is ranked to be one the top five leading causes of death among the Saudi population [1]. Large amount of data about cancer patients are collected in the national Saudi Cancer Registry (SCR) in the Kingdom of Saudi Arabia. Effective technologies need to be used to participate in preventing or at least reducing new breast cancer cases. The Knowledge Discovery from Databases (KDD) is one of those technologies used to extract interesting information from databases or data warehouses [2]. Data mining is considered the most important part in the KDD process provide tools for automated learning from historical data and developing models to predict future trends and behaviors. Data mining has two main models; predictive and descriptive models. Predictive data mining model tries to find out unknown values or future behaviours depending on other variables or historical values. Descriptive data mining techniques extract useful patterns that can give a description of the mined data and find out attributes and properties of the data. For example, in a hospital database, it may be found that a large number of patients, who are infected by disease A, are also infected by disease B. The patterns extracted using data mining help organizations make better and knowledge based decisions.

DOI: 10.5121/hiij.2013.2301

Data mining has many different applications in different fields of science [3]. In bioinformatics, protein sequences can be analyzed to find the relationship between some specific sequence and a specific disease. Data mining is used also in medical applications. Patient records include patient's demographic information, symptoms, blood measurements and laboratory test results. Data mining can for example find the relationship between two different diseases. In this paper, we introduce a solution scenario on how to invest data mining techniques to support the decision making process in the field of cancer management in order to describe, predict, prevent, and control cancer in Saudi Arabia which will consequently participate in reducing cancer cases, preventing new cases to occur, and help in implementing new medical, or social strategies to come with better solutions and decisions to fight this national problem in Saudi Arabia.

This paper is structured as follows: An overview about cancer in Saudi Arabia is introduced in section 2. Related work is presented in section 3. In section 4, we give an overview about the knowledge discovery process, and discuss its phases. Then, in section 5, we will see how the extracted patterns using different Data Mining techniques and more specifically association rule mining can be used to improve the decision making process for cancer management and see how to invest the extracted interesting patterns in the decision making process. Finally, in section 6, we summarize our paper and present future work.

2. CANCER IN SAUDI ARABIA

According to King Hussein Cancer Center (KHCC) [4], Cancer happens when cells that are not normal grow and spread very fast. Normal body cells grow and divide and know to stop growing. Over time, they also die. Unlike these normal cells, cancer cells just continue to grow and divide out of control and do not die. Cancer cells usually group or clump together to form tumors. A growing tumor becomes a lump of cancer cells that can destroy the normal cells around the tumor and damage the body's healthy tissues [4].

The Saudi Cancer Registry (SCR) annual report describes the cancer incidence in Saudi Arabia and provides valuable statistics about the situation of cancer among the Saudi population. SCR collects data from the pathology and hematology laboratories, in addition to the hospitals distributed all over the country from all health sectors [1]. According to [5], Breast cancer led the list of total cancer cases seen from 1975 to 2011 with 11.7%, followed by leukemia (8.6%), non-Hodgkin's lymphoma (7.6%), thyroid (6.7%) and colon, rectum (5.0%). Figure 1 shows the distribution of 20 most common Tumors from 1975 to 2011.

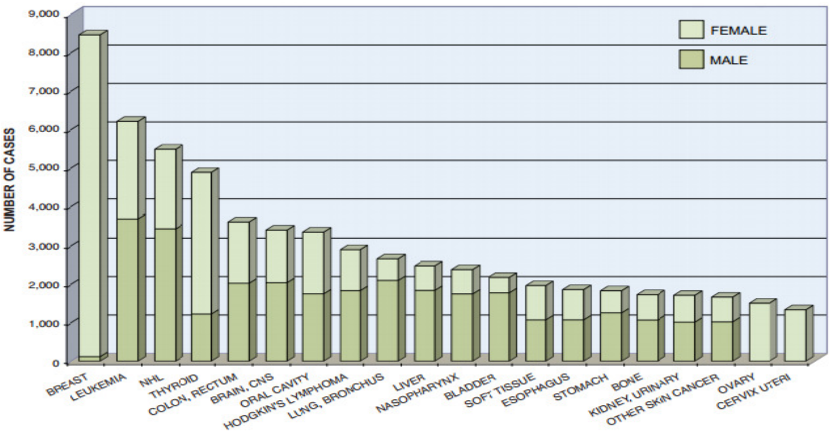

Figure1. Distribution of 20 most common tumours in Saudi Arabia [16]

