ISSN:0975-9646

Data Mining Techniques Using WEKA classification for Sickle Cell Disease

Ashokkumar Vijaysinh Solanki

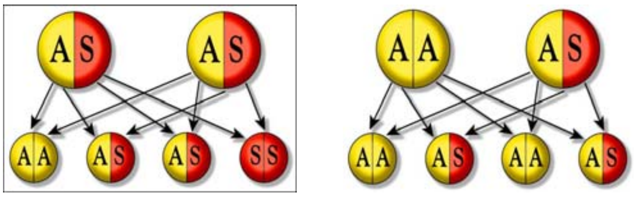
Research Scholar, JJT University,
Rajasthan, India

Abstract- **Data mining is an inter-disciplinary promising field that focuses on access of information useful for high-level decisions and also includes Machine Learning. Data miners evaluate and filter the data as a result convert this data to information and information to knowledge by performing some techniques. This paper is towards the areas containing Data Mining techniques for analysis about the disease highly affected to tribal zone of Gujarat, which is known as Sickle Cell Disease (SCD). Through open source WEKA data mining techniques, we can generate predictive model to classification of blood group.**

***Keywords* -** Data mining, heterozygous, homozygous, Data Preprocessing, Machine Learning.

I INTRODUCTION

Data mining applications are widely use in various areas such as e-business, education, engineering, biotechnology, medical science etc. This paper focus about genetic disease known as SCD. SCD is a hereditary anemia, predominantly seen amongst various tribal populations of India. SCD is found all over the world, particularly amongst people migrated from Malaria. According to the hypothesis, it is a natural mutation in Hemoglobin molecule to protect RBCs from malarial vermin by making them a little rigid, so that malarial vermin cannot enter into RBCs[4]. Sickle Cell gene is mainly present amongst tribal people, who originated from malaria endemic forest areas. SCD occurs due to inherited abnormal hemoglobin (Hb) gene, which produce Hb-S (Hb-Sickle). Due to the presence of Hb-S and because of its abnormality converts RBCs into rigid-brittle half moon and it is known as Sickle. This sickle shaped instead of normal soft round shape, which is the main cause of complication of Sickle Cell disease [5]. The diagnosis of SCD is only possible by carrying out a simple special blood test known as Sickling test on RBCs. Sickle gene is transmitted from parent to child. If either of the parent is having Sickle gene, then the child may be normal or Sickle trait (heterozygous) and if both the parent are having Sickle gene, then the child may be Sickle disease (homozygous) or Sickle Trait or Normal. Prevention of Sickle cell disease is that the child birth is only possible by marriage counseling and prenatal diagnosis.



AA -> Indicate Normal Child, AS-> Sickle Trait Child, SS-> Sickle Disease Child

Fig. 1. Shows Sickle cell Trait/Disease/Normal

II SCOPE AND PURPOSE OF RESEARCH

Massive healthcare data needs to be converted into information and knowledge, which can help to control cost and maintains high quality of patient care. Without data mining it is difficult to realize the full potential of data collected within healthcare organization as data under analysis is massive, highly dimensional, distributed and uncertain. The traditional manual data analysis has become insufficient and methods for efficient computer assisted analysis indispensable. Application of data mining, knowledge discovery and database techniques are very beneficial but highly challenging in the field of medical and health care.

SCD is highly affected in southern region of Gujarat specifically in tribal zone. Through my research we can collect information from various sources eventually we can aware maximum SCD patient regarding disease and kind of prevention that they have to consider for rest of life.

My proposed research will be helpful to the society, medical sector and Government department for future enhancement and improvement in the current system. Proposed system can discover unexpected relationships, discover pattern on the basis of different attributes. Using classification techniques we can blood group, finally we can predict the disease and classify the patients who are more prone to the disease.

III DATA MINING RESEARCH IN HEALTHCARE

In healthcare thousand of records are being captured for healthcare processes in the form of Electronic Records(ER). As a result, data mining has become critical to the healthcare world. In my research work we covered more than 1,00,000 records, which is related to SCD. Through Data Mining Techniques we can do the following.

