Engineering and Scientific International Journal (ESIJ)
Volume 4, Issue 2, April – June 2017
ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

Predicting the Risk Factors of Endometrial Cancer using Data Mining

A. Hency Juliet[#1], Dr.R. Padmajavalli[*2]

[1]*Professor in Department of Computer Application, Mar Gregorios College, Chennai*
[2]*Associate Professor in Department of Computer Application, Bhaktavatsalam Memorial College for Women, Chennai*

Abstract— Data mining act as an imperative part for uncovering new idea in healthcare organization which is supportive for all the parties related with medical field. This paper analyses the effectiveness of Data mining technique in healthcare domain. Cancer is one among the foremost crisis today, diagnosing cancer in earlier period is still challenging for doctors. Detection of hereditary and ecological aspect is very essential in developing novel methods to perceive and stop cancer. Endometrial cancer is one of the most general feminine gynaecologic malignancy, is naturally a curable disease. It is the most wide-ranging of the entire cancers and the main reason for the cancer fatality in women worldwide. This paper also presents an study of the risk factors related with endometrial cancer by means of association rule mining. Here we applied Apriori algorithm to uncover the associations. Women who are extensively heavy weight, hypertension and more estrogens level are increased risk of certain cancers. Heavy weight, hypertension, and more estrogens level were drastically related with an increased risk of endometrial cancer.

Keywords— *Data Mining; Association; Endometrial; Apriori; Healthcare.*

1. Introduction

Data mining is the manner of take out the useful data from the massive dataset [1]. Using Data mining techniques large volumes of data are handled to determine veiled outline and relations helpful in decision making [2]. A range of algorithms and techniques of data mining can be used in medical domain, so that the patients' data can be analyzed and the factors that cause the syndrome can be easily traced [3]. Nowadays using this people can aware of any diseases. Also they can be alert on particular syndrome.

Endometrial cancer is a cancer that occurs from the endometrium, the inside layer of the uterus or womb. It is the effect of the anomalous development of cells that have the ability to occupy or spread to other parts of the body [4]. Hormones changes the endometrium during the woman's menstrual cycle [5]. In the early stage of the cycle, prior to release eggs from ovaries, the ovaries create estrogen hormones. Estrogen is the reason for the endometrium to condense so that it could cultivate an embryo if pregnancy happens. A woman's hormone dependability is an important function in the maturity of most endometrial cancers. The change in estrogen level is depends on the hormone. Many risk factors for endometrial cancer influence estrogen levels [5]. The Knowledge step of classification is used to originate a classification model and a categorization step used to estimate the class labels for a given data [1]. It provide as an evocative form, to discriminate among objects of unlike classes. A Classification model can also serve in prognostic modeling, to determine the class label of unidentified records. This development is primarily appropriate for describing data sets with dual or diminutive types [6]. It is a methodical approach to accumulate a classification models from the input data set [7]. It includes Bayesian, Meta-learning, Lazy, Rule-Based, Decision-Tree and Miscellaneous classifiers. Each method exploits a learning algorithm to identify a model that preeminent fits the liaison between the attribute set and class label of the input data. An vital point of the learning algorithm is to construct the representation with generalization capability i.e., the depiction specifically forecast the class labels of formerly unidentified instances. Clustering is a data mining technique to find groups of objects [8]. In healthcare domain, clustering has been used to group patients according to their symptoms [9]. Association analysis is the innovation of relationship policy showing attribute-value situations that take place frequently in-cooperation in a given set of data. In a transactional database each time the customer purchasing details may be matched with other customer, which can be identified and analyzed using this.

1.1 APRIORI Algorithm

The association rule creation has two steps:

- Find the recurrent item sets in a database by applying minimum support to the candidate item set.
- Construct rules using recurrent item sets and the minimum confidence constraints.

1.1.1 Constructive Terms

Usually large number of rules will generate, from this interesting patterns which are useful to our field are measured. The final rules are the buck sum thresholds on support and confidence.

