International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 11 55 – 60

Breast Cancer Prediction using Data Mining Techniques

Jyotsna Nakte
Student, Dept. of Information Technology
MCT Rajiv Gandhi Institute ofTechnology
Mumbai, India
Email: jyotsnanakte26@gmail.com

Varun Himmatramka
Student, Dept. of Computer Engineering
MCT Rajiv Gandhi Institute of Technology
Mumbai, India
Email: mail.varun09@gmail.com

***Abstract*—**Cancer is the most central element for death around the world. In 2012, there are 8.2 million cancer demise worldwide and future anticipated that would have 13 million death by growth in 2030.The earlier forecast and location of tumor can be useful in curing the illness. So the examination of methods to recognize the event of disease knob in early stage is expanding. Prior determination of Breast Cancer spares tremendous lives, falling flat which may prompt to other extreme issues bringing on sudden lethal end. Its cure rate and expectation depends chiefly on the early identification and finding of the infection. A standout amongst the most well-known types of therapeutic acts of neglect internationally is a blunder in determination.Today,there is huge usage of data mining techniques and process like Knowledge discovery development for prediction. Significant learning can be found from utilization of information mining methods in social insurance framework. In this study, we quickly look at the potential utilization of arrangement based information mining systems, for example, Decision tree classification to huge volume of human services information. The social insurance industry gathers tremendous measures of medicinal services information which, shockingly, are not "mined" to find shrouded data. In this method we make use of see5 algorithm and k-means algorithm for prediction.

Keywords*—*Data mining; breast cancer; classification techniques; k-means algorithim; see5 algorithm.

*****

I. INTRODUCTION

Breast cancer begins with forming tumour in cells of breast forming clusters and spreads in the entire tissue. In this article, we especially focus on ladies bosom growth and procedures for early forecast.

Symptoms of Breast Cancer:

- A lump in a breast
- A pain in the armpits or breast that does not seem to be related to the woman's menstrual period
- Pitting or redness of the skin of the breast; like the skin of an orange
- A rash around (or on) one of the nipples
- A swelling (lump) in one of the armpits
- An area of thickened tissue in a breast
- One of the nipples has a discharge; sometimes it may contain blood
- The nipple changes in appearance; it may become sunken or inverted
- The size or the shape of the breast changes
- The nipple-skin or breast-skin may have started to peel, scale or flake.

Types of breast cancer classified basicallymalignant tumor and benign tumor. The choice of the classification technique is very much important as the accuracy of the classification as malignant or benign varies from algorithm to algorithm. The project involves the use of different classification techniques to classify a particular cancer as malignant or benign. The result is later compared and the most accurate of them is selected. This will help to determine which algorithm is best fit for a particular set of images.

II. LITERATURE SURVEY

C4.5 is a notable choice tree acceptance learning system which has been utilized by AbdelghaniBellaachia and ErhanGauven alongside two different strategies i.e. naive Bayes and Back-Propagated Neural Network. They exhibited an examination of the expectation of survivability rate of breast growth patients utilizing above information mining systems and utilized the new form of the SEER Breast Cancer Data. The preprocessed information set comprises of 151,886 records, which have all the accessible 16 fields from the SEER database. They have received an alternate approach in the pre-grouping process by including three fields: STR (Sarvival Time Recode), VSR (Vital Status Recode), and COD (Cause Of Death) and utilized the Weka toolbox to explore different avenues regarding these three information mining calculations. A few investigations were led utilizing these calculations. The accomplished forecast exhibitions are practically identical to existing procedures. In any case, they discovered that model created by C4.5 calculation for the given information has a vastly improved execution than the other two methods. [1]

Wei-stick Chang, Der-Ming and Liou investigated that the hereditary calculation display yielded preferable results over other information digging models for the examination of the information of breast tumor patients as far as the general

exactness of the patient arrangement, the expression and many-sided quality of the classification rule. The artificial neural system, decision tree, calculated relapse, and hereditary calculation were utilized for the similar studies and the precision and positive prescient estimation of every calculation were utilized as the assessment pointers. WBC database was joined for the information investigation took after by the 10-overlap cross-approval. The outcomes demonstrated that the hereditary algorithm portrayed in the study could deliver exact results in the order of breast disease information and the arrangement run recognized was more adequate and intelligible. [2]

Labeed K Abdulgafoor et al wavelet change and K-means clustering calculation have been utilized for intensity based segmentation. [4]

