IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727Volume 8, Issue 4 (Jan. - Feb. 2013), PP 18-23
www.iosrjournals.org

Data Mining Applications in Medical Image Mining: An Analysis of Breast Cancer using Weighted Rule Mining and Classifiers

A.Kavipriya[1], B.Gomathy[2]
[1]*PG Scholar, Department of CSE, Bannari Amman Institute of Technology, India*
[2]*Research Scholar, Department of CSE, Bannari Amman Institute of Technology, India*

Abstract : *The Association Rule Mining methods are used to mine attribute relationships. The Support and Confidence values are estimated for all item-sets. Minimum Support and Confidence values are used to select frequent patterns. Classification technique is applied to assign labels for the transactions. Learning phase is carried out for transaction pattern identification. Testing process handles the pattern matching and label assignment task.*

Classification and Association Rule Mining techniques are integrated to perform disease prediction process. Decision tree or Naive Bayes Classifiers consider the feature is independent of each other. Associative Classifiers are especially fit to applications where the model may assist the domain experts in their decisions. Prediction model for medical domain needs high accuracy level. Pruning-Classification Association Rule (PCAR) is used to predict cancer diseases. A proficient methodology for the generation of association rules from the cancer disease warehouses is used for Breast Cancer prediction. The Pima Indian Breast Cancer data warehouse is pre-processed in order to make it suitable for mining process. The cancer disease data warehouse is binning with the aid of the modified equal width binning interval approach to discretizing continuous attributes. The width of the desired interval is chosen based on the opinion of medical expert and is provided as an input parameter to the model.

Associative Classifier uses Weighted Association Rule Mining (WARM) model for cancer patient diagnosis. The diseases prediction process is performed by comparing the symptoms of patients with the existing database. Different weights will be assigned for different attributes based on their predicting capability. The attribute with higher predicting capability will have higher weightage.

Keywords- *Associative Classifiers, ARM, Pruning Classification Association Rule (PCAR), Weighted Association Rule Mining (WARM).*

I. Introduction

In statistical and data analysis, the business analyst knows, what the variables are before can start to analyze. It needs knowledge discovery technology and tools. Knowledge discovery has its roots in artificial intelligence and machine learning.

Knowledge discovery may be a non trivial extraction of implicit, unknown and useful information from the data. Knowledge discovery may be the data search process, without a hypothesis or question and still finding either unexpected or interesting information in relationships and patterns among its data elements or important business rules in the full data searched and analyzed. Knowledge discovery may mean to uncover previously unknown business facts in the gigabytes of data in the data warehouse or data mart.

Business managers and analysts are always seeking new and additional business insights so that crucial business decisions, it has a significant impact on the health of a business can be improved. By using the traditional techniques of business queries and data analysis requires asking the right questions. Knowledge discovery technology determines by itself the questions to ask and then keeps on asking questions, digging deeper to unearth the nuggets of knowledge the business seeks. Business analysts do not have the time, attention span and stamina to ferret out all the implicit relationships and patterns that exist in the data warehouse.

A number of data mining algorithms have been recently developed that greatly facilitate the processing and interpreting of large stores of data. One example is the Association Rule Mining algorithm, which discovers correlations between items in transactional databases.

Apriori algorithm is an example of Association Rule Mining algorithm. By this algorithm, candidate patterns that receive sufficient support from the database are considered for transformation into a rule. This type of algorithm works well for complete data with discrete values.

Medical classification or medical coding is the process of transforming descriptions of medical diagnoses and procedures into universal medical codes. Diagnosis codes are used to track diseases and other health conditions to contagious diseases such asnorovirus, the flu, and athlete's foot. These diagnosis and procedure codes are used by government health programs. Medical classification systems are used for a variety of applications in medical informatics, including statistical analysis of diseases and therapeutic actions,

reimbursement e.g., based on diagnosis-related groups, knowledge-based and decision support systems and direct surveillance of epidemic or pandemic outbreaks.

Many different medical classifications into two main groupings:

- Statistical classifications
- Nomenclatures.

A statistical classification brings together similar clinical concepts and groups them into categories. Consider an example, International Statistical Classification of Diseases and Related Health Problems . However, there are several other clinical concepts that are also classified here. Another feature of statistical classifications is the provision of residual categories for "other" and "unspecified" conditions that do not have a specific category in the particular classification.