1.1.2 Support

The supp(X) of an item set X is defined using the below formula. Supp(X) = No. of transactions which include the item set X / Total No. of transactions .

1.1.3 Confidence

To find the confident of the rule, Confidence(X→Y) =supp (XUY)/supp (X)

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

2. Related Work

This section analyses the correlated work on cancer information using data mining algorithms. Shweta Kharya et al. [10] précised different review and technical articles on breast cancer analysis and forecast also they focus on current research being carried out using the data mining techniques to augment the breast cancer analysis and forecast and proposed decision tree algorithm is the best predictor with 93.62%.

Ramya rathan et al. [11] related association rule mining algorithm and classification technique on endometrial cancer data and found when the woman having more than 60 years of age, menopause, vegetarian diet and when these factors are involved the occurrence of cancer is high. Also the occurrence of the disease is not due to heredity.E. Friberg et al. [12] examined the diabetes (largely type 2) and endometrial cancer, based on different learning together with 96,003 partakers and 7,596 cases of endometrial cancer. Some of the learning shows a statistically considerably augmented threat and few a - significant increased threat of endometrial cancer. Their meta-analysis results that diabetes was statistically considerably related with an augmented threat of endometrial cancer. Girija D.K et al. [13] summarizes plentiful data mining techniques, review and technical articles on fibroid finding and forecast and tend to present an delineate of the current research being carried out using the data mining techniques to emphasize the fibroid finding and forecast. The algorithms C4.5, Naive Bayes, ID3 were used and Naïve Bayes gives the highest accuracy 96%.

Elisabete Weiderpass et al. [14] conceded out a population-based, nationwide, case-control study with postmenopausal women aged 50-74 years and contrasted lean women with flabby women had a 50% increase in risk for endometrial cancer. Obesity and diabetes mellitus are associated with endometrial cancer risk. Parazzini.F et al. [15] confirmed that insulin dependent diabetes is related with the risk of endometrial cancer. Connection endures subsequent to taking into report the probable mystifying effect of recognized risk factors for endometrial cancer, particularly overweight, and is consistent diagonally level of main identified covariates. This relationship may be related to prominent estrogen levels in diabetic women.

I. P. Constantinou et al. [16] presents an incorporated structure for underneath the identification of endometrial cancer. It has electronic a patients' record that encloses a system for the early detection of endometrial cancer. The maximum percentage of accurate classifications score for the SVM classifier was 79% & C4.5 model were also maintain with classification rules. Huiqiao Gao, Zhenyu Zhang et al. [17] analytically differentiate the appearance of endometrial cancer (EC) associated genes and to study the purpose, pathways, and networks of EC-linked hub proteins. The data also help to expose the molecular system of EC progress provide suggestion for targeted therapy for EC. C. Kalaiselvi et al. [18] authenticates classifiers effectiveness for the forecasting of cancer and heart disease in diabetic patients. To progress the classification accuracy and to attain better competence a new approach is proposed. Sally Yepes et al. [19] Identify and confirm cancer subtypes, summarize a variety of methodological principles, and highlight representative studies. Cancer sub typing schemes obtained by machine learning strategies and the use of clinically characterized cohorts are contributing to a better understanding of the molecular heterogeneity of cancer.

Priyanka.A et al. [20] recommended the cancer forecast scheme based on data mining. This scheme approximates the risk of the various cancers. This scheme is validated by contrasting its expected results with patient's prior medical information. Ramachandran et al. [21] used classification algorithm to classify the pattern and Clustering algorithm to subdivide the cancer into six types.

3. Materials and Methods

3.1. Dataset Description

The endometrial cancer dataset is collected from the article wrote by Breslow, N.E. and Day, N.E. [22]. The dataset contains 315 instances & 10 attributes with data on cases and controls from the Leisure World study of endometrial cancer as related to treatment with estrogens for menopausal symptoms and other risk factors. Endometrial cancer dataset' description is given in table-1. This study uses 30 patients' details for finding the associations between the symptoms, shown in table-2. That is the patients who have the positive symptoms in obese, hypertension, gall bladder disease, estrogen level and estrogen are considered. The Apriori algorithm is used to find the association between the symptoms, which leads to analyze the risk factors of endometrial cancer.