- Carry out statistical analysis of healthcare data
- Mining healthcare data for improved patient care and it will helpful for cost-reduction
- Data quality assessment and we can do preprocessing, cleaning, missing data treatment etc.
- Pattern detection from observational data
- Health Information exchanges
- Classification trees are used for the kind of Data Mining problem which are concerned with prediction.

My research is related to prediction of Sickle Cell Disease, in that classification tree is most suitable methods using WEKA Data Mining J48 and Random tree Algorithm. In this paper I compared J48 and Random tree classification techniques for mining process.

IV WEKA DATA MINING TOOL

There are several open source Data Mining tools like WEKA, TANAGRA available. WEKA is powerful Data Mining Tool specifically for classification model.

The Waikato Environment for Knowledge Analysis(WEKA) came about as the best in machine learning. WEKA would not only provide a tool for learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation. Nowadays, WEKA is recognized as a landmark system in data mining and machine learning.

Data can be loaded in WEKA from various sources, including files, URLs and databases. Supported file formats include own ARFF format, CSV, and C4.5‟s format. The main interface in WEKA is the Explorer. It has a set of panels, each of which can be used to perform a different task. Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis.

WEKA is a collection of machine learning algorithms for Data Mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. For our purpose the classification tools is applicable. WEKA has four different modes to do work in.

• Simple CLI: It is environment for you to provide a simple command-line interface that allows direct execution of WEKA commands.

• Explorer: It is an environment for exploring data.

• Experimenter: Ii is an environment for performing experiments and conduction of statistical tests between learning schemes.

• Knowledge Flow: Presents a “data-flow” inspired interface to WEKA.

WEKA support various algorithm for generate mining models required by researcher like clustering, classification etc. My research classification techniques are best among all techniques available in WEKA. In Classification, training examples are used to learn a model that can classify the data samples into known classes. For classification we used Decision tree algorithm.

Decision trees are tree-shaped structures that represent decision sets. It generate rules, which then are used to classify data[3]. Decision trees are the favored technique for building understandable models. Decision trees are a way of representing a series of rules that lead to a class or value. A decision tree partitions the input space of a data set into mutually exclusive regions, each of which is assigned a label, a value or an action to characterize its data points.

The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made. A decision tree is a tree structure consisting of internal and external nodes linked by branches. An internal node is a decision making unit that evaluates a decision function to determine which child node to visit next. The external node, on the other hand, has no child nodes and is connected with a label or value that characterizes the given data that leads to its being visited.

Related to most of the other techniques whether it is primarily classification or prediction, the decision tree support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each result leads to a further analysis to help classify or identify the data so that it can be categorized, so that a prediction can be made based on each result.

Using Decision Tree we can discover unexpected relationships, identify subgroup differences, use categorical or continuous data and accommodate missing data. For classification WEKA have different types of algorithm.

J48

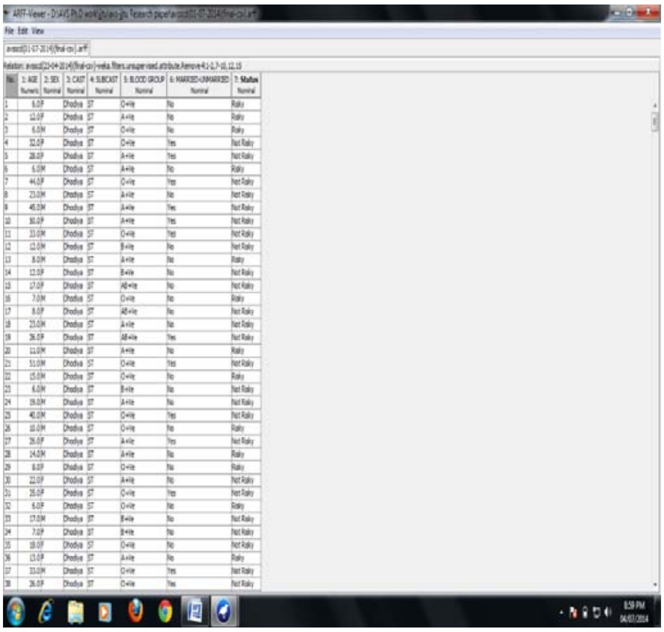
It builds the decision tree from training data set using information gain and it examines the same that results from choosing an attribute for splitting the data. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class. In WEKA J48 Algorithm is also known as C4.5 implementation. For classification in WEKA I go through Knowledge Flow component. Using Knowledge Flow component we can perform various test in a way to do better classification.