In a nomenclature, there is a separate listing and code for every clinical concept. So, in the previous example, the tachycardia listed would have its own codes. This makes nomenclatures unwieldy for compiling health statistics.

II. Related Work

Numerous works in literature related with cancer disease diagnosis uses data mining techniques have. Some of the works are discussed below:

This paper presents a novel and more efficient PCAR algorithm. It analyzes and consider Apriori algorithm. It's the algorithm to mine association rules.Breadth-first search strategy is used to counting the support of item sets and uses a candidate generation function which exploits the downward closure property of support.

Steps to perform Apriori algorithm:

1. Generating item sets that pass a minimum support threshold.
2. Generating rules that pass a minimum confidence threshold.
3. Bottom up approach is used, where frequent subsets are extended one item at a time a step known as candidate generation. When no further successful extensions are found,it terminates.
4. Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently.

Apriori algorithm gets large frequent item sets through the combination and pruning of small frequent item sets. The principle of the algorithm is: firstly calculates the support of all item sets in candidate item set *Ck* obtained by *Lk-1*, the support of the item set is greater than or equal to the minimum support, frequent *k*-item set considered as the candidate *k*-item set , that is *Lk*, then all frequent *k*-item sets combined into a new candidate item set *Ck+1*, level by level, until finds large frequent item sets.

The problem of identifying constrained association rules for cancer disease prediction was studied by Carlos Ordonez. Three constraints were introduced to decrease the number of patterns.

- Necessary attributes to appear on only one side of the rule.
- It segregates attributes into uninteresting groups.
- Number of attributes in a rule is restricted.

Two groups of rules envisaged the presence or absence of cancer disease in four specific cancer arteries. Data mining methods may aid the clinicians in the predication of the survival of patients and in the adaptation of the practices consequently.

III. Problem Statement

The healthcare industry collects huge amounts of healthcare data which unfortunately, are not "mined" to discover hidden information for effective decision making. Hidden patterns and relationships are often discovered. Advanced data mining techniques can help remedy this situation.By using medical profiles, it can predict the likelihood of patients getting a cancer disease. It enables the knowledge relationships between medical factors related to cancer disease, to be established. Data mining technology provides a user-oriented approach for hidden patterns in the data. Those knowledge can be used by the healthcare administrators to improve the quality. Numerous fields associated with medical services like prediction of effectiveness of surgical concepts, medical tests, medication, and the discovery of relationships among clinical and diagnosis data as well employ Data Mining methodologies [7]. Therefore, data mining has developed into a vital domain in healthcare. It is possible to predict the efficiency of medical treatments by building the data mining applications.

The real-life data mining applications are attractive since they provide data miners with varied set of problems, time and again. Working on cancer disease patient's databases is one kind of a real-life application. It tries to utilize the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process [1]. In the recent past, the data mining techniques were utilized by several authors to present diagnosis approaches for diverse types of cancer diseases [5, 2, 6]. This paper presents about the various

effective cancer disease prediction system using PCAR: an Efficient Approach for mining Association Rules. An efficient methodology for the generation of association rules from the cancer disease warehouses for cancer prediction has been presented. Initially, the Pima Indian cancer data warehouse is pre-processed in order to make it suitable for mining process. Once preprocessing gets over, the cancer disease data warehouse is binning with the aid of the modified equal width binning interval approach to discretizing continuous attributes. The width of the desired interval is chosen based on the opinion of medical expert and is provided as an input parameter to the process. First, we have converted numeric attributes into categorical form based on above techniques. The frequent patterns are applicable for cancer disease that are mined with the aid of the PCAR algorithm from the data extracted. In addition, the patterns vital to breast cancer prediction are selected on basis of the computed significant class labels. The techniques are trained with the selected class labels for the effective prediction of breast cancer. A lot of existing algorithms used for mining association rules identify frequent item sets by the method of bottom-up combination of smaller frequent item sets or top-down decomposing of larger infrequent item sets will results the large volumes of candidate item sets.

