ISSN No. 0976-5697

Volume 2, No. 6, Nov-Dec 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Knowledge Discovery From Medical Data:Extracting Influencing Factors Of Breast Cancer Recurrences Through Predictive Apriori Algorithm

Srinivas Murti*
Department Of Information Science and Engg
Basavakalyan Engineering College
Basvakalyan, District: Bidar, India
s.sri108@gmail.com

Gururaj R .Patwari
Department Of Information Science and Engg
Basavakalyan Engineering College
Basavakalyan, District: Bidar, India
gururaj.pat26@gmail.com

Sriranga Joshi
Computer Science and Engg Department
Appa Institute Of technology
Gulbarga, India
sriranga.joshi@gmail.com

Abstract: Breast Cancer is the leading cause of cancer deaths in women .One in nine women is expected to develop breast cancer. Breast cancer can recur at any time, but most recurrences occur in the first three to five years of initial treatment. Breast Cancer can come back as a local recurrence (in the treated Breast or near mastectomy scar) or as a distant recurrence somewhere else in the body. In this work 286 Breast Cancer patient data , obtained from UCI Machine Learning Repository are used to determine the relationship between the Breast cancer recurrences and other attributes through Predictive Apriori algorithm using WEKA(Waikato Environment for Knowledge Analysis) data mining tool

Keywords: Medical data, Data Mining, Association *Rules*, WEKA, Predictive Apriori Algorithm, Predictive Accuracy, Knowledge Discovery in Databases (KDD)

I. INTRODUCTION

Clinical Databases have Accumulated large quantities of information about patients and their Medical Conditions[1].The nature of this Medical Data is Noisy, In complete and Un Certain .Too many disease markers (attributes) are now available for decision making. Relationships and Patterns within this data could provide "New Medical Knowledge" and enhance our knowledge of disease progression and management. Evaluation of stored Medical data using tools like WEKA (Waikato Environment for Knowledge Analysis) may lead to discovery of trends and patterns. Techniques are needed to search large quantities of this Medical Data for these patterns and relationships. In this Context, Medical Data Mining came into existence.

Data Mining also referred to as Knowledge Discovery in Databases or KDD is the search for Relationships and Global Patterns that exist in the large databases but are "hidden" among the vast amount of data [2]. Data mining also refers to extracting information from very large databases [3]

Classification and Association are two mechanisms to represent extracted information [2][3].Association Rules are of type A->B where A and B are the sets of attributes (items)

The purpose of this paper is to establish relationship between Breast Cancer Recurrence Class (Whether "Breast Cancer Recurrence event" or "no Breast Cancer Recurrence event") and other attributes through Predictive Apriori algorithm(A Association Rule Mining Technique) using WEKA data mining tool. The rest of the sections are organized in the following manner – Section II discusses about Predictive Apriori algorithm .Section III gives the details of Experimental Results, Section IV analyses the results and finally section V concludes.

II. PREDICTIVE APRIORI

The major distinction between Apriori Algorithm and Predictive Apriori is their different interesting measure [5]. Apriori sorts the rules according to the confidence .if two rules have the same confidence then one with higher support is preferred. Predictive Apriori however uses Predictive accuracy .it equals a support based correction of the confidence of a rule. In every Experiment the support threshold of Apriori is set to a count that equals 1% of all instances N and the confidence threshold is set to 0.5-the standard threshold for support and confidence .The third parameter is the number of "N "mined association rules, Apriori outputs the first "N" Rules which are above the support and confidence where as Predictive Apriori outputs the "N" Best Rules.

