Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.713 – 719

A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques

P. Saranya
M.Phil Computer Science, Bishop Heber College, Trichy, India
saranresearch16@gmail.com

B. Satheeskumar
Assistant Professor in Computer Science, Bishop Heber College, Trichy, India
satbhc@yahoo.com

Abstract— This paper presents a survey on medical image feature selection using data mining techniques. In medical field there are different kinds of problem in medical imaging like classification, segmentation, extraction and selection. Medical datasets are often categorized by huge amount of disease measurements and comparatively small amount of patient records. These measurements (feature selection) are not relevant, where this irrelevant and redundancy features are difficult to evaluate. On the other hand, the large number of features may cause the problem of memory storage in order to represent the data set. Different kinds of data mining techniques (or) algorithms can convenient with imprecision and uncertainty in data analysis and can effectively remove noisy and redundant information.

Keywords— data mining, feature selection, mining techniques.

1. INTRODUCTION

1.1 Data Mining

Data Mining is the process of discovering interesting knowledge from large amounts of data stored in database. It is an essential process where intelligent methods are applied in order to extract data patterns. Simply stated, data mining refers to “extracting” or “mining” knowledge from large number of data.[1] The major reason for using data mining has attracted a great deal of concentration in the information industry. Growth of the information technology has a huge amount of data and the imminent need for revolving such data into useful information and knowledge. These kinds of information gathered from business management, engineering design, science exploration and medical field.

1.2 Image Mining

Image Mining plays vital role among the researchers in the field of data mining. The image mining mainly focuses the extraction of pattern from large collection of images.

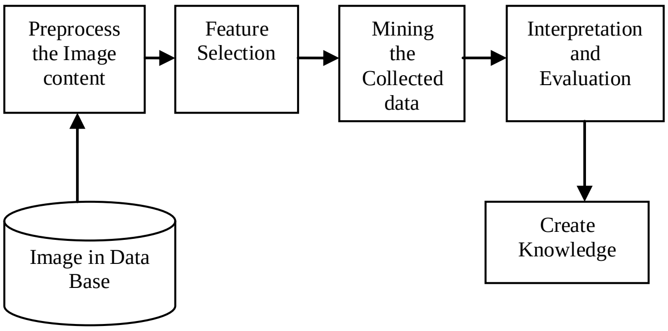
Preprocess the Image content
Feature Selection
Mining the Collected data
Interpretation and Evaluation
Image in Data Base
Create Knowledge

Fig 1. Image Mining Process[2]

The above figure1 shows that the steps involved in image processing. The images from the database are preprocessed to improve the image quality. Then image goes under various transformation, selection and extraction to measure important features. With these features, mining can use data mining techniques to discover the significant pattern. This pattern are evaluated and interpreted to find the knowledge.

Now a day, the rapid development of digital medical device, medical information data bases have included not only the structural information of patients, it also have non - structural medical image information. Even medical images give a small picture of organ in the body. It may take more time to extract the useful information. Because the medical researchers consider effectiveness of the treatment depends upon the information in the medical data base. So that medical information plays vital role in medical field. The researchers have to concentrate more to find out the exact feature selection of the medical images. In this paper section II, the main concept of feature selection, section III, detail about Rough set theory, section IV, describes the literature survey, section V, describes the experimental results using rough set, Finally section VI concludes the paper.

2. Feature Selection

The main problem in medical diagnosis of data is feature selection. Features are selected depends on the property of the data. It deals with the redundant and irrelevant data.

A. Components in FS

There are three components in the feature selection method :a) *search algorithm* – which looks through the space of the feature subsets; b) an *evaluation function* – used to evaluate examined subsets of feature; c) classifier – which constructed based on a final subset features.[3]

B. Advantages

- FS is useful to different computational task, for instance in process of machine learning,
- Improve the data manipulation speed
- To reduce the dimensionality problem
- Improve the classification rate by reduce irrelevant and noisy data
- These are all the some advantages of the feature section.

C. Disadvantages:

- Feature selection method search through the entire subset of features and find the best subset features among the 2^N-1 where N is the total number of feature.

3. Basic Data Mining Techniques

3.1. Rough Set Theory

3.1.1 Description

Rough set was introduced at 1991 by polish computation scientist Zdzislaw I. Palwak is a formal approximation of a conventional set in terms of a pair of sets which give the lower and upper approximation of original data set.

