ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

A Study of Classification Techniques of Data Mining Techniques in Health Related Research

K Prasanna Jyothi[1], Dr R SivaRanjani[2], Dr Tusar Kanti Mishra[3], S Ranjan Mishra[4]

M. Tech Student, Dept. of C.S.E, ANITS, Visakhapatnam, India[1]

Professor, Dept. of C.S.E, ANITS, Visakhapatnam, India[2]

Associate Professor, Dept. of C.S.E, ANITS, Visakhapatnam, India[3]

Assistant Professor, Dept. of C.S.E, ANITS, Visakhapatnam, India[4]

ABSTRACT: Several health related studies used data mining and machine learning techniques for analysis and prediction health risk from patients health records. There are different classification techniques to predict the risk in health related studies. Classification techniques in data mining can process large amounts of data. Different classification algorithms are there to classify the data in data mining. Several classification algorithms including c 4.5, ID3, k-nearest neighbor, SVM, Naive Bayes, and ANN. In this paper the study of the various classification techniques is presented.

KEYWORDS: c 4.5, ID3, K-Nearest neighbor, SVM, Naive Bayes, and ANN

I. INTRODUCTION

Recent advances in health related studies are concentrating on risk prediction of diseases. In the study of risk prediction from patients health records different classification techniques are used. To identify the risk from the patients health records different classification techniques of data mining are used. Large amounts of data is processed using different classification methods. Classification techniques can be used to predict class labels. It classifies data based on training set and class labels. Classification procedure is used to repeatedly make decisions in new situations [1,4,5]. In data mining different classification methods are there to classify data and predict variables. In different situations of classification different classification method is used. There is nothing like a particular classification method is accurate to classify the data in all situations. The accuracy of classification method is depends on the data we want to classify. In health related studies the main research is going on the risk prediction. Predicting the risk in study of any disease is the main goal of health related studies. In health related studies researchers are using data mining techniques especially classification methods to classify data. From the classified data researchers predicting the risk from the patients health records. By doing this the researches want to show the conditions those causes that particular disease. For this research the classification methods are very helpful and giving the best results. In this type of researches, researchers generates classification rules from the classification algorithms. From the classification rules a decision tree can be generated. From the resultant tree attributes are assigned to classes. We can predict the results from the classes.

II. RELATED WORK

In [1] the importance of Data mining techniques and Artificial Intelligence is clearly explained. Only Artificial intelligence or data mining techniques can solve problems up to some extent. If we combine the both, we can break complex problems in computer field. Expert systems will do decision making like a human expert. They can solve

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

complex problems by its knowledge. Expert systems using data mining techniques to solve specialized problems[2]. By symptoms, prediction of disease is not accurate, it can be correctly predicted or wrongly predicted. Symptoms and result both are connected to characteristics of uncertainty[8,9]. In prevention of disease by depending on patient's records is enhanced by combining artificial intelligence and data mining techniques[3].

III. CLASSIFICATION

Classification is very helpful method in predicting the risk in health related studies. Classification categorizes the items into target classes. The aim of the classification is to predict the target class accurately from the data[8]. Classifier can learn from the examples. Modern classification techniques gives more intelligent prediction results[5,7,8].

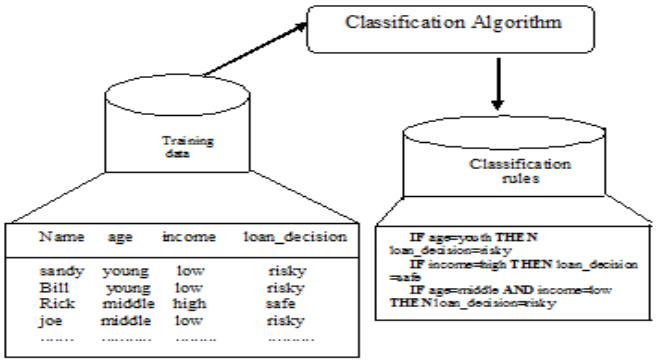

Figure1. Classification Technique in Data Mining

Classification is a technique in which the training data is processed by classification algorithm and it generates classifications rules. From the obtained classification rules a tree can be generated[17]. Figure 1 is clearly representing that in classification of data, training data is processed by classification algorithm. The classification algorithm generates the rules from the training data. The generated rules are known as classification rules.
In Health related analysis the research workers initially collect health records from different sources. From the collected health records they will identify the useful data and takes large portion of useful data for training purpose, that is called as training data. On the training data they apply classification algorithm to get classification rules. These rules will help in the disease prediction and from the rules researches develops the conditions of occurrence of a particular diseases.