J48 is slightly modified as C4.5 in WEKA. The C4.5 algorithm generates a classification and produce decision tree for the given data set by recursive partitioning of data. The decision is developed using Depth-first strategy [6]. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain.

RandomTree

Random Tree is an algorithm for constructing a tree that considers K random features at each node[4]. It performs no pruning. WEKA Random tree generate full classification for each node.

For implementation of J48 and Random tree I used following set of Data. Following Fig. 2 is snapshot of .Arff files for my research.



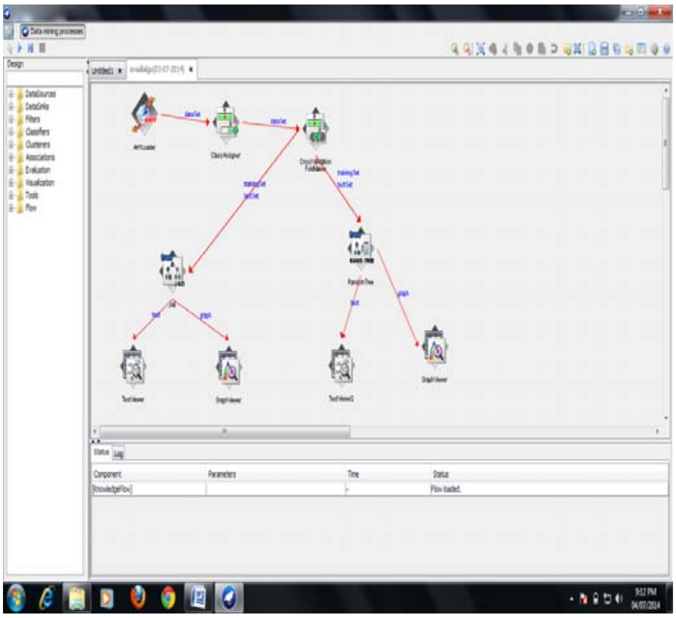
ARFF-Viewer - D:\AVS PhD work\gfu\avs\gfu Research paper\avsscd(01-07-2014)(final-csv).arff

Relation: avsscd(21-04-2014)(final-csv)-weka.filters.unsupervised.attribute.Remove-R1-2,7-10,12,15