Sahar A. Mokhtar et al have examined three diverse classification models for the forecast of the seriousness of breast masses in particular the choice tree, simulated neural system and bolster vector machine.[5]

Rajashree Dash et al a hybridized K-mean algorithm has been proposed which consolidates the means of dimensionality decrease through PCA, a novel introduction approach of group focuses and the means of doling out information focuses to proper clusters. [6]

Ritu Chauhan et al concentrates on clustering algorithm,, for example, HAC and K-Means in which, HAC is connected on K-means to decide the quantity of bunches. The nature of bunch is enhanced, if HAC is connected on K-means. [3]

III. FEATURES OF DATA MINING PROCESS

Knowledge Discovery and Data Mining
This area gives a prologue to knowledge discovery and data mining. We list the different investigation assignments that can be objectives of a discovery procedure and records strategies and research zones that are promising in fathoming these examination errands.

The Knowledge Discovery Process

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low- level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The following fig. 1 shows data mining as a step in an iterative knowledge discovery process.

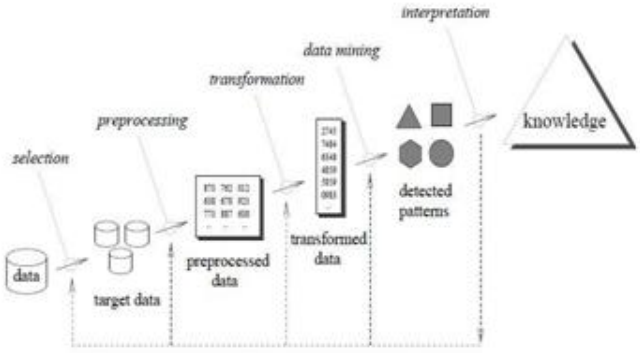

Fig. 1 Steps in KDD

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge [8]. The iterative process consists of the following steps:

(1) *Data cleaning:* also known as data cleansing it is a phase in which noise data and irrelevant dataare removed from the collection.
(2) *Data integration*: at this stage, multiple data sources, often heterogeneous, may be combined in acommon source.
(3) *Data selection:* at this step, the data relevant to the analysis is decided on and retrieved from the datacollection.
(4) *Data transformation:* also known as data consolidation, it is a phase in which the selected data istransformed into forms appropriate for the mining procedure
(5) *Data mining:* it is the crucial step in which clever techniques are applied to extract patterns potentiallyuseful.
(6) *Pattern evaluation:* this step, strictly interesting patterns representing knowledge are identified basedon given measures.
(7) *Knowledge representation:* is the final phase in which the discovered knowledge is visually representedto the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

Data Mining Process

In the KDD procedure, the data mining techniques are for extricating patterns from data. The patterns that can be found rely on the data mining errands connected. For the most part, there are two sorts of data mining errands: graphic data mining undertakings that portray the general properties of the current data, and prescient data mining assignments that endeavor to do expectations in view of accessible data. Data mining should be possible on data which are in quantitative, literary, or sight and sound structures.

Data mining applications can utilize distinctive sort of parameters to look at the data. They incorporate affiliation (patterns where one occasion is associated with another occasion), succession or way investigation (patterns where one occasion prompts to another occasion), classification (recognizable proof of new patterns with predefined targets) and bunching (gathering of indistinguishable or comparable objects).Data mining includes a portion of the accompanying key steps [9]-

(1) *Problem definition:* The first step is to identify goals. Based on the defined goal, the correct series oftools can be applied to the data to build the corresponding behavioral model.

(2) *Data exploration:* If the quality of data is not suitable for an accurate model then recommendations onfuture data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.

(3) *Data preparation:* The purpose of this step is to clean and transform the data so that missing *andinvalid* values are treated and all known valid values are made consistent for more robust analysis.

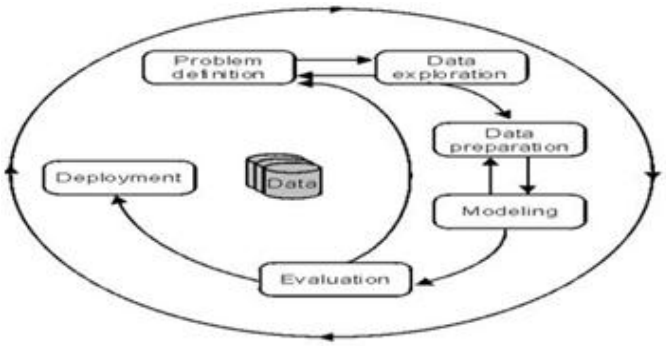

Fig.2. Data Mining Process Representation

(4) *Modelling:* Based on the data and the desired outcomes, a data mining algorithm or combination ofalgorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, *networks* and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.

