IJMEIT// Vol.03 Issue 05//May//Page No:1274-1282//ISSN-2348-196x 2015

International Journal Of Mechanical Engineering And Information Technology

An Official Publication Of IGM Publication

Extraction of Rules For Predicting HIV Infections And Computing Support And Confidence

Authors

Amol Joglekar[1], Dr.G. Prasanna Lakshmi[2], Maunash Jani[3]

[1]Research-Scholar, Pacific University Udaipur

[2]Professor and Guide

Email- [1]amol.joglekar@gmail.com,[2]prasannagandi@yahoo.com,[3]tmaunash08@gmail.com

ABSTRACT

Data mining is a dominant step for the processing of large database. Medical science has a huge database and day by day it increases. We, therefore need an efficient tool which will extract the correct symptoms for identifying the disease. There are many symptoms which might be same for number of diseases. The number of patterns or rules which can be extracted from algorithm may be many out of which we need to select the exact rules and other rules can be deleted. Here we need remarkable Measures which will narrow down our rules and provide the best one. This paper proposes an idea of computing support, confidence and lift for the rules generated using an algorithm which helps us to predict infection of HIV.

Keywords— *HIV, support, confidence, lift, if-rules.*

INTRODUCTION

HIV stands for human immune deficiency virus. This particular virus was identified in the 1980s and belongs to a group of viruses called 'retroviruses'. HIV attacks the immune system, and gradually causes damage. This implies, without a proper treatment and care, a person with HIV is at risk of developing serious infections and may lead to death. HIV cannot be completely removed but if it is detected at early stages we can ensure of good and healthy life in next future. There is a rapid and fast development in medical sector. Day by day new diseases and hence new treatment comes which need to be registered. Hence a field of Computer Science and Information Technology comes into picture. This field has a role of computerizing the medical data and give support to medical field.

One of the trend in Computer Science / IT is data mining which plays a major role in identifying the patterns based on history and provides accurate knowledge. Such automated system is needed in identifying HIV like diseases.

BACKGROUND AND RELATED WORK

The proposed thesis paper is about developing an algorithm using data mining predictive tools and techniques which will extract the correct rules from the given datasets. There are many diseases like cancer, heart attack etc. are having symptoms. Due

to this reason there is a need of studying current techniques in order to propose futuristic scope. Medical science has great potential for finding out hidden information or domain. The data should be extracted in a particular format so that it can be used to dig out the knowledge. G.Parthiban, A. Rajesh , S.K.Srivasta [1] used Naïve Bayes theorem used to predict attributes such as age, sex, blood pressure, blood sugar and chance of a diabetic patient getting a heart problem. They got data from a diabetic research centre Chennai. Using WEKA they did analysis of 50 records of patients. WEKA tool has a collection of machine learning algorithms written in JAVA. It also has many features like data pre-processing, clustering, visualization tool etc. they used Naïve theorem for finding out the exact knowledge and classify accordingly. Their experiment had shown that 74% of results were proper and this method was compared with other methods showed proper results.

Aqueel Ahmed, Shaikh Abdul Hannan [2] had mentioned different types of techniques for predicting Heart Disease. They had suggested some important parameters like age, sex, chest pain etc. Author had just explained various predictive algorithms in theoretical way and claimed Decision Tree and Support vector machine are most effective techniques for the same. They would like to increase the accuracy of proposed model by selecting more number of parameters.

Jyoti Soni, Ujma Ansari, Dipesh Sharma [3] proposed for predicting the heart disease using the association rule data mining technique. Awkwardly they have produced a large number of rules when association rules are applied on medical data set. Most of the rules were not required for the same. Authors had developed a system based on some manual inputs and in future they would like to automate the same. Also they could not able to extract data from real life and therefore if they extract medical data from some real database then it would have an added advantage.

Sellappan Palaniappan and Rafiah Awang [4] proposed a prototype model for identifying Heart Disease using different predictive algorithms. Using this model they wanted to find out solutions for complex questions so that effective solution can be given to patients. Author had decided to work on five different goals and for that they had collected data from Cleveland database. Goals were tested using there different predictive techniques like Decision Tree, Naïve Bayes and Neural Networks. They had selected thirteen attributes and conclude that Naïve Bayes performs well with 96.6% accuracy.

