Volume 3, Issue 9, September 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Approach on Sequential Data Mining Algorithm for Breast Cancer Diseases Management

Dr. P. Ponmuthuramalingam,
Associate Professor,
Department of Computer Science,
Government Arts College, Coimbatore,
India

Mrs. M. Yasodha,
Assistant Professor,
Department of Computer Science,
D.r N.G.P College of Arts & Science, Coimbatore,
India

Abstract: The Breast Cancer has converted a common expiry factor in India. Despite the circumstance, not all overall hospitals have the mammogram services. The extended waiting for identifying a breast cancer may upsurge the option of fatality and the cancer scattering. Therefore a computerized breast cancer may rise diagnosis prototype has been developed to decrease the time taken and circuitously dropping the probability of death. Micro calcification on X-ray mammogram is a noteworthy mark for early discovery of breast cancer. Breast cancer diseases expressively affect the quality of life of people and are among the most common and costly health problems. Due to the difficulty of these diseases, it is problematic for clinicians to analyse trends in patient data and relate these trends with other patient information such as demographic data. Therefore, there is a need for informatics tools to efficiently monitor disease progression and to analyse trends in patient data to improve disease management. Moreover, because this disease has been identified as among the most avoidable diseases, these tools can also be used to identify patients at risk and provide information for early intervention. To this end Sequential Data Mining framework was developed to identify frequently occurring the patterns in patient measurements that may lead to development of diseases.

Keywords: Diseases management, Demographic data, Data mining, Patterns, Sequential data mining.

I. Introduction

Data mining can be defined as an activity that extracts some new nontrivial information contained in large databases. The goal is to discover hidden patterns, unexpected trends or other sub title relationships in the data using a combination of techniques from machine e-learning, statistics and database technologies. This new discipline today finds application in a wide and diverse range of business, scientific and engineering scenarios. Other situations where data mining can be of use include analysis of medical records of hospitals in a town to predict, for example, potential outbreaks of infectious diseases. The list of application areas for data mining is large and is bound to grow rapidly in the years to come.

Data mining is usually defined as the extraction of before unknown and possibly valuable info from a database. With the rising volumes of electronic patient records, data mining has become general to excerpt hidden patterns inpatient data for healthier understanding of relationships within the data. Data mining in medical domain is single from that in other domains due to the special characteristics of medical datasets. Medical datasets are often privacy-sensitive, huge and heterogeneous with data collected from dissimilar sources. The collected data may also need to be branded mathematically. The rest of the section discusses a few data mining educations that have been lead in medical and clinical areas. In clinical domain, data mining methods have been applied to big clinical repositories containing clinical and administrative data collected from electronic sources to identify new disease associations. The techniques applied include pattern detection to identify commonly happening associations in the dataset, predictive analysis to predict upcoming outcome for a patient based on the existing patient records, and association mining to excerpt interesting rules from the recognized associations.

There have been many recent studies to predict the survival of patients with fatal diseases and to predict treatment outcomes. Studies were conducted by Oztekinet. al. to predict the survivability of heart-lung transplantation patients and by Delen et. al. to forecast the survivability of breast cancer patients using prediction models such as neural networks , decision trees , and regression . Decision tree algorithms were also used to efficiently predict the survival period of kidney dialysis patients and bladder cancer treatment outcomes. Decision trees based on rules were shaped and decision making algorithms were used to predict outcomes.

II. Breast Cancer Diseases –An Overview

The Breast Cancer is the one of the foremost cancers. About 10% of all women grow breast cancer and about 25% of all cancers diagnosed in women are breast cancers[21]. Although actual anticipation is not possible, early discovery can at least decrease the chance of breast cancers from becoming hopeless [22]. Mammography has been shown to be the most real and dependable method for early cancer detection. Mammogram clarification is both laborious and problematic, requiring the information of trained radiologists.