3.1.2 Basic Concept of Rough Set

In, Rough set theory data are collected in an information table. Such that table rows are represents the objects and columns are represents the feature if the objects. Let take O, F denote the set of objects and object features. Assume that $B \subseteq F$; $x \in O$. Further, let $x\sim \boldsymbol{B}$ denote

$$x\sim/\boldsymbol{B} = \{ y \in O \mid \forall \emptyset\, B, \emptyset(x) = \emptyset(y)$$

i.e., $x\sim \boldsymbol{B}$ y (description of x matches the description of y).
Rough sets theory defines three regions based on the equivalent classes induced by the feature values: lower approximation $\underline{B}X$, upper approximation $\overline{B}X$ and boundary $BND_B(X)$. A lower approximation of a set X contains all equivalence classes $x\sim/\boldsymbol{B}$ that are proper subsets of X, and upper approximation $\overline{B}X$ contains all equivalence classes $x\sim/\boldsymbol{B}$ that have objects in common with X, while the boundary $BND_B(X)$ is the set $\underline{B}X$ / $\overline{B}X$, i.e., the set of all objects in that $\overline{B}X$ are not contained in $\underline{B}X$. Any set X with a non-empty boundary is roughly known relative, i.e., X is an example of a rough set.[4]

- ➢ Advatntages[5]
  - It is based on the original data only
  - Does not need any external information
  - No assumption about the data are made
  - It is suitable for both quantitative and qualitative data
  - The result of rough set model are easy to interpret

3.2. Decision Tree

A Decision tree has three types of node, internal node denotes the test on an attribute, leaf node denotes the classes or class distributions, root node is the top most node in a tree.C4.5 and ID3 these are the two main algorithms used to construct the decision tree.

- ➢ Advantages
  - Decision tree uses “tree pruning” approach to identify and remove noisy data from the branch and to improve the classification accuracy.
  - The attribute with the highest normalized data is chosen to make a decision.
  - Algorithm used continuous and discrete values.

- ➢ Disadvantages
  - Efficiency and scalability are low when applied to mining of very large data bases.
  - Decision tree construction inefficient when swapping sample data from main memory to cache memory.
  - C4.5 algorithm can contain empty attributes, over fitting.

3.3. Naïve Bayesian Classification

Bayesian classification is derived from the Bayes theorem. It is also known as “Simple Bayesian Classifier”. In this classifier each data sample is represented by an n- dimensional vector and measurements samples are formed by n- attributes. Suppose there are m classes, C_1, C_2 …..C_3 having a unknown data sample, X, Naïve Bayesian classifier will predict that X belongs to class having the highest probability conditioned on X.

$P(C_i| X) > P(C_j|X)$ for $1\leq j\leq m$, $j\neq i$.

- Advantages
  - Comparing other classifier naïve classifier gives high accuracy and speed when applied large data base.
  - Minimum error rate
  - Suitable for large amount of data bases.

- Disadvantages
  - It considering each attribute in each class separately.
  - Zero conditional probability problem.

3.4. Backpropagation Neural Network

Backpropagation is based on the neural network algorithm. Here we are having the three kinds of layer, input layer- the inputs are fed into this layer, hidden layer – weighted outputs can be input to another hidden layer, number of hidden layer's arbitrary, but only one is used, output layer- hidden layers are give the input to the output layer, which emits the network's prediction. This network model helps to classifying new data.

- Advantages
  - High accuracy neural networks are able to approximate complex non-liner mappings.
  - Very flexible with the noisy and inconsistence data
  - Easily updated with the present data

- Disadvantages
  - Major disadvantage is knowledge representation
  - Knowledge interpret is difficult for humans.
  - Removing weighted links that decreases the accuracy of the given network.
  - General method is not available.
  - Difficult to select training dataset.

3.5. Support Vector Machine

Support Vector Machine is one type of learning system algorithm, which is used to perform classification more accurately. SVM used for two class classifier. The essence of SVM is hyper plane also known as "decision boundary or decision surface". This hyper plane separates the positive and negative of training data sample.

- Advantages
  - SVM is easy to extended, useful to pattern reorganization, formulated quadratic optimization problem.

- Disadvantages
  - It is suitable for real valued space.
  - For binary classification, it allow only two classes and for multiple class classification, it apply several strategies
  - Hyper plane is very hard to use by the user

3.6 K- Nearest Neighbor Learning

KNN is another one important learning method to classify the simple data set. Comparing with other learning classifier approaches k-nn method is quite effective. Here the learning occurs when test examples need classified.

3.6.1 Algorithm[k,d,D]

- compute the distance between d and every example in D;
- choose the k examples in D that are nearest to d, denote the set by P(⊆ D);
- assign d the class that is the most frequent class in P[6]

3.7 Classification and Regression

Classification and Regression Tree represents the data as tree structure. It is also known as CART. It displays the data tree relationship. CART uses another analytical technique very effectively.

- ➢ Advantages
- CART does not need any specification, is non parametric
- Easy to identify the most significant variable and eliminate the noisy data
- Does not select variable as earlier stage of classification

Disadvantages

i. This tree is not in the stable state always, splitting variable may decrease or increase the complexity of the decision tree
ii. CART split only one variable.