Rosma Mohd Dom, Sameem Abdul Kareem, Basir Abidin, Adeeba Kamaruzaman, Annapurni Kajindaran [5] mentioned the feasibility of applying predictive data mining technique for survival of AIDS. They used adaptive fuzzy technique. For the survival of AIDS patients authors had used different attributes like CD4, CD8 and viral load count i.e. the amount of HIV infection blood. They measured the ability of FuReA with fuzzy neural network and found that the accuracy of FuReA was 60-100% based on selected datasets.

Dr. K. Rameshkumar [6] developed a model using ARM(Association Rule Mining) to extract valuable information from database. Author has proposed a new algorithm which would take care of missing values for detecting HIV AIDS. With the help of this proposed algorithm author could able to extract information about CD4 cell counts, RNA levels and treatment given for various patient. The model is lacking of handling data with a very good accuracy.

Agbonifo Oluwatoyin C and Ajayi Adedoyin C[7] developed a predictive model which determines a predicative model determining the life of a cattle. The authors have considered various diseases like Antrax, Babesiosis, Backleg etc. They focused on fuzzy logic to identify the type of disease. They have constructed various parametric values for symptoms and hence rules are formed. They claimed that the model is fast and optimized. The working of the system can be enhanced using real life data and Neural Network.

M. Mayilvaganan and K. Rajeswari[8] presented the analysis of risk factor using fuzzy logic. Authors have constructed innovative model using logic gates. The parameters like blood pressure, pulse rate and GFR rate of kidney were taken into consideration. Different values for parameters were taken into consideration and presented in the form of tables. They claimed that their model accurately predicts the depression risk level based on expert knowledge using Sugeno FIS Method.

Mir Anamul Hasan, Khaja Md. Sher-E-Alam and Ahsan Raja Chowdhury [9] proposed a model for identifying Human Disease using fuzzy logic. Authors have developed a model which will work on user interaction. Based on the selection of problems there will be problem defined area like chest, legs etc. For the selected area only there will be the listing of symptoms. The rules are constructed using Fuzzy Logic to predict the disease. They have claimed the confidence of system using rules and charts.

RESEARCH METHODOLOGY

A. Block Diagram / System Architecture

The model of proposed system is shown in figure: 1. It consists of three major sections called input section, logic section and prediction section. This system defines the mapping of input data into prediction rules with the help of a proposed algorithm. The input section will accept data in the form of symptoms associated with a disease. Based on the symptoms there will be weights allocated to each symptom. The proposed algorithm will perform function prescribed in a logic section so that the corresponding prediction will determine the output. In order to produce some predictions we need to take the help of rules which has the combination of symptoms and it will help to produce the forecast.

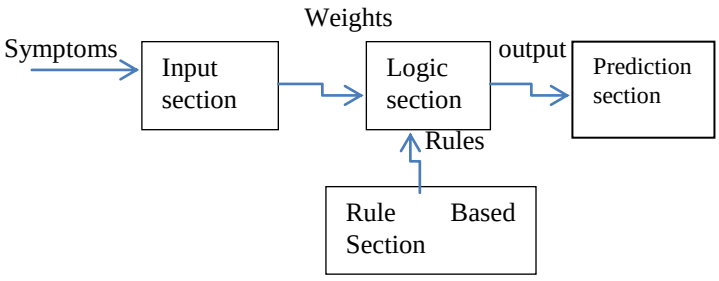

Figure 1: Schematic diagram

B. Rule Based System

Data mining is a step of discovering association rules between items in large databases. The main reason is to find hidden relations between items of various transaction of database. The most salient area is processing of knowledge and the analysis of interestingness of discovered patterns. Association rule is very important tool for mining process it has two special characteristics support and confidence. Support gives total number of transaction of any particular item are occurring in datasets while confidence gives strength of a data in a dataset, we can say support is probability of A and B while confidence is conditional probability. Association rule can be constructed with the help of support and confidence. In order to improve the performance of an association rule one more term can be used called lift.