Breast Cancer as a Domain

Breast tumour may be benign or evil lesions. Benign is a non-cancerous scratch and malignant is cancerous lesion[23] .Benign lesions in the breast tend to be well defined and restricted, while malignant lesions appeared ill defined with speculated margins. Classification of breast lesions can be performed based on border characteristics of lesions, by using the fact that benign lesions have smooth borders, while malignant lesions have rougher border.

Breast cancer is a malignant tumor that has established from cells of the breast. A malignant tumor is a group of cancer cells that may involve surrounding tissue or spread to distant areas of the body.

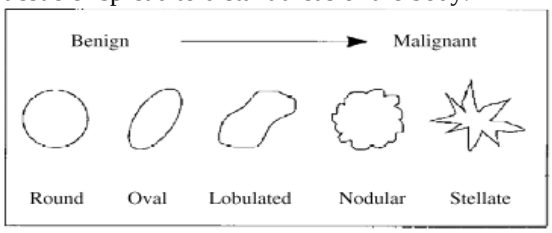

FIGURE 1: Morphologic Spectrum of Masses

The etiologies of breast cancer remain unclear and no single dominant cause has emerged. Defensive way is motionless a mystery and the only way to help patients to survive is by early discovery. If the cancerous cells are detected before dispersal to other organs, the survival rate for patient are more than 97%. Therefore it has been our incentive factor to develop such system .

Some early signs of having high possibility of cancerous cells, like micro classifications, mass, architectural misrepresentation and breast irregularities; might be detected from mammogram images. Mammogram[24], introduced in 1969 is the first digital steps in detecting cancerous cell in breast tissues, and it has been a useful tool in diagnosing breast cancer ever since.

A closer inspection of the mammograms reveals several difficulties for the irregularity approach. First, the global presence (brightness, contrast, etc.,) of the two breasts may differ, usually due to variations in the recoding method. This can be solved in the improvement phase. Second, due to natural irregularity, and due to the mammography recording method, the shapes of the left and right breast do not match. Defining corresponding positions in both breasts becomes therefore a nontrivial task. Third, asymmetries exist not only at the tumor position but also in the appearance of healthy breast tissue at corresponding locations in the two breasts. The asymmetry method thus has to discriminate tumor related asymmetries from obviously occurring asymmetries.

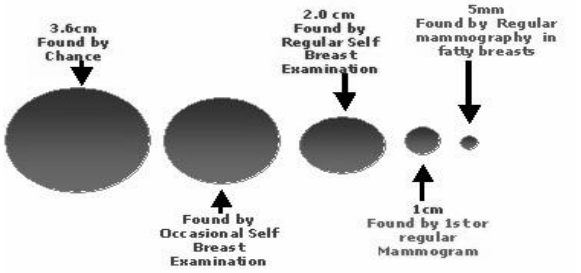

FIGURE 2: Dimensions of mass that may be sensed by mammogram or hand

As some of cancerous tissues can be very aggressive [24], it is very important to identify them as early as possible. But the waiting time from capturing mammogram images to biopsy result is 2 weeks to a month, on average. However, 10-30% visible abnormalities are usually detected, due to technical or human error. The whole procedure in diagnosing breast cancer in woman is shown in Figure 2 below.