A Rule belongs to "N" Best, if its predictive accuracy is among the "N" Best rules and there is no Rule in Best "N" which is more general and at least equally accurate

Let D be a Database whose individual records are generated by a static process P, Let $X \Rightarrow Y$ be an Association.Rule. The predictive accuracy $c(X \Rightarrow Y) = P$ r(r satisfies Y |r satisfies X) is the Conditional Probability of $Y \subseteq r$ When the distribution of r is governed by P[5]

Divide the predictive Accuracy c in 100 discrete intervals Every time the midpoint of the each interval is taken for calculation .For discrete values of c and association rule r of the form $X \Rightarrow Y$, the Predictive Accuracy is calculated by

formula below (Keeping in mind that s(r) = s(X | Y) = ˆc(r)s(X))

$$E(c(r)|\hat{c}(r), s(X)) = \frac{\sum_{c_i} c_i B[c_i, s(X)](\hat{c}(r))P(c_i)}{\sum_{c_i} B[c_i, s(X)](\hat{c}(r))P(c_i)}$$

$$= \frac{\sum_{c_i} c_i \binom{s(X)}{s(r)} c_i^{s(r)}(1-c_i)^{s(X)-s(r)}P(c_i)}{\sum_{c_i} \binom{s(X)}{s(r)} c_i^{s(r)}(1-c_i)^{s(X)-s(r)}P(c_i)} \quad (1)$$

Drawing the 1000 random association rules for each possible length. For every association rule its confidence is measured (given the support is greater zero) and a histogram $\pi_i(c)$ for every length **i** will be calculated where i Corresponds to number of items a specific association rule has. Consequently for every discretized value c, Prior distribution P(c) is given by

$$\pi(c) = \frac{\sum_{i=1}^{k} \pi_i(c)\binom{k}{i}(2^i - 1)}{\sum_{j=1}^{k} \binom{k}{j}(2^j - 1)} \quad (2)$$

Where $\pi_i(c)$ = $\frac{|\{X \Rightarrow Y \,|\, c(X \Rightarrow Y) = c\}|}{|\{X \Rightarrow Y\}|}$ (3)

The only input parameter is the number n of desired association rules. The output is a list containing the n best rules. We will refer to this list from now on as best[n]. It is implemented using a priority queue. Like apriori, predictive apriori uses frequent item sets, but the difference is that apriori de- composes a frequent item set into a rule body and a rule head and therefore the support computed for the frequent item set corresponds to the support of the whole rule. Predictive apriori, on the other hand, uses a frequent item set as a rule body and joins it with a separately computed rule head. Hence the support of a frequent item set equals the support of the rule body. The first important step in the algorithm (see Figure 1) is the estimation of the prior using equation (2).

1. Input: number of desired association rules n, database D with items $a_1, \ldots, a_k$
2. Set the support threshold of the rule body $s_{body\ min} = 1$
3. . **For i** $= 1, \ldots, k$ **DO:**
   Construct a number of association rules of length **i** at random and measure their confidence $\hat{c}$ provided $s(X) > 0$
   Let $\pi_i(c)$ be the distribution of confidences
4. **For** all c, compute $\pi(c)$ using equation (2)
5. Let $F_0 = \{\emptyset\}$ be the set of frequent item sets of length 0
6. **For i** $= 1$, TO K-1 **Do: While (i** $= 1 || F_{i-1} = \emptyset$)
   (a)Determine all frequent item sets X of length **i** with $s(X) > s_{body\ min}$

   (b)For all $X \in F_i$ call **RuleGen**(X)

   (c)If best[n] has changed in RuleGen **Then** Increase $s_{body\ min}$ so that $(c|1, s_{body\ min}) > E(c(best[n])|\hat{c}(best[n]), s(best[n]))$.