4. LITERATURE SURVEY

The overview of feature selection and different data mining classification techniques is described in the previous section. In this section various researchers are used different data mining techniques to identify most relevant features from the medical image data. In medical point of view different techniques is used to identify the most important features/attributes influencing the treatment of patients.

A.E. Hassanien has proposed in his paper rough set theory approach for attribute reduction and generate rule to identify breast cancer. [7]

Tsumoto, in his paper proposed a rough set algorithm to generate diagnostic rule based on the hierarchical structure of different medical diagnostics. The induced rules can correctly represent perfect decision process.[8]

Komorowski and Ohrn, they suggested using rough set approach for identifying a patient group in need scintigraphic scan for subsequent modeling. [9]

Bazen, in his paper compares rough set-based methods with other methods. His diagnosis medical data, i.e., lymphography, breast cancer and primary tumors and find out the errors for rough set are fully comparable as well as often lower than other techniques. [10]

J. Zhou, K.L. Chan, V.F.H. Chong and S.M. Krishnan in their paper they proposed support vector machine classification for brain tumor extraction. This technique gives great potential and usefulness in MRI tumor segmentation. [11]

Gopala Krishna and Murthy Nookala[et al] made a comparative study about the 14 different classification algorithms by using the 3 different types of cancer data sets. Most of the algorithms give better result when size of the attributes is increased. However accuracy level is depends on the kind of the datasets to be used. Finally they realize that algorithms are not gives the better accuracy level, user try to choose the best data set. [12]

Priyanga and Prakasam in their paper, they proposed a cancer prediction system based on data mining technology. Here they collect the user's genetic and non-genetic factor, which is helps to predict the breast cancer at early stage. Main drawback of this system is cost effective to the user. Weka system is used to analyze the medical information. Once the attributes are finalized, then the range of the risk can be determined by the prediction system. Here we are having four levels low level, intermediate level, high level and very high level. The above system was successfully applied with the datasets of breast cancer, it gives better accuracy level comparing to the existing system. This system gives earlier stage warning to the users, cost and time benefits to the user.[13]

ShwetaKharya have proposed some effective data mining techniques to classify the breast cancer. By using different kind of data mining techniques, soft computing approaches and decision tree she was found to be best predictor with 93.62% accuracy on benchmark and SEER data set[14]. The predictor can be used to design for the web page application. [15]

Cheng-Mei Chen and Chien Yen Hsu, they proposed survival prediction model for liver cancer using data mining techniques. They collect dataset from the medical data center in North Taiwan between the years 2004 and 2008. They extract nine variables to liver cancer survival. Artificial Neural Network(ANN) and Class and Regression Tree(CART) were used prediction model. The model was tested under three conditions: One Variable(Clinical Stage), Six significant variable and all nine variable (Both significant and Non significant). The results shows that ANN model with nine inputs gives 0.915% accuracy, 0.87% sensitivity and 0.88% specificity respectively. Finally they conclude that ANN model gives more accuracy than the CART model. [16]

Senthil [et al] they analyzed liver cancer DNA sequence data using data mining techniques. It focused only based on the biological modules rather than the individual genes. They tested by using different states. Finally all stages results are give the same approximate percentage. [17]

Krishnaiah in his paper, he used One Dependency Augmented Naïve Bayse Classifier(ODANB) and Naïve Creedal Classifier2(NCC2) for data preprocessing and effective decision making. He used generic ling cancer disease symptom like age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can predict the disease at early stage.[18]

Ada and Rajneet used data mining technique neural network and SVM to perform medical image mining, segmentation, data processing, feature extraction and classification. In their paper, they suggested that classify digital X-ray chest films into two categories: Normal and Abnormal. The normal ones are characterizing the healthy patient dataset and abnormal one includes the lung cancer patient dataset. The different datasets are trained by the SVMs. Finally the result was compared and reported.[19]

Dursun[et al] used popular data mining algorithms such as artificial neural network and decision tree and along with logistic regression to develop the prediction model for the breast cancer. It used the large data base. Comparing the prediction models, that gives 93.6% accuracy with decision tree, 91.2% accuracy with ANN and worst accuracy 89.2% with logistic regression.[20]

Vasantha[et al] have proposed a image classifier to classify the mammogram images. It classified into normal image, benign image and malignant image. There are 26 features are extracted from the mammogram image. They used hybrid approach of feature selection which reduces 75% of features. After reducing the features decision tree algorithm is applied to the mammography classification. This method proves that easier and less computing time than existing method.[21]

Rajendran and Madheswaran suggested an improved image mining technique for brain tumor classification. Here they used association rule with MARI algorithm. The experimental result shows the 96% and 93% sensitivity and accuracy respectively. [22]

5. CONCLUSION

Thus the survey helps to identify the data mining techniques to predict the cancer disease at earlier stage. Different researchers have proposed different techniques to predict the cancer disorder and different kinds of accuracy level as per used techniques. These techniques help to minimize the irrelevant data of patient's data from the data bases in medical center. Algorithms such as decision tree, ANN, Support Vector Machine, Naïve bays, Backpropagation and Regression are consider for the study. These algorithm gave the various result based on speed, accuracy, performance and cost. Also these effective classification data helps to find the treatment to the patient. In future a better method to predict the cancer disease can be found out with improvements in existing methods.