(5) *Evaluation and Deployment:* Based on the results of the data mining *algorithms*, an analysis isconducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

IV. ALGORITHM USED FOR PREDICTION SYSTEM

K-Means Algorithm

According to MacQueen, the k- means algorithm is one among the algorithms of partitioning methods [10]. It is exceptionally straightforward and it can be effortlessly utilized for taking care of a large portion of the useful issues. It is the best squared error-based clustering algorithm.

Consider the data set with 'n' objects, i.e.

$$S \quad \{xi : 1 \le i \le n\}$$

1. Initialize k-partition based on some prior knowledge or randomlyi.e.

$$\{C1, C2, C3, ..., Ck\}$$

2. Evaluate the cluster prototype matrix M i.e the distance matrix of distances between k-clusters and data objects. Where, m is a column matrix 1× n .

$$M = \{m1, m2, m3, ..., mk\}$$

3. Each object in the data set are assigned to the nearest cluster - Cm

$$xj \in Cmif \; || xj - Cm || \le || xj - Ci ||$$

i.e.

$$\forall 1 \le j \le k, j \ne m$$

where, j=1,2,3,……..n

4. Compute the mean of each cluster and change the k-cluster centers by their means.
5. Once more calculate the cluster prototype matrix M.
6. Reiterate steps 3, 4 and 5 until there is no change for each cluster.

Global K-Means Algorithm

According to Likas [11] the global k-means clustering algorithm does not depend upon the initial parameter values and utilize the k-means algorithm as a local search procedure that constitutes a deterministic global optimization method. This technique proceed in an incremental way of attempting to optimally include one new cluster center at each stage instead of randomly selecting starting value for all clusters [12].

More particularly, to solve a clustering problem with k clusters the method proceeds as follows:
Step 1: We begin with one cluster (k=1) and cluster center corresponds to the centroid of the data set X.
Step 2: We perform N executions of the k-means to find two cluster (k=2) after the initial positions of the cluster centers: For k=1, the first cluster center is constantly placed at the

optimal position. The second center at execution n is placed at the position of the data point x_n (n-1,…..,N). The best solution is obtained after N executions of the k-means algorithm.

Step 3: For k-1 clustering problem, the final solution is denoted by (c_1,c_2,….,c_{k-1}). We perform N execution of the k-means algorithm with initial positions (c_1, c_2….. c_{k-1}, x_n) here n varies from 1 to N to find solution for k-clustering problem. The best solution obtained in this process is considered as the final solution.

K-Means++ Algorithm

K-means ++ was proposed in 2006 by RafailOstrovsky, Yuval Rabani, Leonard Schulman and Chaitanya Swamy, and independently in 2007 by David Arthur and Sergei Vassilvitskii [13]. It is an algorithm for choosing the initial values for the k-means clustering and an approximation algorithm for the NP –hard k-means problem. It provides away of avoiding poor clustering's observed in the standard k-means algorithm.

The k-means algorithm starts with an arbitrary set of cluster centers. We recommend a particular way of choosing these centers. At any time, let D(x) represents the shortest distance from a data point x to the closest center as we have previously selected. We define the k-means++ algorithm as,