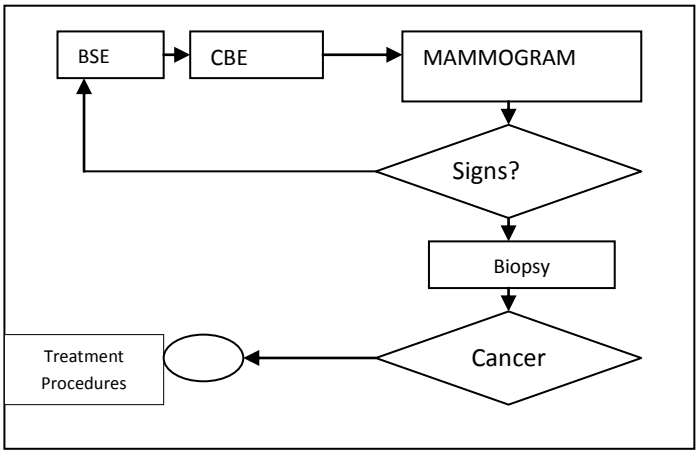

FIGURE 3: Procedures of diagnosing breast Cancer

The efficiency of early discovery has been confirmed to reduce a lot of mortality among breast cancer patients[25]. As a proof, 80% of American society detected cases are still in early stage, but the mortality among them is only 30%in 2006, believed as a result of early detection and better treatment[26].

3. DATA MININGAPPROACH

Since Data mining brings together techniques from different fields such as statistics, machine knowledge and databases, the literature is scattered among many different sources. We mainly concentrate on algorithms for pattern detection in sequential data streams. By sequential data, we mean data that is well-ordered with respect to some index. For example, time series constitute a popular class of sequential data, where records are indexed by time. Other examples of sequential data could be text, gene sequences, protein sequences, lists of moves in a chess game etc. Here, although there is no idea of time as such, the ordering among the records is very important and is central to the data description/modelling.

2.1 Models and Patterns

The types of constructions data mining algorithms appearance for can be classified in many ways. Models and patterns are structures that can be estimated from or matched for in the data. These constructions may be utilized to attain various data mining objectives. A model is a global, high-level and often abstract picture for the data. Typically, models are specified by a collection of model parameters which can be estimated from the given data. Often, it is possible to further classify models based on whether they are predictive or descriptive. Predictive models are used in prediction and classification applications while descriptive models are useful for data summarization.

III. Data Mining With KDD

Data mining is actually an essential part of Knowledge Discovery in Database (KDD) process, which is the overall process of converting raw data into use-ful information. A typical KDD process which consists of five steps :

1. Data collection and cleaning: selecting attributes, dealing with errors, identification of the necessary background information, etc.
2. Choice of pattern discovery method: deciding on the types of knowledge to be exposed, parameter selection, etc.
3. Discovery of patterns (data mining): running algorithms for discovering different types of patterns
4. Pattern Presentation: selecting stimulating patterns, visualisation of results, etc.
5. Putting information into use.

Data mining refers to the third step in this process and is defined as the application of specific algorithms for extracting patterns from data. The KDD process is interactive and iterative. Depending upon the extent to which the results satisfy the user's goals, iteration can be performed between any two steps.

IV. Sequential Data

In sequential data, events notonly can occur at an instant (time point), but also can occur over a time interval.In this thesis, an ordered list of events occurring at an instant is called an eventsequence and an ordered list of events occurring over a time interval is called aninterval sequence.

4.1 Issues

A large number of studies have been concentrated on analysing event sequencedata, while the analysis of interval sequence data has received relatively little attention. As a result, there are still many issues that need to be addressed in the area of analysing interval sequence data. Three main issues that will be addressed in this thesis are as follows.

1. The first important issue is that of what constitutes an interesting pattern in data. As an example, the notions of sequential patterns or frequent episodes represent the currently popular structures for patterns in event sequence data. Therefore, the problem of defining structure for interesting patterns in interval sequence data would be a problem that deserves attention.
2. In all data mining applications, the primary constraint is the large volume of data. Hence there is always a need for efficient algorithms. Therefore, de-signing efficient algorithms for the discovery of patterns in interval sequence data is a problem that would continue to attract attention.
3. Another important issue is that of the analysis of discovered patterns so that one can find interesting patterns from a large number of pat-terns generated by data mining algorithms. Making sense of such a large number of patterns presents a significant challenge. Therefore, there is a need for tools to assist the user finding interesting patterns from a set of discovered patterns.