IV. TAXONOMY OF CLASSIFICATION

This section is about the main things of classification, what are all the things involved in classification of data.

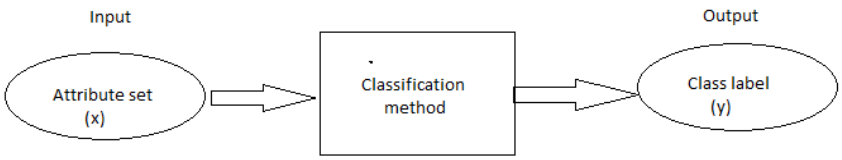

Figure2. Classification process in Data Mining

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Classification mainly includes three things, 1. Attribute set, 2. Classification method, 3. Class label.

1. Attribute set: Attribute set consists of relevant data stored in a particular format as attributes. The attribute set consists of data that is collected from health records of different sources. The important attributes from health records from different sources are arranged into attribute set. The data in the attribute set is ready for classification process[14].Before going for classification we need to prepare data. Following are the things involved in preparing data for classification.
    - Data Cleaning: Data Cleaning involves removing noise from data and missing values treatment. By applying smoothing techniques we can remove noise from data And missing values can be replaced by most relevant attribute.
    - Relevance analysis: Attribute set may contain irrelevant attributes. Correlation analysis will give the relevant attributes by comparing two attributes.
    - Data transformation and reduction: Data transformation can be done by two methods.
        (i) Normalization: Scaling all values for attribute in order to make them fit in specified range. Neural network methods are used to do normalization.
        (ii) Generalization: Data is transformed to higher concept. Concept hierarchy is used for generalization.
2. Classification method: Classification methods are capable to process huge amount of data in data mining. Some classifications methods are there to classify data in health records, Those are ID3 algorithm, C4.5 algorithm, Naive Bayes classification, K-nearest neighbor and SVM.
3. Class label: Class label is known as target class. On the attribute set classification method is applied to get the target class. The resultant class is known as class label[15,17,18].

V. ANALYSIS OF CLASSIFICATION ALGORITHMS

A brief description of the classification algorithms is provided in this Analysis of classification algorithms. Many algorithms are there in classification. In this section the main algorithms implementing in health related research are discussed. Those are ID3 algorithm, C 4.5 algorithm, K-nearest neighbor algorithm, Naive Bayes and SVM [7].

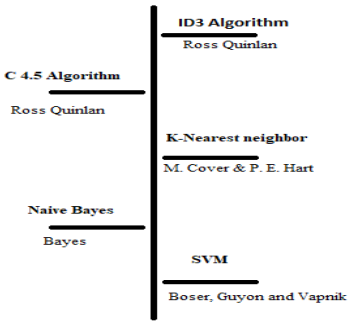

Figure3. Classification algorithms used in health related research

Figure3 illustrates the algorithms used in classification of data in health related researches and the inventors who invented those algorithms.

A. **ID3 Algorithm**: ID3 stands for Iterative Dichotomiser 3. ID3 is Invented by Ross Quinlan. ID3 generates decision tree from data. It takes initial data set S as root node. Iterations are takes place on the data set S. The unused attributes iterates on S and computes Entropy(H(s)) or Information Gain(IG(s)). Now selects the value that is having highest information gain or lowest entropy. Then the dataset S is split into sub sets by the selected attribute [12,18].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Algorithm:

- On the initial data set S, Calculate entropy for every attribute of S.
- Based on the attribute for which entropy is minimum split S into subsets
- Construct decision tree node, which is containing the attribute.
- Iterates through subsets using leftover attributes.

Metrics:

- Entropy: Entropy is denoted by H(S). It is the measure of uncertainty in the data set S.

$$EntropyH(S) = -\sum_{i=1}^{n} p_i \, log_2 p_i = \sum_{i=1}^{n} p_i log_2 (\frac{1}{p_i})$$

- Information gain:

Information gain is denoted with IG(A).

$$IG(A, S) = H(S) - \sum_{t \in T} p(t) H(t)$$

Where T= Subset of S
H(t)=Entropy of t
p(t)=Number of elements in t/Number of elements in S
H(S)=Entropy of S

Figure4 illustrates the generation of decision tree using ID3 algorithm. The attributes with highest gain will be root nodes and child nodes will be generated from those root nodes. Leaf nodes will be called as class labels.