Actually as the supersets of infrequent items are infrequent item sets, a new efficient method as Pruning-Classification Association Rule (PCAR). PCAR combines minimum frequency items with minimum frequency item sets. It first deletes infrequent items and then classifies item sets based on frequency of item sets and discovers frequent item sets. The candidate item sets are greatly reduced and item sets need not to be combined or decomposed then operation time and memory requirement could be decreased accordingly. The significant advantage in mining association rule at large volumes of items and small frequency of item sets. The experimental results are that PCAR outperforms the well-known Apriori algorithm. Then the efficiency of the designed system in predicting the breast cancer is illustrated by the acquired results.

IV. Breast Cancer

4.1. Signs and symptoms

Early breast cancer can in some cases present as breast pain or a painful lump. Breast cancer is most frequently discovered as an asymptomatic nodule on a mammogramA lump under the arm or above the collarbone that does not go away may be present.

When breast cancer has invaded the dermal lymphatics - small lymph vessels of the skin, it can resemble skin inflammation is known as inflammatory breast cancer where it is blocking lymphatic vessels and this can cause some symptoms around the breast, as well as an orange peel texture to the skin referred to as peau d'orange. It may also have been no previous signs of breast cancer and the cancer might be missed in screening mammograms.

Changes in the appearance or shape of the breast can raise suspicions of breast cancer. One of the symptom of breast cancer is Paget's disease of the breast. It presents as eczematoid skin changes at the nipple and it is a late manifestation of an underlying breast cancer. Some breast symptoms do not turn out to represent underlying breast cancer. Some breast diseases such as fibrocystic mastopathy, functional mastodynia, mastitis and fibroadenoma of the breast are more common causes of breast symptoms. A new breast symptom should be taken seriously by both patients and their doctors by the possibility of an underlying breast cancer at almost any age.

Occasionally, breast cancer presents as metastatic disease i.e cancer that has spread beyond the organ. Metastatic breast cancer will cause symptoms that depend on the location of metastasis. Some common sites of metastasis include bone, liver, lung, and brain. Weight loss can occasionally herald an occult breast cancer as symptoms of fevers or chills. Joint pains or bone can sometimes be manifestations of metastatic breast cancer as jaundice or neurological symptoms. The uncommon symptom with metastatic breast cancer is Pleural effusions. Since these symptoms can also be manifestations of many other illnesses.

4.2 Screening

Breast cancer screening is an attempt to find cancers disease. The most common screening methods are self and clinical breast exams, x-ray mammography, Breast Magnetic resonance imaging (MRI), ultrasound, Miraluma and genetic testing.

4.2.1 X-ray mammography

Mammography is still the modality of choice for screening of early breast cancer. It is fast, accurate and widely available in developed countries. Detecting breast cancer by using mammography are usually much smaller than those detected by patients or doctors as a breast lump.

Due to the high incidence of breast cancer, screening is now recommended in many countries. Screening methods include mammography and breast self-examination. Mammography has been estimated to reduce breast cancer-related mortality by 20-30%. Routine mammography of women older than age 40 or 50 is recommended by numerous organizations as a screening method to diagnose early breast cancer and has

demonstrated a protective effect in multiple clinical trials. Mammographic screening comes from eight randomized clinical trials from the 1960s through 1980s. These trials have been criticised for methodological errors and the results were summarized in a review published in 1993.

Improvements in mortality due to screening are hard to measure; similar difficulty exists in measuring the impact of Pap smear testing on cervical cancer, then the impact of that test is likely enormous. National mortality due to cancer before and after the institution of a screening test is a surrogate indicator about the effectiveness of screening and results of mammography are favorable.

4.2.2 Breast MRI

Magnetic Resonance Imaging (MRI) has been shown to detect cancers not visible on mammograms. For example, although it is 27-36% more sensitive, it is less specific than mammography.MRI studies will have more false positives as a result, which may have undesirable financial and psychological costs. Its relatively expensive procedure which requires the intravenous injection of a chemical agent to be effective.

Fig 1: Identifying Suspicious area by using Breast MRI

The Proposed indications for using MRI screening include:

- Strong family history of breast cancer.
- Patients with BRCA-1 or BRCA-2 oncogene mutations.
- Estimation of women with breast implants
- History of previous lumpectomy or breast biopsy surgeries.
- Axillary metastasis with an unknown primary tumor.
- Very dense or scarred breast tissue.