V. Proposed Work

Sequential Pattern Mining pattern discovery model, a dynamic programming algorithm is used for finding optimal information preserving decomposition and optimal lossy decomposition. A closed relationship is discovered between the decomposition of time period associated with the document set and the significant information computed for analysis, the problem of identifying suitable time decomposition for a given document set which does not seem to have received adequate attention. So the time point is defined in interval and decomposition. Time point is given by base granularity such as seconds, minutes, day etc. The time interval between t1 and t2 is defined as $t1 \leq t \leq t2$.

Decomposition of time interval T is given as sequence of time intervals T1, T2, T3, T4… Tn And 'T' is computed by T=T1*T2*T3*T4*…….*Tn. The information is mapped with the keyword 'wi' and document dataset 'D' as, fm(wi,D) = v.

5.1. LAPIN Sequential Pattern Mining

Definitions, Lemmas, and Theorem Definition (Prefix border position set) Given two sequences, For a prefix sequence, C, in a sequence, i, if the prefix border position, Sc, i, is smaller than, or equal to the last position of a candidate IE item,

β, in the same sequence, then C can be extended to C ¦ β as an Item set Extension in the sequence, i. Proof: Since the last location of the candidate IE item β is larger than or equal to Sc, i , at least one β appears behind the prefix sequence C in the sequence i, which means the Item set Extension C ¦ α exists in the sequence i.

Algorithm: Frequent pattern sets generation

```
Generate difference-table DT from I1 and I2;
generate an array Ar;
for each transaction tr 2 DT
for each ith attribute At 2 U
if (tr(At) ==0 00)
then Ar[i × 3] + +;
elseif (tr(At) ==0 −10)
then Ar[i × 3 + 1] + +;
else Ar[i × 3 + 2] + +;
All ;;
k 1;
Lk all patterns with counter _ min sup;
while(Lk¬;){
All = All [ Lk;
Ck =generate next itemsets(Lk);
Ck.prune(min sup);
Lk++ = Ck;
}
return All;
```

［Theorem 1］（Frequent sequence） Given a user specified mini- mum support, ε, a sequence, S, is frequent if, by Sequence Extension checking, its support, Sup(S),is > = ε, or, by Item set Extension checking, its support, Sup(S),is > = ε. 2.2 LAPIN: Design and Implementation In this section, we describe the LAPIN algorithms used to mine sequential patterns in detail. As in other algorithms, certain key strategies were adopted, i.e., candidate sequence pruning, database dividing, and customer sequence reducing. Combined with the LAPIN strategy, our algorithms can efficiently find the complete set of frequent patterns. We used the Depth First Search (DFS).

VI. Conclusion

In the paper, we proposed a mining outline in different groups of patients to classify and study frequent temporal changes in measurements that lead to a disease. The mining model developed in this research can be used in various clinical settings to monitor the progression of cancer diseases, to analyze trends in patient data, to identify patients at risk, and to provide information for early interventions. To accommodate the diversely structured clinical data collected from different sources, we attempted to make the mining model as generic as possible. Using sequential model and frequent pattern set algorithm we have to identify the cancer cells easily.

Reference

[1] Centers for Disease Control and Prevention (CDC).(2009, October). Chronic Disease Prevention and Health Promotion. Available: http://www.cdc.gov/nccdphp/

[2] American Cancer Society. (2009, October).Breast Cancer Facts & Figures 2009- 2010. Available: http://www.cancer.org/downloads/STT/F861009_final 9-08- 09.pdf

[3] W. Mahamaneerat, C-R.Shyu, S. Ho, and C. Chang, "Domain-Concept Association Rules Mining for Large Scale and Complex Cellular Manufacturing Tasks," Journal of Manufacturing Technology Management, vol. 18, pp. 787-806, 2007.

[4] M. H. Dunham, Data Mining: Introductory and Advanced Topics, 1st ed., New Jersey: Prentice Hall/Pearson Education, 2003.

[5] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," ArtifIntell Med, vol. 26, pp. 1-24, 2002.