3. RELATED WORK

A lot of research has been done for cancer control, prediction, and management using statistical approaches as well as different data mining techniques. The authors in [6] demonstrated the use of an association rule mining approach to discover associations between selected socioeconomic variables and the four most leading causes of cancer mortality in the United States. The work in [7] presented the mining processes and results of discovering trends and patterns from a set of health and living habit questionnaire data. The task was to discover the primary factors of cancer patients with the questionnaires. These factors where helpful to control cancer and decrease cancer incidence. The authors in [8] presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques and compared the performance of different data mining algorithms. The work in [9] discussed several data mining algorithms and techniques and proposed an architecture for medical knowledge information systems to permit data mining across several medical information sources. The authors in [10] discussed the process of designing a prototype that can help in the management of Childhood *Acute Lymphoblastic Leukemia (ALL).* They used a decision tree algorithm to mine patients medical history and diagnosis in order to improve the process of disease management. In western countries, data mining usage for cancer management has been implemented and reported as many and variable.

The authors in [11] presented a comparison study of three data mining classification models: multi-layer perception neural networks, *C4.5* decision trees and *Naive Bayes*. The classification models are built for breast cancer survivability prediction. The data set used is collected from registries across Saudi Arabia. The result showed that Decision tree is the most accurate predictor for breast cancer survivability in Saudi Arabia (Accuracy 0.979%). In [12], the authors compared different classification learning algorithms.

4. THE KNOWLEDGE DISCOVERY IN DATABASES PROCESS

The following are the main steps of the knowledge discovery process:

- Understanding the Application Domain: Understanding the application domain and identifying the goals of the data mining process is very important for a successful knowledge discovery project [13].

- Data integration and Gathering: In the data integration and gathering step the target data sets are gathered from different data sources for example from heterogeneous databases and data warehouses and combined in a suitable manner [14]

- Data Preprocessing: the main tasks of this step include solving the problem of missed and repeated data, Eliminating errors and transforming the selected data to format that are appropriate for the data mining procedure[15].

- Data Mining: In this step, the suitable data mining technique(s) for the studied data are decided to be applied [14] based on the type of the data and the type of patterns expected to be extracted. In the next section we discuss briefly different data mining techniques.

- Visualization: presenting the extracted data in a visual manner improves greatly the process of pattern readiness and analysis as the visualization techniques provide a better way to see the extracted information for example, giving every extracted group or cluster a different color can help the data analyst recognize the extracted groups [16].

- Pattern Evaluation: In this step, based on pre-defined measures, all interesting patterns representing meaningful knowledge are identified. It is common to combine some of these steps together. Furthermore, the data analyst can jump within different steps by using another algorithm or changing the format of the target data. In the following section, we discuss the basic data mining tasks: Data characterization, association rule mining, sequential pattern mining, classification, and clustering.

5. APPLYING DATA MINING ON CANCER DATA

In order to get valuable information about the relationships between both patients and cancers, several data mining techniques can be applied on that information system, where the result of those data mining techniques (i.e. patterns) are that kind of information that is not available directly in the information system as records or fields and cannot be extracted directly by running some query or traditional statistical tool. This extracted information can give the data analyst a better view about the patients, cancers, the relationship between patients and cancers, and the behavior of both patient and cancers with respect to time or other environmental, social, or bibliographic attributes such as:

- Who are the frequent patients?
- What kind of breast cancers are they usually infected by?
- What kinds of breast cancers occur together?
- If the patient is infected by breast cancer x may he be infected by breast cancer y directly or within a period of time?
- What kind of effects has some bibliographic attribute on the development of breast cancer?

5.1. Data characterization

Understanding the target data is important for an efficient data mining process. Aggregate functions can be used to summarize the target data and give a statistical overview about the data. Furthermore, Online Analytical Processing technique can also be used to explore the data with respect to different aspects and conceptual levels. As mentioned before, data visualization can also be used to show target data and also the extracted patterns in a visualized manner. Data characterization is the summarization of the general characteristics or features of a target class of data [17] such as: Describe and summarize the characteristics of patients who react positively to the diagnoses process within one year. The result could be a general profile saying that they are 20-40 years old, and have religious attitude.

5.2. Sequential pattern mining

Sequential pattern mining is the mining for frequently occurring patterns with respect to time [18] for example; Patients who are infected by Lung cancer and are Singles seem to be infected by Leukemia cancer within 9 months.