Table 1: Endometrial Cancer Data Description

Name	Description	Codes/Range
SET	Matched set indicator	1-63
CASE	Case-control indicator	0 = Control, 1 = Case
AGE	Age in years	55-83
GALL	Gallbladder disease	0 = No, 1 = Yes
HYP	Hypertension	0 = No, 1 = Yes
OB	Obesity	0 = No, 1 = Yes; 9 = Unknown
EST	Estrogens usage	0 = No, 1 = Yes
DOSE	Dose of conjugated	0 = 0 1 = 0.3 2 = 0.301-0.624 3 = 0.625 4 = 0.626-1.249 5 = 1.25 6 = 1.26-2.50

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

		9 = Unknown
DUR	Duration of estrogens use (months)	0-95 96=96+ 99=Unknown
NON	NON-estrogens drug	0= No, 1 = Yes

Table 2: Endometrial Cancer Data

CASE	AGE	GALL	HYP	OBS	EST
Case	74	No	No	Yes	Yes
Control	67	No	No	No	Yes
Case	76	No	Yes	Yes	Yes
Control	70	Yes	No	No	Yes
Case	69	Yes	No	Yes	Yes
Case	70	No	Yes	Yes	Yes
Case	65	Yes	No	No	Yes
Case	68	Yes	Yes	Yes	Yes
Control	61	No	No	Yes	No
Case	64	No	No	Yes	Yes
Case	68	Yes	No	Yes	Yes
Case	74	No	No	No	Yes
Case	67	Yes	No	Yes	Yes
Case	62	Yes	No	No	Yes
Case	71	Yes	No	Yes	Yes
Case	83	No	Yes	Yes	Yes
Case	70	No	No	Yes	No
Case	74	No	No	No	Yes
Control	70	No	Yes	No	Yes
Case	66	No	Yes	Yes	Yes
Case	77	No	No	Yes	Yes
Case	66	No	Yes	No	Yes
Case	71	No	Yes	Yes	Yes
Case	80	No	No	No	Yes
Case	64	No	No	Yes	Yes
Case	63	No	No	No	Yes
Case	72	Yes	Yes	Yes	No
Case	57	No	No	No	Yes
Case	74	Yes	No	Yes	No
Case	62	No	Yes	Yes	Yes

3.2. Overview of Endometrial Cancer

The main proverbial type of ADENO carcinoma is referred as endometriosis cancer. Endometriosis cancers are created by cells in glands that appear a lot like the usual uterine inside layer (endometrial). Endometrial cancer is classified into two categories [23]. Type-I endometrial cancers are deliberation to be origin by surfeit estrogen. They industrialized from an archetypal hyperplasia occasionally, an anomalous augmentation of cells in the endometrium. These cancers are typically not very destructive and are slow to widen to new tissues. Type-II endometrial cancers create a small number of endometrial cancers. It doesn't give the impression to be cause by a huge quantity of estrogen [24].

4. Results and Findings

In this study among 315 instances, 30 patients' data have been taken as sample. We have established the alliance between the symptoms occurred in the patients. The frequent-item-sets are set up by the frequent incidence of positive symptoms of each patient. List of patients have positive results are listed in table-3. Then the rules are reveal which can be obtained from data bases. The minimum support count taken as 30%, and minimum confident rate as 70%.