No.	1: AGE Numeric	2: SEX Nominal	3: CAST Nominal	4: SUBCAST Nominal	5: BLOOD GROUP Nominal	6: MARRIED-UNMARRIED Nominal	7: Status Nominal
1	6.0	F	Dhodiya	ST	O+Ve	No	Risky
2	12.0	F	Dhodiya	ST	A+Ve	No	Risky
3	6.0	M	Dhodiya	ST	O+Ve	No	Risky
4	32.0	F	Dhodiya	ST	O+Ve	Yes	Not Risky
5	28.0	F	Dhodiya	ST	A+Ve	Yes	Not Risky
6	6.0	M	Dhodiya	ST	A+Ve	No	Risky
7	44.0	F	Dhodiya	ST	O+Ve	Yes	Not Risky
8	23.0	M	Dhodiya	ST	A+Ve	No	Not Risky
9	45.0	M	Dhodiya	ST	A+Ve	Yes	Not Risky
10	50.0	F	Dhodiya	ST	A+Ve	Yes	Not Risky
11	33.0	M	Dhodiya	ST	O+Ve	Yes	Not Risky
12	12.0	M	Dhodiya	ST	B+Ve	No	Not Risky
13	8.0	M	Dhodiya	ST	A+Ve	No	Risky
14	12.0	F	Dhodiya	ST	B+Ve	No	Not Risky
15	17.0	F	Dhodiya	ST	AB+Ve	No	Not Risky
16	7.0	M	Dhodiya	ST	O+Ve	No	Risky
17	8.0	F	Dhodiya	ST	AB+Ve	No	Not Risky
18	23.0	M	Dhodiya	ST	A+Ve	No	Not Risky
19	26.0	F	Dhodiya	ST	AB+Ve	Yes	Not Risky
20	11.0	M	Dhodiya	ST	A+Ve	No	Risky
21	51.0	M	Dhodiya	ST	O+Ve	Yes	Not Risky
22	15.0	M	Dhodiya	ST	O+Ve	No	Risky
23	6.0	M	Dhodiya	ST	B+Ve	No	Not Risky
24	19.0	M	Dhodiya	ST	A+Ve	No	Not Risky
25	40.0	M	Dhodiya	ST	O+Ve	Yes	Not Risky
26	10.0	M	Dhodiya	ST	O+Ve	No	Risky
27	26.0	F	Dhodiya	ST	A+Ve	Yes	Not Risky
28	14.0	M	Dhodiya	ST	A+Ve	No	Risky
29	8.0	F	Dhodiya	ST	O+Ve	No	Risky
30	22.0	F	Dhodiya	ST	A+Ve	No	Not Risky
31	25.0	F	Dhodiya	ST	O+Ve	Yes	Not Risky
32	6.0	F	Dhodiya	ST	O+Ve	No	Risky
33	17.0	M	Dhodiya	ST	B+Ve	No	Not Risky
34	7.0	F	Dhodiya	ST	B+Ve	No	Not Risky
35	18.0	F	Dhodiya	ST	O+Ve	No	Not Risky
36	13.0	F	Dhodiya	ST	A+Ve	No	Risky
37	33.0	M	Dhodiya	ST	O+Ve	Yes	Not Risky
38	26.0	F	Dhodiya	ST	O+Ve	Yes	Not Risky

Fig.2 .Arff files

Using WEKA knowledge flow we get following mining model for comparison of two classification methods.



Component	Parameters	Time	Status
[KnowledgeFlow]		-	Flow loaded.

Fig. 3 Predictive model of J48 and Randomtree

After generating predictive model we can produce graph viewer and text viewer classifier model using WEKA. Fig. 4 and 5 demonstrate classifier model of J48 and Random tree and it shows that Random tree algorithm produced details decision tree compare to J48, which is very much useful for further classification of each node.

```
=== Classifier model ===

Scheme:   J48
Training Fold: 10

J48 pruned tree
------------------

AGE <= 15
|   BLOOD GROUP = O+Ve: Risky (148.51)
|   BLOOD GROUP = A+Ve: Risky (114.0)
|   BLOOD GROUP = B+Ve: Not Risky (85.0)
|   BLOOD GROUP = AB+Ve: Not Risky (35.0)
|   BLOOD GROUP = A-Ve: Not Risky (3.0)
AGE > 15: Not Risky (377.49/0.49)

Number of Leaves:        6

Size of the tree:   8
```