Support is defined on item sets and gives the proportion of transactions which contain X. It is used as a measure of significance (importance) of an item set. Since it basically uses the count of transactions called a frequency constraint. An itemset with a support greater then a set minimum support threshold, $supp(X)>\sigma$, is called a frequent or large item set. The term was introduced by R. Agrawal, T. Imielinski, and A. Swami. Mining associations between sets of items in large databases. In Proc. of the ACM SIGMOD Int'l Conference on Management of Data, pages 207-216, Washington D.C., May 1993.

$Supp(X)=|\{t\in D;X\subseteq t\}||D|=P(EX)$

Hence support is the percentage of transactions that demonstrate the rule.

Confidence is defined as the probability of seeing the rule's consequent under the condition that the transactions also contain the antecedent. Confidence is directed and gives different values for the rules $X\rightarrow Y$ and $Y\rightarrow X$. Association rules have to satisfy a minimum confidence constraint, $conf(X\rightarrow Y)\geq\gamma$. The term was introduced by R. Agrawal, T. Imielinski, and A. Swami. Mining associations between sets of items in large databases. In Proc. of the ACM SIGMOD Int'l Conference on Management of Data, pages 207-216, Washington D.C., May 1993.

$conf(X\rightarrow Y)=supp(X\rightarrow Y)supp(X)=supp(X\cup Y)supp(X)=P(EX\cap EY)P(EX)=P(EY|EX)$

Hence The confidence is the conditional probability that, given X present in a transition , Y will also be present.

Confidence measure, by definition:

Confidence(X=>Y) equals support(X,Y) / support(X)

Lift measures how many times more often X and Y occur together than expected if they were statistically independent. It was introduced by S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic item set counting and implication rules for market basket data. In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '97), pages 265-276, 1997.

$lift(X\rightarrow Y)=lift(Y\rightarrow X)=conf(X\rightarrow Y)supp(Y)=conf(Y\rightarrow X)supp(X)=P(EX\cap EY)P(EX)P(EY)$

Hence lift is a measure which predicts the performance of an association rule in order to boost response.

C. Proposed Algorithm

The proposed approach is helpful in identifying the presence of HIV infection in a patient. As a result, medical conclusions, treatment procedures and decisions can be made by practitioners accurately.

S : Stage

D : Disease

L : Symptom

W : Weight of Symptom

S_v : Scaling Variable

t1,t2 : Temporary Variable

l_i : Likert Scale

Input L for S

Select S_v for L

Scan all D ϵ L

Choose W for L and accumulate it with S_v

for S=1 to n

do

for all D ϵ S_n

do

for all L ϵ D

do

t1= $\sum_{i=1}^{k} W * Sv$

for each subset l_i ϵ L

do

t2= $Max(\sum_{i=1}^{m} li) * Lw$

end for

for each S

do

suspicion= (t1 / t2) * 100

end for

end for

end for

D. Experimental Data

The proposed algorithm uses transaction tables. There are numerical values allocated to each symptom and severity. The aim is to classify the diseases and predict the presence of HIV infection into the patient by analyzing and computing the item sets on the basis of proposed algorithm. The rules are formed and computation of support, confidence and lift is measured. Following are the sample transaction table and computation of support, confidence and lift.

Table 1: Sample Table

Transaction ID	Vomiting	Headache	Fever
1	0	0	0
2	1	1	1
3	1	1	1
4	0	1	1
5	0	1	1
6	0	1	1
7	0	1	1
8	0	1	1

Table 2: Sample Table

Transaction ID	Joint Pain	Muscle Pain
1	1	1
2	0	0
3	0	0
4	0	0
5	1	1
6	0	1
7	1	1
8	0	1

Table 3: Sample Table

Transaction ID	Diarrhoea	Rash
1	0	0
2	0	0
3	1	0
4	1	0
5	1	0
6	1	1
7	1	1
8	1	1

Table 4: Sample Table

Transaction ID	Swollen Glands	Weight Loss
1	0	0
2	0	0
3	0	0
4	0	0
5	1	0
6	0	0
7	1	1
8	0	1

Table 5: Sample Table

Transaction ID	Weight Loss	Rash
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	1	1
8	1	1

Table 6: Sample Table

Transaction ID	Weight Loss	Rash	Swollen Glands
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	1	1
6	0	1	0
7	1	1	1
8	1	1	0