1. Uniformly choosing an initial center c1 at random from X.
2. Choose the next center ci, selecting

$$ci = x' \in c \text{ with probability } \frac{D(x')2}{\sum_{x \in c} D(x)2}$$

3. Reiterate Step 1 until we have chosen a total of k centers.

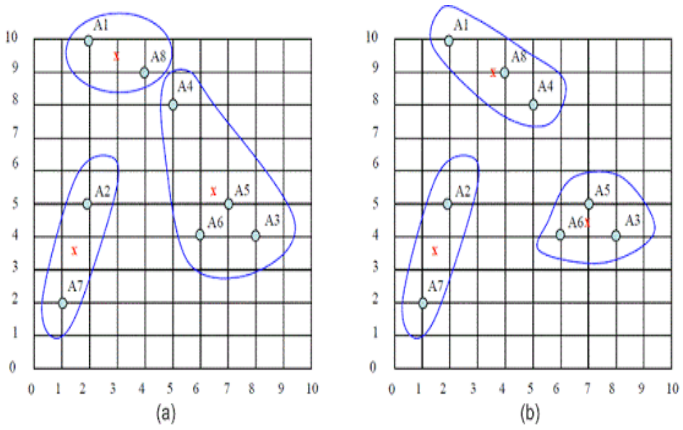

Fig.3.K-means ClusteringAnalysis

See5 Algorithm

The decision tree algorithm is very robust and learning efficiency with its learning time complexity of O(n$\log_2$n). The outcome of this algorithm is a decision tree that can be easily represented as a set of symbolic rules (IF….THEN….)[14]. This rule can be directly interpreted and compared with available biological knowledge and providing useful information for the biologist and clinicians.

According to Quinlan, the learning algorithm applies a divide-and-conquer strategy to construct the tree. The sets of instances are associated by a set of properties (attributes). A decision tree comprises of node and leaves where nodes represent a test on the values of an attribute and leaves represent the class of an instance that satisfies the tests [9]. The outcome is 'yes' or 'no' decision. Rules can be derived from the path from the root to a leaf and utilizing the nodes along the way as preconditions for the rule, to predict the class at the leaf. Pruning is necessary to remove unnecessary preconditions and duplications.

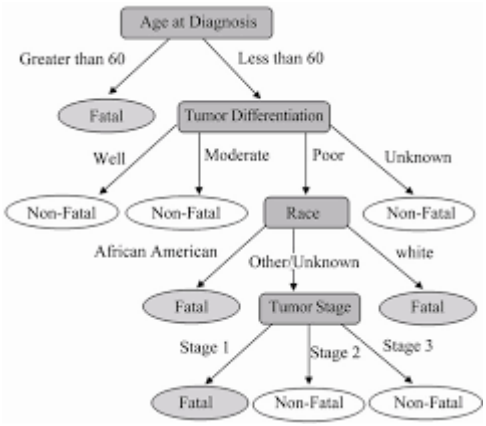

Fig. 4 Classification using decision tree

Gain Criterion

Suppose a possible test with n outcomes that partitions the set T of training cases into subsets T_1,T_2,…,T_n [15]. The only information available for guidance is the distribution of classes in T and its subsets, if this test is to be evaluated without exploring subsequent divisions of the T_i's. Some notation may be useful. For any set of cases S, let freq(C_j,S) represent for the number of cases in S that belong to class C_j. |S| denotes the numbers of cases in set S.

The original ID3 used criterion called Gain [8]. It is defined by imagining and selecting one case at random from a set S of cases and announcing that it belongs to some class C_j. This message has probability

$$\frac{freq(Cj,s)}{|S|}$$

And so the information it conveys is

$$-\log 2 \frac{freq(Cj,S)}{|S|}$$

On summing over the classes in proportion to their frequencies in S, providing

$$\inf o(S) = -\sum_{j=0}^{k} \frac{freq(Cj,S)}{|S|} * \log 2 \frac{freq(Cj,S)}{|S|}$$

On applying to the set of training cases, info (T) measures the average amount of information required to identify the class of a case in T. The expected information requirement can be establish as the weighted sum over the subsets, as

$$\inf ox(T) = -\sum_{1}^{n} \frac{|Ti|}{|T|} * \inf o(Ti)$$

The quantity i=

$$Gain(X) = \inf o(T) - \inf ox(T)$$

V. Comparison of K-Means based Algorithms & See5 Algorithms

In this paper[10] we used dataset to make a comparison study between k-means and See5 algorithms. The Colon dataset is a collection of gene expression measurements from 62 Colon biopsy samples from colon-cancer patients reported by Alon. Among them, 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were selected