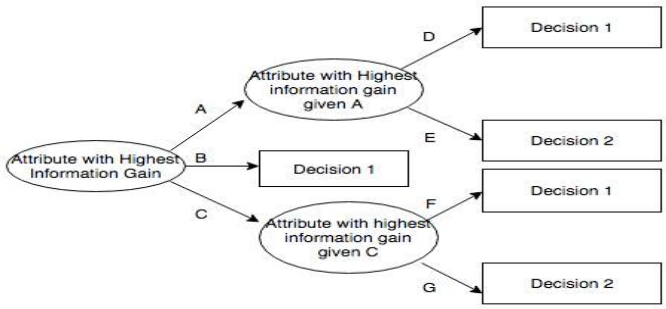

Figure4. Decision tree generation in ID3 algorithm

B. **C 4.5 algorithm:**C 4.5 algorithm builds decision tree from the training data set S. S contains classified sample, which are classified before. It uses ID3 concept in construction of decision tree by calculating entropy and information gain [11,18].The training set S= s1,s2,s3,..... Si consists of p-dimensional vector x1i,x2i,....xpi. These are values of features of sample, Class in which si fails.

Base cases:

- When all the samples belongs to same class, It simply creates leaf node to the decision tree.
- If no attribute provides any gain then, It creates decision tree node with expected class node.
- If previously-unseen instance of class appeared, Then also it creates a node with expected value of class.

Algorithm:

1. Check for C 4.5 bases cases.
2. Consider attribute a, and find out normalized information gain of a.
3. Assume the highest normalized attribute as a_best.
4. Create decision tree by making a_best as root node.
5. Iterate the process through Childs of a_best.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

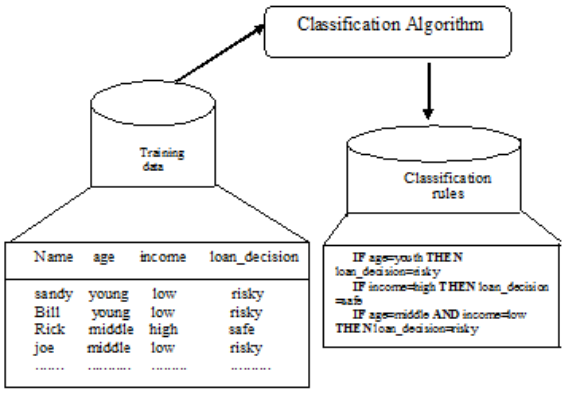

Figure5. Example of C4.5 algorithm

C. **K-Nearest neighbor algorithm:** K-nearest neighbor algorithm was proposed by M. Cover and P. E. Hart. In this algorithm class is known. On the concept of nearest neighbor it identifies the category of attribute to which class it belongs to. It considers more than one nearest neighbor to identify the class to which the data point it belongs to. At run time the data points will be in memory[2]. The training points will be assigned with some weights based on their distance from data points. To improve memory limitations the NN training set can be used to structure using different techniques. To defeat the memory limitations data set size is trim down. In this algorithm the object's group to which it belong to is unknown. To determine the group nearest neighbor technique is used [2,9,13].

Algorithm:

K: Count of nearest neighbors

For each object A in the test set do

Calculate the distance D(A,B) between A and every object B in the training set

neighbor ⇐ K-neighbors in the training set closest to A

A.class ⇐ SelectClass(neighbor)

End For

In figure 6 the example of k-nearest neighbor is explained. To reach point C from A we should choose shortest path. A's neighbors are B and E. In AB and AE, AB is shortest path so we should choose AB. B's neighbors are C and E. In BE and BC, BE is shortest path so we should choose BE. E's neighbors are C and D. In EC and ED, EC is shortest path. So nearest neighbor path is A-B-E-C.

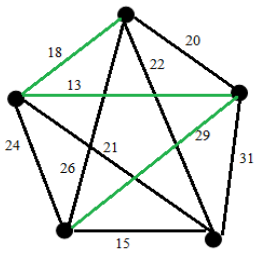

Figure6. Example of K-nearest neighbor algorithm

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

D. **Naive Bayes classification algorithm:**Naive Bayes classifier is a probabilistic classifier, Which considers probability values to classify attributes. It is a independent feature model. It calculates probabilities to hypothesis. And also it is robust to noise in data. This classifier assumes that presence /absence of a particular feature is lineal to any feature's presence/absence. It is a supervised learning classifier [6,19]. It is easy to construct and easy to interpret. So users who are don't have prior knowledge in classification can also construct Naive Bayes classifier [9,11,13,21].
In case of objects to be classified P(X/Y) is the probability distribution Where X denotes classes, Y denotes descriptions. Consider a description d of particular object , We appropriate the class ***argmaxc P*(X=x|Y=y).** A Bayesian distribution splits the distribution as follows **P(X)** and **P(X/Y):P(Y=y|X=x)P(X=x).*argmaxc P*(X= x|Y= y),**