   (d) **If** $s_{body\ min} >$ size of database D **Then** Exit

   (e) **If** $s_{body\ min}$ has been increased in step6(d)**Then** Delete all item sets X from F_i with $s(X) < s_{body\ min}$

Figure 1 The Predictive Apriori algorithm

We use a dynamically increasing support of the rule body $s_{body\ min}$ starting with threshold 1. We loop over the length of the frequent item sets, while they are non empty (step 6).Thus in the first iteration all frequent item sets of length 1 with $s_{body\ min}$=1 are constructed. In subsequent passes the frequent which have at least $s_{body\ min}$ are constructed in step6 (a). For all these items we call the rule generation procedure which is explained more detail below. If best[n] changes during the rule generation step we increase $s_{body\ min}$ .Equation 1 used to determine of perfect confident rule (a rule with confidence 1) must have to get into best[n] (step6 (c)). This is our new threshold for **$s_{body\ min}$**. If the new minimum support is greater than number of instances in the dataset, the algorithm terminates. If the support threshold has increased, all items from a frequent item sets which have support below minimum support are deleted. In step 6(b) we call the rule generation procedure which receives as input one frequent item set .Figure 2shows the Pseudo code of the procedure.

RuleGen(X) finds the best rules with rule body X

10. Set $s_{rule\ min}$ so that
    $E(c|s_{rule\ min}/s(X), s(X)) > E(c(best[n])|\hat{c}(best[n]), s(best[n]))$
11. **For** $j = 1, \ldots, k - |X|$ (number of items not in X) **Do**
    (a) **If** $j = 1$ **Then**
    Set $Y_1 = \{\{a\}|a \in \{a_1, \ldots, a_k\}, a \notin X\}$
    Else generate Y_j analogous to the generation of candidate item sets.
    (b) For all $y \in Y_j$ **Do**
    i. Calculate $s(X \cap y)$.
    ii. **If** $s(X \cap y) \leq s_{rule\ min}$ **Then**
    Delete y from Y_j and continue with the next y at 11b.
    iii. Calculate the predictive accuracy of $X \Rightarrow y$ using equation (2)
    iv. **If** the predictive accuracy of $X \Rightarrow y$ is among the best n AND
    (there is no other rule in best[n] which is at least equally accurate AND
    which subsumes $X \Rightarrow y$) **Then**
    update best[n],
    Remove rules which are subsumed by other at least equally accurate rules. Set $s_{rule\ min}$, so that
    $E(c|s_{rule\ min}/s_{body\ min}, s_{body\ min}) \geq E(c(best[n])|\hat{c}(best[n]), s(best[n]))$.
    12. **If** any rule has been removed out of best[n] in step 11(c)iv **Then** recur from step 10

Figure 2 -The rule generation method [6]

III. EXPERIMENTAL RESULTS

A. About the Dataset:

The medical Data Records are collected from UCI machine Learning Repository. The data set in our study consists of 286 cases, of which 201 instances are of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.The attributes are as follows

1. Class: no-recurrence-events, recurrence-events

2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
3. menopause: lt40, ge40, premeno.
4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
6. node-caps: yes, no
7. deg-malig: 1, 2, 3.
8. breast: left, right.
9. breast-quad: left-up, left-low, right-up, right-low, central
10. irradiat: yes, no

B. Using WEKA(Waikato Environment for Knowledge Analysis) for Mining the Dataset:

Weka was created by researchers at the University of Waikato in New Zealand. It's a collection of open source of many data mining and machine learning algorithms, including[7]

a. pre-processing on data
b. Classification
c. Clustering
d. Association rule extraction

The Steps of Data Mining using WEKA for finding out the Association Rules are as follows

Step 1- Click the Explorer on WEKA GUI

Step 2-: On the Explorer Window, Click button open file to open data file from where Breast cancer data file is stored in ARFF (attribute relation file format) format.

Step3- After loading a data file click, associate and under Associate Choose Predictive Apriori algorithm algorithm and click start .The fig 3 shows Results when Predictive Apriori Algorithm is applied over Breast Cancer Data Set all Generated Rules are shown in the output pane.