4.3 Diagnosis

Breast cancer is diagnosed by the examination of surgically removed breast tissue. Some procedures can obtain tissue or cells prior to definitive treatment for histological or cytological examination. Those procedures includes nipple aspirates, fine-needle aspiration, core needle biopsy,ductal lavage and local surgical excision. These diagnostic steps are coupled with radiographic imaging usually accurate in diagnosing a breast lesion as cancer.Pre-surgical procedures such as fine needle aspirate may not yield enough tissue to make a diagnosis or may miss the cancer entirely. Sometimes an imaging tests are used to detect metastasis and include chest X-ray, bone scan, Cat scan, MRI, and PET scanning.Ca 15.3 is a tumor marker determined in blood which can be used to follow disease activity over time after definitive treatment. By blood tumor marker testing which is not routinely performed for the screening of breast cancer and has poor performance characteristics for this purpose.

V. Disease Prediction Using Rule Mining

The extraction of significant patterns from the cancer disease data warehouse is presented in this section. The cancer disease data warehouse contains the screening clinical data of cancer patients. Initially, the data warehouse is preprocessed to make the mining process more efficient. In our proposed study, it uses preprocessing in order to handle missing values. Then applying equal interval binning with approximate values based on medical expert advice to pima Indian breast cancer data. PCAR algorithm is applied to generate the rules. Also consider important measure confidence. Calculating the significant items for all frequent patterns with the aid of the approach proposed. The frequent patterns are selected with confidence greater than a predefined threshold. These frequent patterns can be used in the design and development of breast cancer prediction system.

5.1. Approximate equal binning techniques based on expert advice

After the preprocessing the below attributes are included. Classification of associations requires categorical data. Data variables are binned into small number of categories. The approximate equal intervals are

used for binning and also taken advice from medical experts. It also summarizes the cut-off values along with the names of the bins for the variables.

5.2 Applying Pruning-Classification Association Rule

Pruning-Classification Association Rule (PCAR). PCAR combines minimum frequency items with minimum frequency item sets. It first deletes infrequent items from item sets and then it classifies those item sets based on the frequency of item sets to discovers frequent item sets. The candidate item sets counts are greatly reduced and item sets need not to be combined or decomposed then the operation time and memory requirement could be decreased accordingly. It has significant advantage in mining association rule at large volumes of items and small frequency of item sets.

5.3 Applying Association Rules

Association rules are nothing different from classification rules except that does not predict only class labels but also predict any other attribute. It is used to produce a combination of attributes. Those different association rules convey different regularities that trigger in the dataset and generally predict the different things and so many association rule generated from even the data set is small. By keeping such rules which are applicable reasonably large number of instances based on coverage and accuracy criteria. An association rule is the number of instances for which it predicts correctly this is often called its support. Confidence is the number of instances that it predicts correctly and expressed as proportion of all instances to which it applies called accuracy. The user have to specify the minimum coverage and accuracy values and look for only those rules whose values are at least of the specified minimum value.

VI. Associative Classifier

Associative Classifier uses Weighted Association Rule Mining (WARM) model for cancer patient diagnosis. The diseases prediction process is performed by comparing the symptoms of patients with the existing database. Different attributes will be assigned different weights based on their predicting capability. The attribute with higher predicting capability will have higher weightage. The diseases prediction system is designed to forecast cancer disease severity levels. The cancer symptoms and its class labels are used in the learning process. Learned class patterns are used in the prediction process. The system is divided into four major modules. They are Data preprocess, approximation process, disease prediction and rule mining. Data preprocess module is designed to clean noisy transactions. The approximation process is designed to convert the data values into categorical values. Disease prediction module is designed to forecast the disease severity levels. Rule mining module is designed to fetch symptom patterns.

6.1 Data Preprocess

The data preprocess module is designed to import data from textual data collection. The cancer patient diagnosis details are imported and updated into Oracle database. Redundant data values are removed from the database. Incomplete data values are assigned by the system. Aggregation based data substitution technique is used for the incomplete data assignment process. Cleaned dataset is referred as optimal dataset. The training set is selected from the optimal dataset. The testing is carried out on the unlabeled transactions.