P(Y=y/x=c)P(X=x)→(1)

Equation (1) represents the likelihood of given description of given class. In this attribute vector is represented by a1,....,an for attributes A1,...,An. **P(Y=y|X=x)** requires the estimation probability **P(A1=a1,....,An=an|X=x)**. It is joint probability.Here in this Naive Bayes classification we should assume that all attributes are independent[12]. Given the class:

***P* (*A*1 = *a*1, . . . , *An* = *an*|*C* = *c*)*Π*^{n}*i*=1*P*(*Ai*=*ai*|*C*=*c*) →(2)**

This assumption is called Naive Bayes assumption.Figure 7 illustrates the Naive Bayes classification algorithm.

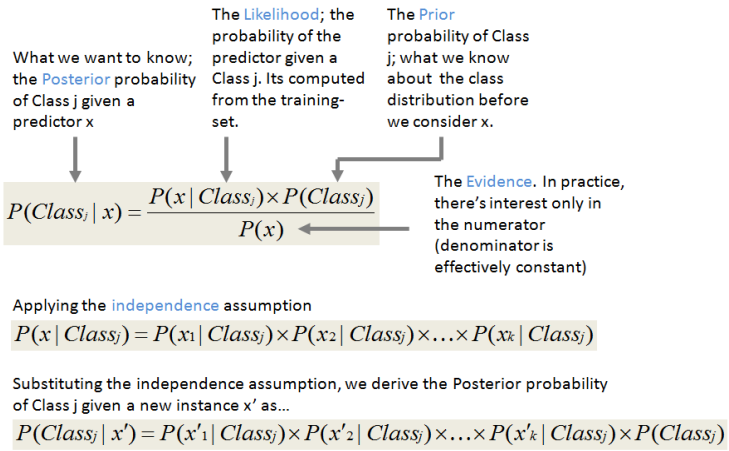

Figure 7: Example of K-nearest neighbor algorithm

E. SVM(Support Vector Machine) classification: SVM was invented by Boser, Guyon and Vapnik. The Support vector machine deals with pattern classification [3]. There are two types of patterns linear and non-linear. Linear patterns can be easily distinguishable and non-linear patterns are not easily distinguishable. The principal concept behind SVM is to develop optimal hyper plane. That hyper plane should be used for classification of linearly separable problems [3,16,20]. The optimal hyper plane means that the hyper plane selected classifying patterns, which is having the maximum size [13,21]. It will be helpful to classify patterns correctly. If the margin size is large then there will be more correctly classified patterns.
Hyper plane Equation:Hyper plane, **aX + bY = C**
The kernal function used to map given function is Φ(x).
x→Φ(x).
The kernel functions used are SIGMOID, POLY, LINEAR and RBF.
Poly kernel function equation is:
K(x, y) =<x, y>^p
The SVM gives the identically distributed, independent training samples

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

{(xi ,yi)}N i=1, where x ϵ Rd and yi ϵ {−1,1}.
Here the main goal is to find hyper plane is: wT.x + b = 0.

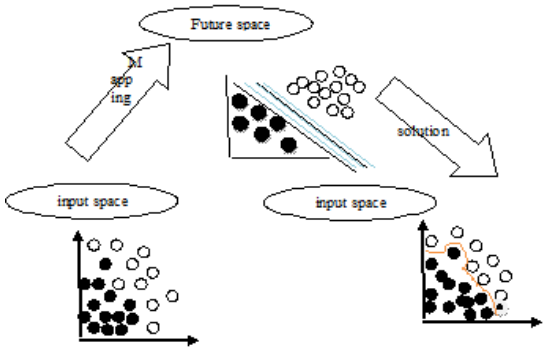

Figure8. Flow of SVM algorithm

VI. CONCLUSION

In this paper, some classification techniques related to health research have been studied. In health related researches data collected will be in the form of health records. On the collected data of health records of patients if we apply the appropriate classification algorithm, it will reflects the correct class labels. If we take class label such that it should be the predicting variable then the result will be the risk prediction of any disease. Here we discussed some of those classification algorithms which will be useful in health related researches.