Table 3: List of patients has positive results

Set	List Of Patients Have Positive Results On The Following Symptoms
1	OBS, EST
2	EST
3	HYP,OBS,EST
4	GALL,EST
5	GALL,OBS,EST
6	HYP,OBS,EST
7	GALL,EST
8	GALL,HYP,OBS,EST
9	OBS
10	OBS,EST
11	GALL,OBS,EST
12	EST
13	GALL,OBS,EST
14	GALL,EST
15	GALL,OBS,EST
16	HYP,OBS,EST
17	OBS
18	EST
19	HYP,EST
20	HYP,OBS,EST

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

21	OBS,EST
22	HYP,EST
23	HYP,OBS,EST
24	EST
25	OBS,EST
26	EST
27	GALL,HYP,OBS
28	EST
29	GALL,OBS
30	OBS,EST

The 1–item set from the database has obtained for 5 symptoms. These obtained data represents C_1 candidate item set of Apriori algorithm. Table-4 shows the candidate item set.

Table-4 C_1 Candidate item set

Symptom	Support Count
GALL	10
HYP	09
OBS	19
EST	24

Then, the number of times, each of the five symptoms repetitive has been obtained. The frequent-item-sets L1 be able to find from table-4. It contains candidate-1 item-sets fulfilling the minimum-support-count. In this analysis the minimum support count is taken as nine. Match up to candidate-support-count with minimum-support-count and produce the frequent item set L1. It is listed in table -5.

Table-5 L_1 frequent item set

Symptom	Support Count
GALL	10
HYP	09
OBS	19
EST	24

By means of L_1, we have produced C_2. There are four frequent-items in L_1. Then the number of pairs in C_2 should be 4 * 3 / 2 = 6. It is also specified as $L_1 * L_1$, where * is a special concatenation operation. Generate C2 candidate item from L1. Number of pairs in C2 should be n * (n-1)/2 that is 4*3/2=6, L1 X L1. Scan the table-3 for count of each candidate. The C2 candidate-item-set is represented in table-6.

Table 6: C2 candidate item set

Symptom	Support count
GALL X HYP	02
GALL X OBS	07
GALL X EST	07
HYP X OBS	07
HYP X EST	08
OBS X EST	15

We have examined the database once and established the support count of every combination. From the table-6, compare candidate support count with minimum support count which is nine, and generate the frequent item set L2, and represent in table -7. There are four item sets in L_2.

Table 7: L2 Frequent Item Set

Symptom	Support count
OBS X EST	15

4.1. Finding the rules

To find the confident of the rule, Confidence (A→B) =P (B/A) =P (A∩B)/P (A)

Discovering Rule (Association rule generation) for the frequent-item-set X= {OBS, EST}. The association rules can be produced from X is, the empty subsets of X are {OBS, EST}, {OBS}, {EST}, {}

Variety of rules acquired from L3 and the confidence rate in percentage specified in the table-8.

Table 8: Confidence rate for rules

Rule	Confidence
OBS → EST	15/19 = 78%

The resulting Association rules are OBS → EST, Confidence = 15/19 =78% If the minimum confidence threshold is 70% then the above one only output, since these are produced only once and that are well-built.

5. Conclusion

When the rule is observed, we got the results that, 78% of patients who are put up with obesity and more estrogen level has the possibility of having the endometrial cancer. From the end of Apriori algorithm, the women who are having more estrogens level and considerably overweight are increased risk of endometrial cancers. Many results affirm that Obese and overweight women have additional opportunity to have the risk of budding endometrial cancer disease than women of a normal weight, even with the consequences of menopausal kind. Also through this study it was observed that the women who are considerably obese and who affected by increased estrogens level, can protect themselves from this are recommend to use the non estrogens drugs.

References

[1] J.Han, M.Kamber, J.Pei. Data Mining concepts and Techniques. 3rd Edition, Simon Fraser University.