Fig. 4 Classifier model using J48

```
=== Classifier model ===

Scheme:   RandomTree
Training Fold: 10

RandomTree
==========

MARRIED-UNMARRIED = No
|   AGE < 15.5
|   |   BLOOD GROUP = O+Ve : Risky (148.74/0)
|   |   BLOOD GROUP = A+Ve : Risky (114/0)
|   |   BLOOD GROUP = B+Ve : Not Risky (85/0)
|   |   BLOOD GROUP = AB+Ve : Not Risky (35/0)
|   |   BLOOD GROUP = A-Ve : Not Risky (3/0)
|   AGE >= 15.5
|   |   SEX = F : Not Risky (62/0)
|   |   SEX = M
|   |   |   BLOOD GROUP = O+Ve
|   |   |   |   SUBCAST = ST
|   |   |   |   |   CAST = Dhodiya : Not Risky (17.26/0.26)
|   |   |   |   |   CAST = Koli : Risky (0/0)
|   |   |   |   |   CAST = Halpati : Not Risky (1/0)
|   |   |   |   |   CAST = Nayka : Not Risky (4/0)
|   |   |   |   |   CAST = Gayakvad : Risky (0/0)
|   |   |   |   |   CAST = Kukana : Not Risky (4/0)
|   |   |   |   |   CAST = Varli : Not Risky (4/0)
|   |   |   |   |   CAST = Mushlim : Risky (0/0)
|   |   |   |   |   CAST = Gamit : Not Risky (1/0)
|   |   |   |   |   CAST = Chodhri : Risky (0/0)
|   |   |   |   |   CAST = Khunbhar : Risky (0/0)
|   |   |   |   |   CAST = Kotvadiya : Risky (0/0)
|   |   |   |   SUBCAST = OBC : Not Risky (1/0)
|   |   |   |   SUBCAST = SC : Risky (0/0)
|   |   |   BLOOD GROUP = A+Ve : Not Risky (22/0)
|   |   |   BLOOD GROUP = B+Ve : Not Risky (11/0)
|   |   |   BLOOD GROUP = AB+Ve : Not Risky (9/0)
|   |   |   BLOOD GROUP = A-Ve : Risky (0/0)
|   |   SEX = f: Risky (0/0)
|   |   SEX = m: Risky (0/0)
MARRIED-UNMARRIED = Yes : Not Risky (241/0)

Size of the tree: 34
```

Fig. 5 Classifier model using Random tree

V CONCLUSION

This paper emphasis on comparison of two classification techniques J48 and Random tree using WEKA. Implementation of both classification algorithms, I can classify specific blood group with respect to Age as dependent variable. Random tree produce depth decision tree respect to J48 and that will helpful to researcher. From tested data we can conclude that those specific blood groups have more chances of SCD.

Future scope and enhancement in this research we can develop predictive model that can analysis tested data and it will be helpful for medical science and government sector.

ACKNOWLEDGMENT

I wish to acknowledge and thanks very much to my Research guide Dr. Jigar V. Patel, Associate Professor, Kalol Institute of Technology & Research Centre, Kalol for his kind support and encouragement.

Without support and encouragement of my entire CASPS family it won't be possible. I wish to acknowledge my entire CASPS family which helps me to work hard towards producing this work.

REFERENCES

[1] "Data Mining - Typical Data mining Process for Predictive Modeling",BPB Publications, First Edition 2004 -REPRINTED 2007, ISBN 81-7656-927-5

[2] Dr. Carlo kopp, "Data Mining and Knowledge Discovery Techniques"

[3] Yongheng Zhao and Yanxia Zhang , "Comparison of decision tree methods for finding active objects", Accepted for publication in Advances of Space Research

[4] "Sickle Cell Anemia", Gujarati Version, Hani Printers, Ahemedabad

[5] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald and David Scuse, "WEKA Manual for Version 3-7-10" ,2013

[6] Mohd Fauzi bin Othman, Thomas Moh Shan Yau, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer", IFMBE Proceedings Vol. 15, pp. 520-523, 2007,

[7] Trilok Chand Sharma and Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013, pp 1925-1932, ISSN (Print) : 2319-5940

[8] P ooja Sharma and Asst. Prof. Rupali Bhartiya, "Implementation of Decision Tree Algorithm to Analysis the Performance", International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 10, December 2012, ISSN (Print) : 2319-5940 , pp 831-864.

[9] "Sickle Cell Disease and Sickle Cell Anaemia" by Dr. Louise Newson, emis press, pp 1-8