The implementation follows expect for adding a rule to the output of the 'n' best rules. A rule is added if: the expected predictive accuracy of this rule is among the 'n' best and it is not subsumed by a rule with at least the same expected predictive accuracy [6]

IV. RESULTS ANALYSIS

The mining rules are shown in Fig 3, which sort on the basis of Predictive Accuracy. A Significant part of Association Rules are as follows:

Rule 1: age=40-49 node-caps=yes breast=left ==>Class=recurrence-events acc:(0.98356)

Rule 2: node-caps=yes breast-quad=right_low ==>Class=recurrence-events acc:(0.9796)

Rule 3: age=30-39 tumor-size=35-39 ==> Class=recurrence-events acc:(0.96405)

Rule 4: menopause=premeno inv-nodes=15-17 ==> Class=recurrence-events acc:(0.96405)

Rule 5: menopause=premeno tumor-size=15-19 node-caps=yes 3==> Class=recurrence-events 3 acc:(0.96405)

Rule 6: age=40-49 tumor-size=15-19 inv-nodes=0-2 breastquad=left_up 2 ==> Class=recurrence-events acc :(0.94778)

Rule 1 means that, the studied patients whose age is 40 to 49 and whose node caps are present and whose tumor location is at the left breast have the probability of recurrence (predictive accuracy=98.35%).

Rule 2 means that, the studied patients whose node caps are present and whose Quadrant location is at the right lower part of the breast have probability of recurrence (predictive accuracy -97%).

Rule 3 means that, the studied patients whose age is 30 to 39 and tumor size is 35-39 have possibility of getting breast cancer recurrence (predictive accuracy 96.40%).

Rule 4 means that patients who are Pre-Menopausal and whose tumor size is 15 to 19 have possibility of breast cancer recurrence is 96.04% (Predictive Accuracy).

Rule 6 can be interpreted as the patients whose age is 40 to 49 and whose tumor size is 15 to 19(mm) and whose inv-nodes (number of Lymph nodes Invasion) are 0 to 2 and location of quadrant tumor(breast quad) is at the left upper part of the breast have possibility of recurrence with predictive accuracy as 94.778%.