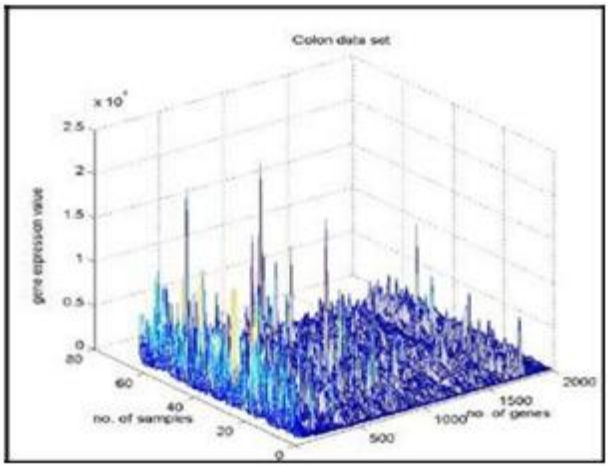

Fig. 5 Graphical representation of Colon dataset

The analysis of different variants of algorithm is done with the help of colon dataset. Variants of algorithm used in this study are k-means, global k-means, k-means++ and See5. In this case the see5 algorithms average accuracy is comparatively better than the k-means, global k-means and k-means++. The result of Average accuracy rate for over 2000-gene colon dataset are shown in below fig. 3.

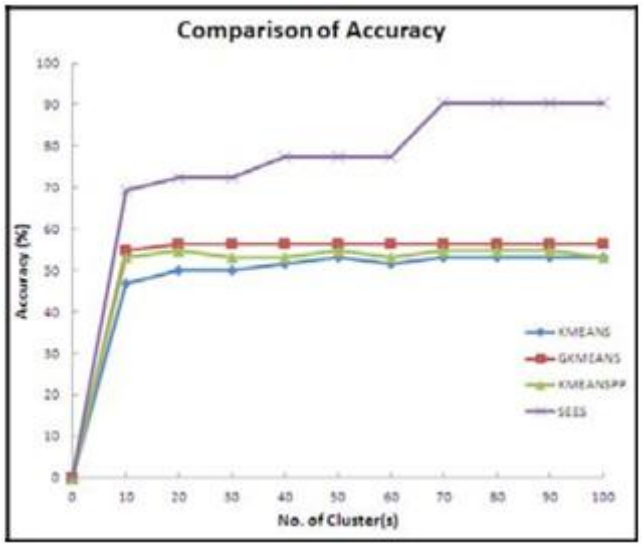

Fig.6 Comparison of Algorithm

VI. Conclusion

Restrictions in every algorithm are mulled over and see5 is demonstrating the improved results. In clustering exceptionally significant articles should have been assembled then our framework is said to function admirably. Taking the colon cancer disease dataset and tossed the algorithm it is obviously outstanding that see5 is having the higher precision on correlation with the quantity of bunches. Precision plays a key element for the social affair data from dataset. It is achievable through navigating different algorithms.

VII. References

[1] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining," 2006.

[2] Chang Pin Wei and Liou Ming Der, "Comparision of three Data Mining techniques with Ginetic Algorithm in analysis of Breast Cancer data".[Online].Available:http://www.ym.edu.tw/~dmliou/Paper/compar_threedata.pdf.

[3] Ritu Chauhan "Data clustering method for Discovering clusters in spatial cancer databases" International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010.

[4] Labeed K Abdulgafoor "Detection of Brain Tumor using Modified K-Means Algorithm and SVM" International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Applications NCRTCA 2013.

[5] A. Sahar "Predicting the Serverity of Breast Masses with Data Mining Methods" International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org.

[6] Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66.

[7] Guha, Rastogi.R, Shim.K, "CURE: An Efficient Clustering Algorithm for Large Databases", Proceedings of the ACM SIGMOD Conference, 1998, pp.73-84

[8] Osmar R. Zaïane, *Principles of Knowledge Discovery in Databases*. [Online]. Available:*webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf.*

[9] The Data Mining Process. [Online]. Available:http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=/com.ibm.im.easy.doc/c_dm_process.html.

[10] Parvesh Kumar, Siri Krishnan Vasan, "Analysis of X-means and global k-means using tumour classification", 2010; Volume V: pp. 832–835.

[11] Likas,A.,Vlassis.M, Verbeek.J,"The Global k-means Clustering Algorithm", Pattern Recognition, 36, pp. 451-461

[12] Dan Pelleg, Andrew Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters",ICML Proceedings of the Seventeenth International Conference on Machine Learning, 2000

[13] David Arthur,Sergei Vassilvitskii,"K-Means++: The Advantages of Careful Seeding",SODA Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007

[14] Quinlan J.R, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Inc., California

[15] Parvesh Kumar, Siri Krishnan Vasan, "Analysis of Cancer Datasets using Classification Algorithms", IJCSNS, 2010, Vol. 10, No.6, pp. 175–178.