[6] I. M. Mullins, M. S. Siadaty, J. Lyman, K. Scully, C. T. Garrett, W. G. Miller, R. Muller, B. Robson, C. Apte, S. Weiss, I. Rigoutsos, D. Platt, S. Cohen, and W. A. Knaus, "Data mining and clinical data repositories: Insights from a 667,000 patient data set," ComputBiol Med, vol. 36, pp. 1351-77, 2006.

[7] A. Oztekin, D. Delen, and Z. J. Kong, "Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology," Int J Med Inform, vol. 78, pp. 84-96, 2009.

[8] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," ArtifIntell Med, vol. 34, pp. 113-27, 2005.

[9] S. Haykin, Neural networks: a comprehensive foundation. New Jersey: Prentice Hall, 1998.

[10] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, pp. 81- 106, 1986. 61

[11] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning New York: Springer-Verlag, 2001.

[12] A. Kusiak, B. Dixon, and S. Shah, "Predicting survival time for kidney dialysis patients: a data mining approach," ComputBiol Med, vol. 35, pp. 311-27, 2005.

[13] S. C. Shah, A. Kusiak, and M. A. O'Donnell, "Patient-recognition data-mining model for BCG-plus interferon immunotherapy bladder cancer treatment," ComputBiol Med, vol. 36, pp. 634-55, 2006.

[14] G. Richards, V. J. Rayward-Smith, P. H. Sonksen, S. Carey, and C. Weng, "Data mining for indicators of early mortality in a database of clinical records," ArtifIntell Med, vol. 22, pp. 215-31, 2001.

[15] S. P. Imbermana, B. Domanskia, and H. W. Thompsonb, "Using dependency/association rules to find indications for computed tomography in a head trauma dataset, " Artificial Intelligence in Medicine, vol. 26, pp. 55-68, 2002.

[16] M. Toussi, J.-B.Lamy, P. Toumelin, and A. Venot, "Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes," BMC Medical Informatics and Decision Making, vol. 9, 2009.

[17] B. Honigman, P. Light, R. M. Pulling, and D. W. Bates, "A computerized method for identifying incidents associated with adverse drug events in outpatients," Int J Med Inform, vol. 61, pp. 21-32, 2001.

[18] M. S. Siadaty and W. A. Knaus, "Locating previously unknown patterns in data- mining results: a dual data- and knowledge-mining method," BMC Med Inform DecisMak, vol. 6(13), 2006.

[19] H. Mannila, H. Toivonen, and A. Verkamo, "Discovery of Frequent Episodes in Event Sequences," Data Mining and Knowledge Discovery, vol. 1, pp. 259-289, 1997.

[20]. Ajay NageshBAravanhalty, ShridarGancean, SHANNAN Agner, James Peter Monaco, " Computerised Image based detection and grading of tymphocytic infiltration in HER2 breast cancer histopathology ", IEEE Transactions on biomedical Engieeringvol 57, no-3 march-2010.

[21]. Yu –Len Huang and Dar-RenChen , 2004 " Watershed Segmentation for Breast tumor in 2-D Sonography" , Ultrasound in Med & Biol.,vol.30, No.5 , pp 625 – 632.

[22]. H.S.Sheshadri and A.Kandaswamy , 2004 , " Detection of Breast Cancer tumor based on Morphologicalwatershed Algorithm",www.icgst.com

[23]. Chen DR , Chang RF , H unag YL. 2000,"Breast cancer diagnosis using self organizing map for Sonography". Ultrasound Med Biol ., Vol.26, No.3 , pp 405-411

[24]. Chen DR , Chang RF , Hunag YL. 2000, " Texture analysis of breast tumors on sonograms", Semin ultrasound CT MR, Vol.24,No.4, pp 308-316

[25]. Rafel C. Gonzalez and Richard E.Woods, 2002, " Digital image processing" second edition, by Pearson Education Inc.

[26]. Haris K, Efstrafiadis SN,1998, "Hybrid image segmentation using watershed and fast region merging" , IEEE Trans Image Processing Vol.