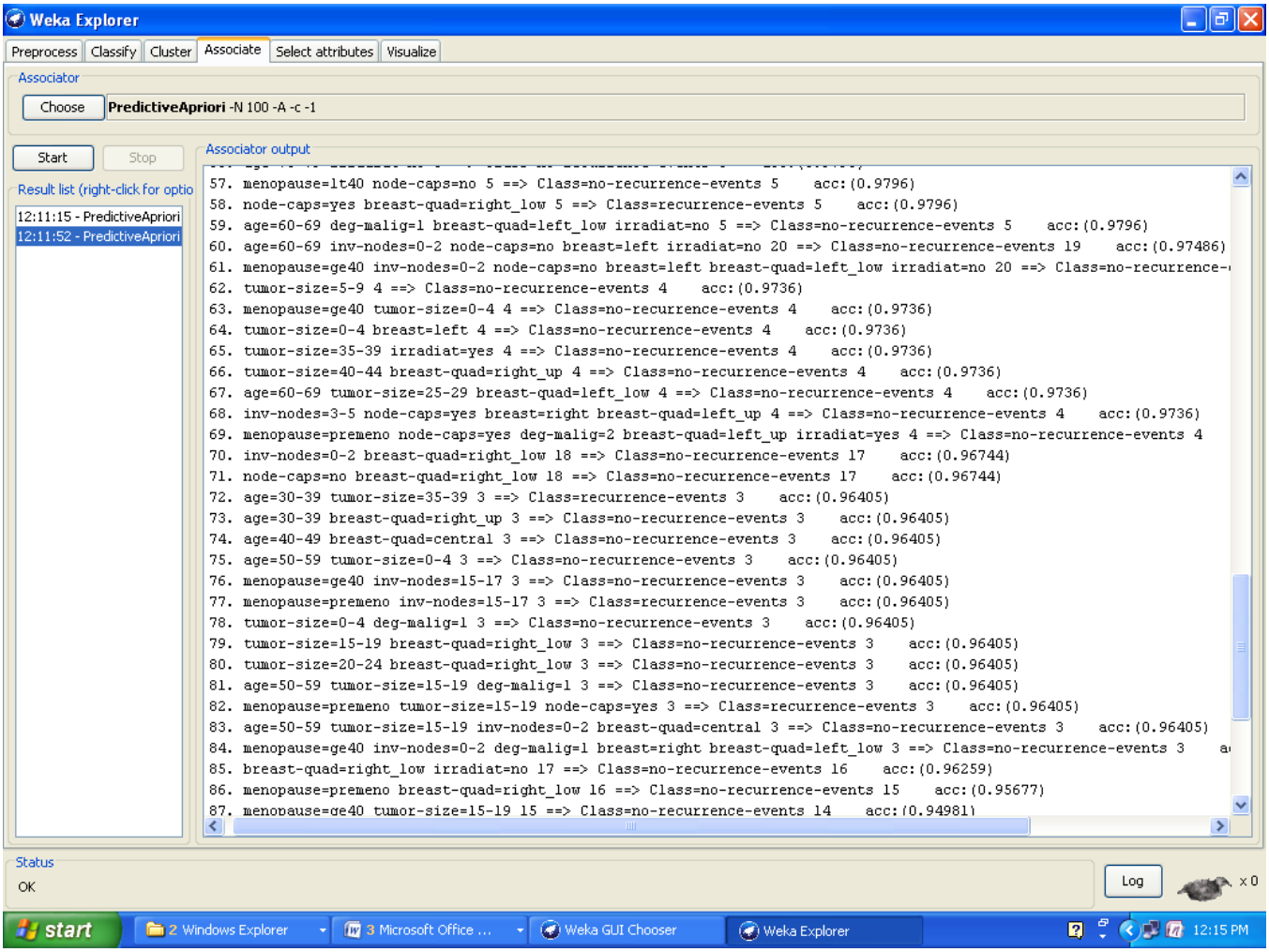

Figure.3-All Generated Rules shown in the output pane

V. CONCLUSION

In this Paper "the need for data mining in Medicine "is discussed". The characteristics of the Predictive Apriori algorithm are reviewed and the complete algorithm is presented, the same is implemented in WEKA (Waikato Environment for Knowledge Analysis) for Extracting Relationship between the Breast Cancer Recurrence Class and other attributes. The implementation follows expect for adding a rule to the output of the 'n' best rules. A rule is added if: the expected predictive accuracy of this rule is among the 'n' best and it is not subsumed by a rule with at least the same expected predictive accuracy. Conclusions are made on the Association Rules Generated by WEKA. These Results can be used for Generating Medical Hypothesis for predicting and preventing Breast cancer Recurrence, which is the leading cause of cancer deaths in women

VI. REFERENCES

[1] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond:"Medical data mining: knowledge discovery in a clinical data warehouse" Proc AMIA Annual Fall Symposium, 1997: pp.101–105

[2] Feelders, A., Daniels, H. and Holsheimer, M. "Methodological and Practical Aspects of Data Mining', Information and Management", 2000:pp.271-281.

[3] Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining:Concepts and Techniques 3rd edition, Morgan Kaufmann,2011.
Roy Levy"Pharmaceutical Industry: A discussion of legislative and Antitrust issues in an environment of Change", Federal trade commission report ,March 1999:pages 1-211

[4] Scheffer T. "Finding Association Rules That Trade Support Optimally against Confidence". In L. De Raedt and A. Siebes, editors, Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg, Germany, September 2001. Springer-Verlag :pp.424–435,

[5] Scheffer T. " Finding Association Rules That Trade Support Optimally against Confidence- algorithm's pruning method changed" Unpublished

[6] http://www.cs.waikato.ac..nz/ml/weka,accessed 02/10/2011

[7] Sheng, O. R. Liu 'Decision support for healthcare in a new Information age', Decision Support Systems, (2000) pp101-103