Table 7: Sample Table

Transaction ID	Vomiting	Headache	Fever	Diarrhoea
1	0	0	0	0
2	1	1	1	0
3	1	1	1	1
4	0	1	1	1
5	0	1	1	1
6	0	1	1	1
7	0	1	1	1
8	0	1	1	1

Table 8: Sample Table

Transaction ID	Headache	Fever	Diarrhoea	Rash
1	0	0	0	0
2	1	1	0	0
3	1	1	1	0
4	1	1	1	0
5	1	1	1	0
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1

Table 9: Sample Table

Transaction ID	Headache	Fever	Joint Pain
1	0	0	1
2	1	1	0
3	1	1	0
4	1	1	0
5	1	1	1
6	1	1	0
7	1	1	1
8	1	1	0

Table 10 : Support and Confidence Table

Rules	Symptoms	Support	Confidence	Lift
R1	Headache, Fever, Joint Pain	0.25	0.285	0.761
R2	Headache, Fever, Diarrhea, Rash	0.375	0.5	1.333
R3	Vomiting, Headache, Fever, Diarrhea	0.125	0.5	0.666
R4	Weight Loss, Rash, Swollen Glands	0.125	0.5	2
R5	Vomiting, Headache, Fever	0.25	1	4
R6	Joint Pain, Muscle Pain	0.375	1.5	1.60
R7	Diarrhea, Rash	0.375	1	1.33
R8	Swollen Glands, Rash	0.25	1	2
R9	Weight Loss, Rash	0.25	1	4

Rules: Some of the rules showing the relationship between the sets as shown in above tables are discussed below.

Rules used in Algorithm

Rule 1: If Headache=TRUE AND Fever=TRUE AND Joint Pain=TRUE THEN D1=STIFFNESS AND D2=MENINGITIS AND D3= RHEUMATIC FEVER AND D4=' RHEUMATOID ARTHRITIS' AND D5='INFLUENZA'

Rule 2: If Headache=TRUE AND Fever=TRUE AND Diarrhea=TRUE AND Rash=TRUE THEN D1=MENINGITIS AND D2= INFLUENZA D3= MALARIA AND D4= PNEUMONIA AND D5=DENGUE FEVER AND D6='SHINGLES' AND D7= 'MYCOBACTERIUM AVIUM COMPLEX'

Rule 3: If Vomiting=TRUE AND Headache=TRUE AND Fever=TRUE AND Diarrhea=TRUE THEN D1=MIGRAINE AND D2=INFECTIUOS DIARRHEA AND D3= DEATH CAP POISONING AND D4= ALCOHOL WITHDRAWAL SYNDROME AND D5='INFLUENZA' AND D6='HERBICIDE POISONING'

Rule 4: If Weight Loss=TRUE AND Rash=TRUE AND Swollen Glands=TRUE THEN D1= MYCOBACTERIUM AVIUM COMPLEX AND D2= INFECTIUOS DIARRHEA AND D3= PHARYNGITIS AND D4= SHINGLES AND D5='INFLUENZA' AND D6='HUMAN PAPILLOMA VIRUS'

Rule 5: If Vomiting=TRUE Headache=TRUE AND Fever=TRUE AND THEN D1=MIGRAINE AND D2=ELECTROLYTE DISTURBANCES AND D3= THYROTOXICOSIS AND D4= MENINGITIS AND D5='INFLUENZA' AND D6='INFECTIOUS DIARRHEA'

Rule 6: If Joint Pain=TRUE AND Muscle Pain=TRUE THEN D1=STIFFNESS AND D2=MENINGITIS AND D3= RHEUMATIC FEVER AND D4= 'RHEUMATOID ARTHRITIS' AND D5='INFLUENZA'

Rule 7: If Diarrhea=TRUE AND Rash=TRUE THEN D1= OBSTRUCTING JAUNDICE AND D2= INFECTIUOS DIARRHEA AND D3= MYCOBACTERIUM AVIUM COMPLEX AND D4= PRIMARY BILIARY CIRRHOSIS AND D5=SHINGLES AND D6= 'HMF DISEASE' AND D7= 'CANDIDIASIS'

Rule 8: If Swollen Glands=TRUE AND Rash=TRUE THEN D1= INFLUENZA AND D2=PHARYNGITIS AND D3= COMMON COLD AND D4=SHINGLES AND D5= HMF DISEASE AND D6= 'CANDIDIASIS' AND D7='INFLUENZA'