6.2 Approximation Process

The approximation process is initiated to convert data values into categorical attributes with experts advice. The data values are divided into small intervals. The interval data are assigned with category information. Classification is applied on the categorical data values.

6.3 Disease Prediction

Pruning-Classification Association Rule (PCAR) combines minimum frequency items with minimum frequency item sets. It first deletes infrequent items from item sets, and then classifies item sets based on frequency of item sets. The number of candidate item sets is greatly reduced and item sets need not to be combined or decomposed.

6.4 Rule Mining

The association rule mining technique is applied to predict the disease severity. The disease symptom rules are also identified with label information. The rules are updated and used in future prediction process. The rule base is automatically updated by the frequent patterns with its labels.

VII. Weighted Classifiers

7.1 Data Sets

The cancer disease severity prediction system is tested using the breast cancer diagnosis datasets. The dataset is downloaded from the UCI (University of California, Irwin) machine learning repository [4]. It provides information about the breast cancer patient diagnosis information. The class information and associated symptom details are provided in the dataset. The dataset is also constructed with noise records. Noise elimination process is performed on the data sets. The dataset attribute details are given in table 7.1. Classification, clustering and rule mining operations can be tested using the dataset.

TABLE 1: ATTRIBUTE DETAILS FOR BREAST CANCER DATASET

S.No.	Attribute Name	Description
1	Pid	Patient identification number
2	Ct	Clump thickness
3	UCSize	Uniformity of cell size
4	UCShape	Uniformity of cell shape
5	Ma	Marginal adhesion
6	Sece	Single epithelial cell size
7	Bn	Bare nuclei
8	Bc	Bland chromatin
9	Nn	Normal nucleoli
10	M	Mitoses
11	Class	Class

7.2 Cancer Prediction with Weighted Classifiers

The Weighted Rule Mining techniques are used to fetch the rules with frequency and weight values. The frequency based rule mining methods uses the item set count values. The support and confidence values are used in the frequency based rule mining method. The attributes are considered with same priority. In the case of weighed rule mining each attribute is assigned with a weight value. Weighted support and weighted confidence values are used in the rule mining process. The minimum support and minimum confidence values are used to filter the rules. The associative classifier based diseases prediction scheme is enhanced with weighted rule mining method. The weighted associative classifier method uses the weighted rule mining for diseases prediction process.

The cancer patient diagnosis information is used in the classification process. The rule mining methods are used to filter the rules with support and confidence values. The rule bases are used for the disease severity prediction process. The severity and associated symptom details are maintained under the rule base. The rule base is prepared with the transactions. The learning transactions are maintained with labels. The symptom patterns are identified from the labeled transactions. The prediction process is initiated with unlabeled transactions. The symptom collections are matched with the rule base details. The severity levels are assigned with weighted rule mining based pattern information.

VIII. Conclusion

New classification approach that use association rule mining and classification has become a significant tool for knowledge discovery. The association rule mining and classification techniques are integrated under the associative classification process. The weighted association rule mining is carried out on the data with class labels. By using PCAR algorithm, the system is designed to predict cancer severity levels. The system performs weighted rule mining on labeled data values. Symptom based weight assignment is performed. Support based confidence based filtering model. Disease and its severity level are predicted.

References

[1] Andreeva P., M. Dimitrova and A. Gegov, "Information Representation in Cardiological Knowledge Based System", SAER'06, pp: 23-25 Sept, 2006.

[2] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

[3] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, "Knowledge Management, Data Mining, and Text Mining In Medical Informatics", Chapter 1, eds. Medical Informatics: Knowledge Management And Data Mining In Biomedicine, New York, Springer, pp. 3-34, 2005.

[4] Newman.D, Hettich.J.S,Blake.C.L.S, and C.J.Merz, "UCI Repository of machine learning databases" Irvine, CA: University of California, Department of Information and Computer Science.1998, last accessed: 1/10/2009.

[5] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007.

[6] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008.

[7] Tzung-I Tang, Gang Zheng, Yalou Huang, Guangfu Shu, Pengtao Wang, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis", IEMS, Vol. 4, No. 1, pp. 102-108, June 2005.