REFERENCES

1. Delveen Luqman Abd Al.Nabi, Shereen Shukri Ahmed, “Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)” (ISSN 2222-2863),Vol.4, No.8, 2013
2. Nitin Bhatia, Vandana,” Survey of Nearest Neighbor Techniques” (IJCSIS) Vol. 8, No. 2, 2010, ISSN 1947-5500.
3. Ashis Pradhan., “Support Vector Machines a survey, ISSN 2250-2459, Volume 2, Issue 8, August 2012
4. SMs. Aparna Raj, Mrs. Bincy, Mrs. T.Mathu “Survey on Common Data Mining Classification Techniques”, International Journal of Wisdom Based Computing, Vol. 2(1), April 2012
5. Raj Kumar, Dr. Rajesh Verma,” Classification Algorithms for Data Mining P: A Survey” IJIET Vol. 1 Issue August 2012, ISSN: 2319 – 1058.
6. Vidhya.K. G.Aghila. “A Survey of Naïve Bayes Machine Learning approach in Text Document Classification”,(IJIST) Vol. 7, No. 2, 2010.
7. B. Kotsiantis · I. D. Zaharakis · P. E. Pintelas, “Machine learning: a review of classification and combining techniques”, Springer Science10 November 2007
8. S.Archana, Dr. K.Elangovan, “Survey of classifications techniques in Data Mining”, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014
9. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu,Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, " Top 10 algorithms in data mining", Knowledge and Information Systems, January 2008, Volume 14, Issue 1, pp 1–37
10. Brijesh Kumar Bhardwaj, Saurabh Pal, "Data Mining: A prediction for performance improvement using classification"; Vol. 9, No. 4, April 2011, pp 136-140.
11. ParneetKaur, ManpreetSingh, Gurpreet SinghJosan, " Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector", Procedia Computer Science, Volume 57, 2015.
12. ID3 Algorithm, http://en.wikipedia.org
13. Classification in Data Mining, https://www.tutorialspoint.com
14. Ms. Aparna Raj, Mrs. Bincy, Mrs. T.Mathu “Survey on Common Data Mining Classification Techniques”, International Journal of Wisdom Based Computing, Vol. 2(1), April 2012.
15. Vivek Agarwal, Saket Thakare, Akshay Jaiswal, "Survey on Classification Techniques for Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 132 – No.4, December2015.
16. M.A. Hearst, ”Support vector machines,” IEEE Intelligent Systems, pp.18-28, 1998.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

17. Han, J., & Kamber, M. (2006). "Data Mining: Concepts and Techniques" (2nd ed.). Morgan Kaufmann Publishers.
18. Barros, R. C., Basgalupp, M. P., Carvalho, A. C., &Freitas, A. A. (2010, Jan). "A Survey of Evolutionary Algorithms for DecisionTree Induction". IEEE Transactions on Systems,Mans and Cybernetics, Vol. 10,No. 10,pp. 1-22.
19. K.P.Murphy. (2006). Naive Bayes classifiers. [Online]. Available: http://www.cs.ubc.calmurphyk/TeachinglCS340-Fall06/readingiNB. pdf.
20. K.P.Bennet and C.Campbell. "Support Vector Machines: Hype or Hallelujah?" in Proc. SIGKDD Explorations, 2000, vol. 2, no. 2, pp 1-13.
21. Supreet Kaur, Amanjot Kaur Grewal, " A Review Paper On Data Mining Classification Techniques For Detection Of Lung Cancer", International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 11 | Nov -2016

BIOGRAPHY

K Prasanna Jyothiis a M.Tech student in the Computer Science Department, College of Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. Her research interests are Data Mining, Artificial Intelligence, Neural Networks etc.

Dr R SivaRanjaniis Professor & HOD ofComputer Science Department,Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. She received her Ph.D from Andhra University. She is life time member of ISTE and CSI. Her research interests are Cryptography & security, Cyber forensics and image processing etc.

Dr Tusar Kanti Mishrais Associate Professor inComputer Science Department,Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. He received his Ph.D from NIT Rourkela. He is having membership in ACEEE. His research interests are Pattern Recognition, Image Processing, Computational Intelligence, Machine Learning, Optical Character recognition etc.

S Ranjan Mishrais Assistant Professor inComputer Science Department,Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. He is pursuing Ph.D from NIT Durgapur. He is having membership in CSI. His research interests are Pattern image processing, Computer vision, Machine intelligence, WNS etc.