Rule 9: If Weight Loss=TRUE AND Rash=TRUE THEN D1= OBSTRUCTING JAUNDICE AND D2= INFECTIUOS DIARRHEA AND D3=SHINGLES AND D4= 'HMF DISEASE' AND D5= 'CANDIDIASIS'

RESULTS AND FUTURE SCOPE

The system is tested with the help of GUI based model which consists of various modules and testing of the same. The patient has to undergo a particular procedure where he has to select the symptoms. Based on his choice there will be the set of selected

rules from the database and predict the disease with percentage. The future work will be the design of complete module which would include patient history and recording of different laboratory tests. Also to test the accuracy of proposed algorithm on different platform will be a challenge.

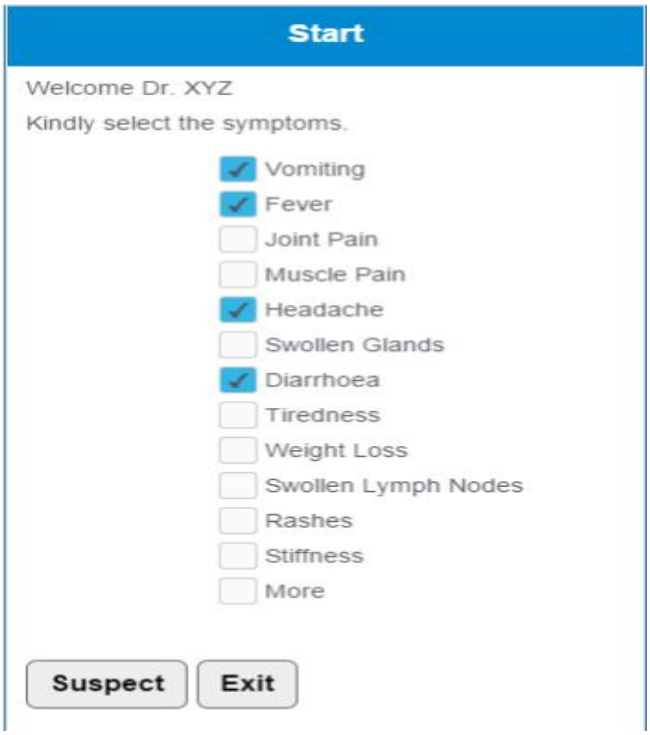

Figure 2: GUI interface 1

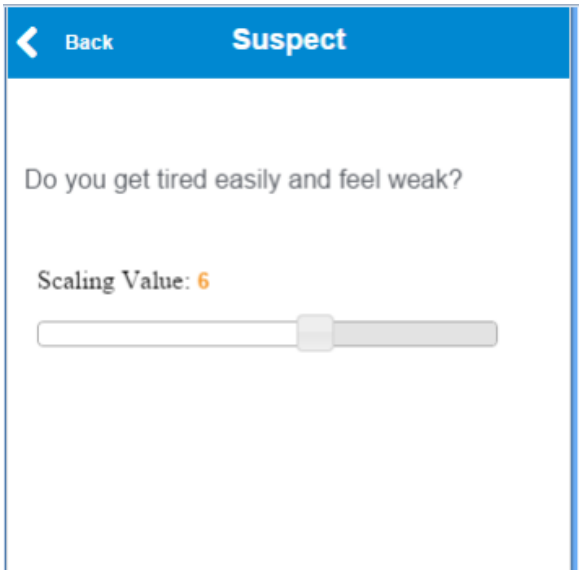

Figure 3: scalar module

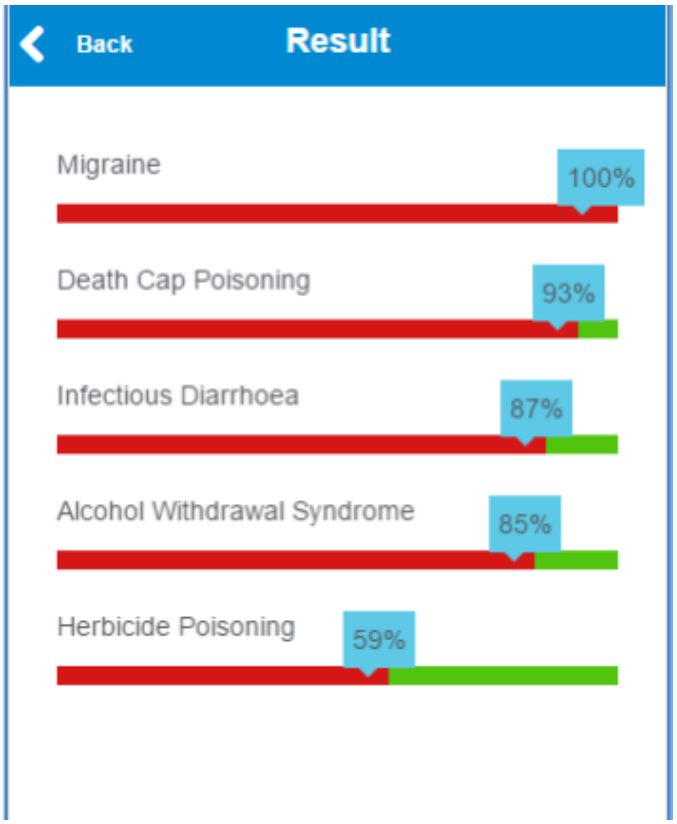

Figure 4: Suspect Module

CONCLUSION

The focus of this paper is to provide fast and user friendly module for the diagnosis of HIV. To achieve this we have proposed if-rules which will generate some rules for identifying the disease. The research work gives an idea of support, confidence and lifts which are necessary for the analyzing. More rules can be formed and generated which will accurately predict the disease.

REFERENCES

1. G.Parthiban, A. Rajesh , S.K.Srivasta "Diagnosis of Heart Disease for Diabetic Patients Using Naïve Bayes Method". International Journal of Computer

Applications(0975-8887), Vol.24, No.3, June 2011.

2. Aqueel Ahmed, Shaikh Abdul Hannan "Data Mining Techniques to Find Out Heart Disease: An Overview" International Journal of Innovative and Exploring Engineering (IJITEE), ISSN:2278-3075, Vol. 1, Issue 4, September 2012.