5.3. Classification

Classification is the process of finding a set of models or functions that describe and distinguish data classes or concepts where the models derived based on a set of training data [17]. The following pattern can be a result of using classification data mining technique: Classifying patients with respect to their age into three classes. The first class contains patients who are less than 20 years of age. The second class contains patients who are between 20 and 40 years old. The third class contains patients who are more than 40 years old. After analyzing the characteristics of each class the result could be that patients who are less than 20 years old are usually infected by Lymphoma, Leukemia, and Brain cancers. While patients between 20 and 40 years old are usually infected by Prostate, Stomach, and Kidney cancers. Patients who are more than 40 years of old are usually infected by Breast, and Lung cancers.

5.4. Clustering

In clustering data objects have no class label. The objects are clustered or grouped based on the principle that objects in one class have high similarity to one another but are very dissimilar to objects in other clusters [13]. In clustering there are no predefined class labels. The following rule is an example of interesting patterns using a clustering approach:

- Patients who are living in Jeddah are mostly in risk of infection of collection A of cancers which contains Urinary Bladder, Lymphnodes, Brain, Lung, and Prostate. On the other hand, female patients are mostly in risk of infection of a subset of collection B of cancers which includes Breast, Ovary, Thyroid, and Hodgkin Lymphoma.

5.2. Association rule mining

Association rule mining (ARM) is the discovery of association rules showing attribute values that occur frequently together in a given set of data [17]. According to [13], an association rule is an expression of the form X »Y, where X and Y are sets of items and have no items in common. This rule means that given a database of transactions D where each transaction T є D is a set of items. X »Y denotes that whenever a transaction T contains X then there is a probability that it contains Y, too. The rule X »Y holds in the transactions set T with confidence c if c% of transactions in T that contain X also contain Y. The rule has support s in T if s% of the transactions in T contains both X and Y [13][17].

According to [9], Association rule mining is the process of searching for frequent itemsets, ARM algorithms employ one of two common approaches: Breadth-first search Approach (BFS), and Depth-first search approach (DFS). The Apriori algorithm [14], The *AprioriTID* Algorithm, *AprioriHyprid* Algorithm [15], and *FP-Growth* Algorithm [5] are common Association Rule Mining algorithms.

Association Rule Mining is finding all Association Rules that are greater than or equal a user-specified minimum support and minimum confidence.. The first step of association rule mining is finding all itemsets that satisfy minimum support (known as Frequent-Itemset generation). The second step is generating all Association Rules that satisfy minimum confidence using itemsets generated in the first step. The following rules can be a result of applying association rule mining techniques on on a database that contains patients' history and demographic information:

- If a patient record R contains Kidney cancer there is a 70% chance that it contains Bone cancer as well, and 4% of all records contain both.
- 80% of the patients who are infected by Breast cancer is also infected by Lymphoma cancer and 8% of all patients are infected by both Breast and Lymphoma cancers.

That rules are called interesting association rules. Suppose that we have three breast cancer types *BC1, BC2*, and *BC3* and the factors *F1, F2, F3, F4,* and *F5* represent different factors that are related to patient history or demographic information such as sex, age, weight, social level, address, and marital status. For example, the relationship between breast cancers *BC1*, and *BC3* is created depending on a rule that says:

BC1 » BC3 [support = 4%, confidence = 60%]

This rule means that 60% of patients who are infected by Breast Cancer *BC1*, are also infected by BC3 cancer type, and 4% of all patients are infected by both.
The following rule can also be created:

F5 (X, ”20...35”) BC2

This rule means that patients who are between 20 and 35 years of old are infected by *BC2* cancer type. The extracted patterns from applying data mining and more specifically association rule mining techniques on the cancer information system database give the data analyst an overview about the behavior of the patients, the relation between patients and cancers, what kinds of cancers are usually occur together, and about cancer occurrence habits with respect to different bibliographic attributes such as patients age, or address.

Thus, the decision maker can come with some decisions that will improve the treatment strategies, improving preventive medicine techniques, reducing cancer cases, preventing new cases to occur, and help in implementing new medical or social strategies to come with better solutions and decisions to fight this national problem in Saudi Arabia. Data mining as an advanced computational technology can has a significant impact on health care field in term of its application. Using data mining as a descriptive and predictive technique in health care systems can identify health risk behaviors and has a significant economic impact. Investing this technology in Saudi Arabia, will enhance setting well-structured and designed plans to deal effectively with such risky behaviors, particularly cancer health risk behaviors.

Furthermore, it helps characterize cancer patient behavior to predict office visits, identify successful medical therapies, and predicts the effectiveness of therapeutic procedures or medical tests used in cancer management. The primary health care system that is mainly concerned with health promotion and disease prevention is the principal benefit, where various health care strategies can be adopted to decrease cancer incidence and prevalence in Saudi Arabia.