References

[1] **Jiewai Han and Micheline Kamber** ,"Data Mining Concepts and Techniques", second edition.

[2] **A. Kannan, Dr. V. Mohan, Dr. N. Anbazhagan,** "Image Clustering and Retrieval using Image Mining Techniques", IEEE International Conference on Computational Intelligence and Computing Research, 2010.

[3] **Maciej Komosinski and Krzysztof Krawiec,** "Evolutionary Weighting of image features for diagnosing of CNS tumors", Artificial Intelligence In Medicine · June 2000.

[4] **Aboul Ella Hassanien, Ajith Abraham, James F. Peters, Gerald Schaefer, Christopher** Henry, "Rough Sets and Near Sets in Medical Imaging: A Review", IEEE Trans. On Information Technology In Biomedicine, Vol. X, No. X, Nov. 2008.

[5] **Yudel Gómez, Rafael Bello, Amilkar Puris, María M. García, Ann Nowe,** "Two Step Swarm Intelligence to Solve the Feature Selection Problem" , Journal of Universal Computer Science, vol. 14, no. 15 (2008), 2582-2596.

[6] **Bing Liu**, " Web Data Mining Exploring Hyperlinks, Content and Usage Data".

[7] **A.E. Hassanian,** " Rough set approach for attribute reduction and Rule generation: a case of patients with suspected breast cancer, J.Am. Soc.Inform.Sci Technol.2004.

[8] **S. Tsumoto**, "Mining Diagnostic rules from clinical databases using rough sets and medical diagnostic model", Inform.Sci. 162(2004)

[9] **J.Komorowski, A.Ohrn,**" Modelling Prognostic Power of Cardiac tests using Roughsets", Artif,Intell.Med.15(1999).

[10] **J. Bazan,**" A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables", in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, Physica-Verlag, Heidelberg, 1998.

[11] **J. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak, J. Wr´oblewski**, "Rough set algorithms in classification problems, in: L. Polkowski, T.Y. Lin, S. Tsumoto (Eds.)",Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems. Studies in Fuzziness and Soft Computing, vol. 56, Physica-Verlag, Heidelberg, Germany, 2000.

[12] **Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, NagarajuOrsu, Suresh B. Mudunuri**, " Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification" , (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013.

[13] **A.Priyanga and Dr,S.Prakasam** " The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness" International Journal of Computer Science and Engineering Communications- IJCSEC. Vol.1 Issue.1, December 2013.

[14] http:// seer.cancer.gov/data/

[15] **ShwetaKharya**, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[16] **Cheng-Mei Chen, Chien-Yeh Hsu, Cheng-Mei Chen andChien-Yeh Hsu**," Prediction of Survival in Patients with Liver Cancer using Artificial Neural Networks and Classification and Regression Trees" Seventh International Conference on Natural Computation 2011.

[17] **K.Senthamaraikannan,N.SenthilvelMurugan, V.Vallinayagam and T. Viveka**, "Analysis of Liver Cancer DNA Sequence Data using Data Mining"International Journal of Computer Applications (0975 – 8887) Volume 61– No.3, January 2013.

[18] **V.Krishnaiah, Dr.G.Narsimha and N.Subhash Chandra**, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013.

[19] **Ada and RajneetKaur**, "A Study of Detection of Lung Cancer Using Data Mining Classification Techniques",International Journal of Advanced Research in Computer Science and Software Engineering 3(3), March – 2013

[20] **DursunDelen, Glenn Walker, Amit and Kadam**, "Predicting breast cancer survivability: a comparison of three data mining methods" Department of Management Science and Information Systems, Oklahoma State University, 700 North Greenwood Venue, Tulsa, OK 74106, USAReceived 13 January 2004.

[21] **M. Vasantha, Dr. V. Subbiah bharathi, R. Dhamodharan,**" Medical Image Feature, Extraction, Selection and Classification, International Journal of Engineering Science and Technology, Vol.2(6),2010,2071-2076

[22] **P.Rajendran , M. Madheswaren**, " An Improved Image Mining Technique For Brain Tumour Classification Using Efficient Classifier, International Journal of Computer Science and information Security, Vol.6, No.3,2009.