3. Jyoti Soni, Ujma Ansari, Dipesh Sharma "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Application (0975-8887), Vol. 17, No.18, March 2011.
4. Sellappan Palaniappan and Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques" International Journal of Computer Science and Network Security (IJCSNS) , Vol.8, No. 8, August 2008.
5. Rosma Mohd Dom, Sameem Abdul Kareem, Basir Abidin, Adeeba Kamaruzaman, Annapurni Kajindaran "The Prediction of AIDS Survival: A Data Mining Approach" Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering" ISBN: 978-960-474-083-3.
6. Dr. K Rameshkumar "Association Rules Mining from HIV/AIDS patient's case history database with missing values" International Journal on Data Mining and Intelligent Information Technology Applications" Vol.2 , No. 1,pp. 18-24, March 2012.
7. Agbonifo Oluwatoyin C and Ajayi Adedoyin C " Design of a Fuzzy Expert Based System for Diagnosis of Cattle Diseases" International Journal of Computer Applications and Information Technology Vol. I, Issue III, November 2012 (ISSN :2278-7720)
8. M. Mayilvaganan and K. Rajeswari "Risk Factor Analysis to Patient Based on Fuzzy Logic Control System" International Journal of Engineering Research and General Science Volume 2, Issue 5, August-September,2014 ISSN 2091-2730.
9. Mir Anamul Hasan, Khaja Md. Sher-E-Alam and Ahsan Raja Chowdhury "Human Disease Diagnosis Using a Fuzzy Expert System" Journal of Computing, Volume 2, Issue 6, June 2010,ISSN 2151-9617.
10. www.avert.org
11. www.health.com
12. www.aidsprogramme.ukzn.ac.za
13. Priyanka Sharma, DBV Singh, Manoj Kumar Bandil, Nidhi Mishra "Decision Support System for Malaria and Dengue Disease Diagnosis(DSSMD) "International Journal of Information and Computation Technology ISSN0974-2239 ,Vol.3, Number 7(2013) pp.633-640.