6. SUMMARY AND FUTURE WORK

In this paper we provided a knowledge discovery approach to improve and the decision making process for breast cancer management in the Kingdom of Saudi Arabia, by applying association rule mining on the contents of the cancer information system in Saudi Arabia. From the extracted patterns we can recognize the information need to be considered in the decision making process which yields to knowledge based decisions. Although data mining is becoming a complementary application for many clinical researches in the medical and bioinformatics fields, there are still limitations that cannot be ignored.

In the experimental work that followed our previous work in [19], we made an exploratory study on the data available in the Jordan cancer registry [4], and we found that the data was not well-structured and formatted to be mined. Therefore, we were not able to mine the data. From this point our future work will start from preparing the cancer information database in Saudi Arabia and restructure it to get efficient and valuable results. We also plan to build a database from scratch for cancer patients' information in Hail region in Saudi Arabia that can be later expanded to include other regions in the country. This database should be the seed for well-structures cancer data in Saudi Arabia

REFERENCES

[1] Saudi Ministry of Health. Health statistical Year Book 2011. Last viewed 10.06.2013 from: http://www.moh.gov.sa/Ministry/Statistics/book/flash/1432/MOH_Report_1432.ht

[2] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, pages 1–34, 1996.

[3] B. J. Read. Data Mining and Science: Knowledge Discovery in Science as Opposed to Business. In Proceedings of the 12th ERCIM Workshop on Database Research, Amsterdam, 1999.

[4] King Hussein Cancer Center. Cancer Information. http://www.khcc.jo/cancerinformation.asp, 2009.

[5] King faisal specialist hospital and research centre, Oncology Centre Research Unit, Tumor registry annual report 2011.

[6] S. Vinnakota and N. S. Lam. Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. In International Journal of Health Geographics, volume 5. BioMed, 2006.

[7] X. Zhang and T. Narita. Discovering the Primary Factors of Cancer from Health and Living Habit Questionnaires. In Discovery Science DS99, pages 371–372. Springer Verlag, 1999.

[8] E. G. Abdelghani Bellaachia. Predicting Breast Cancer Sur¬vivability Using Data Mining Techniques. Scientific Data Mining workshop in SIAM Conference on Data Mining, pages 1–4, 2006.

[9] Andrea Houston, Hsinchun Chen, Susan Hubbard, Bruce Schatz, Tobun Ng, Robin Sewell and Kristin Tolle. Medical Data Mining on the Internet: Research on a Cancer Infor¬mation System. Artificial Intelligence Review, 13:437–466, 2000.

[10] N. Labib and M. Malek. Data Mining for Cancer Manage¬ment in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia . In Proceedings of the World Academy of Sci¬ence, Engineering, and Technology, 2005.

[11] Ghofran Othoum1 and Wadee Al-Halabi. Predicting Breast Cancer Survivability Rates For data collected from Saudi Arabia Registries. PROCEEDINGS OF THE 2011 INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE volume 1 pp 918-923.

[12] Adel Aloraini. DIFFERENT MACHINE LEARNING ALGORITHMS FOR BREAST CANCER DIAGNOSIS. International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.6, November 2012.

[13] Ian H. Witten and E. Frank. Data Mining: Practical Ma¬chine Learning Tools and Techniques. Morgan Kaufmann Publishers, San Francisco, 2005.

[14] Asem Omari. Data Mining for Retail Website Design and Enhanced Marketing. VDM Publishing, Saarbruecken, Ger¬many, 2008.

[15] R. Cooley, B. Mobasher, and J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1(1):5–32, 1999.

[16] Tom Soukup and Ian Davidson. Visual Data Mining: Tech¬niques and Tools for Data Visualization and Mining. John Wiley & Sons, first edition, 2002.

[17] J. Han and M. Kamber. Data Mining Concepts and Tech¬niques. Morgan Kaufmann Publishers, San Francisco, 2001.

[18] J. Vel´asquez, H. Yasuda, and T. Aoki. Combining the Web Content and Usage Mining to Understand the Visitor Behav¬ior in a Web Site. In Proc. 3rd IEEE International Confer¬ence on Data Mining (ICDM 2003), pages 669–672. IEEE Computer Society Press, 2003.

[19] A. Omari and I. Hweidi, On The Usage of Data Mining as a Descriptive and Predictive Tool for Cancer Management in Jordan: A Scenario. In Proceeding of DMIN, 2009, pp.396-402.