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

[2] F.Angiulli and C. Pizzuti, "Outlier Mining in Large High-Dimensional Data Sets", IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 2, pp. 203-215, 2005.
[3] Christopher.T. (2016) study of Classification prediction for Lung cancer prediction, international Journal of Innovative Science, Engineering & Technology, Vol. 3 (2), ISSN 2348 – 7968.
[4] Faysal A Saksouk, MD; Chief Editor: Eugene C Lin, MD. Endometrial Carcinoma Imaging.
[5] A detailed guide – endometrial cancer. www. cancer. org/ cancer/ endometrial cancer/cancer-risk-factors.
[6] Hency Juliet, Dr.R. Padmajavalli, Comparative Analysis Of Classification Algorithms On Endometrial Cancer Data, Indian Journal Of Science And Technology, Volume 9(28),DOI: 10.17485 /IJST/2016/v9i28/93846, July 2016.
[7] Classification basic concepts. E-Book. Chapter 4. Pages 145 – 149.
[8] Faramarz Karamizadeh and Seyed Ahad Zolfagharifar. Using the Clustering Algorithms and Rule-based of Data Mining, Indian Journal of Science and Technology. 2016 February; 9 (7).
[9] Hency Juliet, Dr. R. Padmajavalli, Comparative Analysis Of Clustering Algorithms On Endometrial Cancer Data, International Conference on Viable Synergies in Mathematical and Natural Sciences.
[10] Shweta Kharya, Using Data Mining Techniques For Diagnosis And Prognosis Of Cancer Disease, International Journal Of Computer Science, Engineering And Information Technology, Vol.2, No.2, April 2012.
[11] Ramya Rathan, Sridhar R, Balasubraminian S, Association Rule-Spatial Data Mining Approach For Exploration Of Endometrial Cancer Data, International Journal Of Advanced Research In Computer Science And Software Engineering, Volume 3, Issue 10, October 2013.
[12] E. Friberg & N. Orsini & C. S. Mantzoros & A. Wolk, Diabetes Mellitus And Risk Of Endometrial Cancer: A Meta-Analysis, Springer-Verlag 2007.
[13] Smt Girija D.K, Dr. M.S. Shashidhara, Data Mining Techniques Used for Uterus Fibroid Diagnosis and Prognosis, ©2013 IEEE.
[14] Weiderpass E, Persson I, Adami Ho, Magnusson C, Lindgren A, Baron Ja, Body Size In Different Periods Of Life, Diabetes Mellitus, Hypertension, And Risk Of Postmenopausal Endometrial Cancer, Cancer Causes And Control - Published By Springer.
[15] Parazzini F, La Vecchia C, Negri E, Diabetes And Endometrial Cancer: An Italian Case-Control Study. International Journal For Cancer, 1999, 81:539–542,
[16] I. P. Constantinou, C. A. Koumourou, M. S. Neofytou, V. Tanos,C. S. Pattichis, E.C. Kyriakou, An Integrated Cad System Facilitating The Endometrial Cancer Diagnosis, Information Technology And Applications In Biomedicine, Larnaca, Cyprus, 5-7 November 2009.
[17] Huiqiao Gao And Zhenyu Zhang, Systematic Analysis Of Endometrial Cancer-Associated Hub Proteins Based On Text Mining, Biomed Research International, Volume 15 (2015).
[18] C. Kalaiselvi And G. M. Nasira, Prediction Of Heart Diseases And Cancer In Diabetic Patients Using Data Mining Techniques, Indian Journal Of Science And Technology, Vol 8(14), Ipl 018, July 2015.
[19] Sally Yepes and Maria Mercedes Torres, Mining Datasets For Molecular Subtyping In Cancer, Journal Of Data Mining In Genomics & Proteomics Data Mining Genomics Proteomics 2016.
[20] A. Priyanga, Dr.S.PRAKASAM, Effectiveness of Data Mining - based Cancer Prediction System, International Journal of Computer Applications (0975 – 8887) Volume 83 – No 10, December 2013.
[21] P.Ramachandran, N.Girija, Bhuvaneswari, Early Detection and Prevention of Cancer using Data Mining Techniques, International Journal of Computer Applications Volume 97– No.13, July 2014.
[22] The Cancer Genome Atlas, A pilot project of the National Cancer Institute.
[23] The Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Endometrial Carcinoma. Nature, May 2, 2013.
[24] A detailed guide – endometrial cancer. http://www.cancer.org /cancer /endometrialcancer/what-is-endometrial-cancer.