International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Android Based Questionnaires Application for Heart Disease Prediction System

Nan Yu Hlaing[1], Phyu Pyar Moe[2]

[1]Assistant Lecturer, Myanmar Institute of Information Technology, Mandalay, Myanmar
[2]Assistant Lecturer, Sagaing Education College, Sagaing, Myanmar

How to cite this paper: Nan Yu Hlaing | Phyu Pyar Moe "Android Based Questionnaires Application for Heart Disease Prediction System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.1988-1991, **https://doi.org/10.31142/ijtsrd26750**

IJTSRD26750

ABSTRACT

Today classification techniques in data mining are most popular to prediction and data exploration. This Heart Disease Prediction System (HDPS) is using Naive Bayesian Classification with a comparison for simple probability and that of Jelinek-Mercer (JM) Smoothing. It is implemented as an Android based application: user must be feedback and answers the questions then can be seen the result as user desired in different ways: exactly heart disease is present or not and then with predictions (No, Low, Average, High, Very High). And the system will be provided required suggestions such as doctor details and medications to patients could be able. It will be also proved that enhanced Naive Bayes with Jelinek-Mercer smoothing technique is also effective to eliminate the noise for prediction the heart disease. This system can also calculate classifier accuracy by using precision and recall.

KEYWORDS: *Heart disease, Naïve Bayes, Smoothing, Android, precision, recall*

1. INTRODUCTION

Data mining is the technique to search the relationships and patterns from the large database. Heart diseases remain the main cause of death worldwide and possible detection at an earlier stage will prevent the attacks. Sometime clinical decisions that are based on doctors' intuition and experience. But it can cause unwanted biases, large amount of medical costs which affects the quality of service provided to patients, so need to predict heart disease before they occur in their patients. For this purpose, HDPS converts the unused data into a dataset for modeling using mining techniques, can be used by practitioners and also patients themselves. This android based HDPS system helps and advices the patient to safe, decrease unwanted practice variation, to improve patient outcome, how to care and information about doctors and clinic.

2. THEORY BACKGROUND

Data mining is a knowledge discovery technique to analyze data from large data sets. It is a computational process of finding patterns in large data sets including methods at the intersection of machine learning, artificial intelligence, statistics and database systems. One of the main focuses of data mining process is to obtain information from the data and converted it into can knowledgeable and reasonable structure for further use [1].

Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known [3].

2.1 Handling Missing Attribute Values

Methods of handling missing attribute values in data mining are categorized into sequential and parallel.

1. In sequential methods, known value is replaced at the missing attribute values.
2. In parallel methods, knowledge from the original data sets is used directly.
3. In sequential methods, missing attribute value is converted into completer data set to handle missing attribute values original in-complete data sets e.g., rule induction, is conducted.

Table 1 is represented as a simple example with the attributes *OldPeak*, *Slope*, and *CA* and with the Predicted Attribute. However, data sets are incomplete, so that missing attribute values are presented by "?"S [2].

In this method, an arithmetic mean of known values is used for all missing attribute value for a numerical attribute. In Table 1, the mean of known attribute values for Old Peak is 2.2; hence all missing attribute values for *Old Peak* should be replaced by 2.2. This mean of known attribute values for slope is 2 and CA is 1. Table 2 is represented after replacing with mean value for missing attribute values.

Table 1 Sample data set with a numerical attribute before replacing

Case	Old Peak	Slope	CA	Predicted Attribute
1	2.3	3	?	0
2	1.5	2	3	2
3	?	?	2	1
4	3.5	3	0	0
5	1.4	1	?	0
6	0.8	1	0	0
7	3.6	3	2	3
8	?	1	?	0
9	1.4	?	1	2
10	3.1	?	0	1

Table2. After replacing with mean value for missing attribute values

Case	Old Peak	Slope	CA	Predicted Attribute
1	2.3	3	1	0
2	1.5	2	3	2
3	2.2	2	2	1
4	3.5	3	0	0
5	1.4	1	1	0
6	0.8	1	0	0
7	3.6	3	2	3
8	2.2	1	1	0
9	1.4	2	1	2
10	3.1	2	0	1

2.2 Introduction of Naive Bayes

The Bayes theorem was developed and named for THOMAS BAYES."Naive" because it is based on independence assumption. Describes what makes something "evidence" and how much evidence it is. Bayesian Classifiers are *statistical* classifiers. They can *predict the probability* that a data item is a member of a particular class.

Original Belief + Observation = New Belief (1)

Naive Bayes or Bayes' Rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. Why naive Bayes implementation is most prefer:

1. When the data is high
2. When the attributes are independent of each other
3. When we expect more efficient output, as compared to other methods output

2.2.1. Attributes' Probability of Naive Bayes

Bayes' theorem can be stated and estimate
$P(a_i|v_j)$ using m-estimates [3]:

$$P(ai\ |vj) = \frac{nc+mp}{n+m} \quad (2)$$

Where:n = the number of training examples for which $v=v_j$
n_c = number of examples for which $v=v_j$ and $a=a_i$
p = a priori estimate for $P(a_i|v_j)$
m = the equivalent sample size

2.2.2. Naive Bayes Classifier

The Bayes Naive classifier selects the most likely classification V_{nb} given the attribute values a_1 , a_2 , …, a_n results in [4]:

$$V_{nb} = arg\ _{max} v_j \in v,\ P(v_j) \prod P(a_i|v_j) \quad (3)$$

2.3 Smoothing Methods

To remove and avoid noise or other fine-scale structures/rapid occurrences in data smoothing is attempts to capture those noise by using an approximating function. It refers to the adjustment of maximum likelihood estimator for the model so that it will be more accurate. The estimator generally under estimate the probability of unseen data. The main purpose of the smoothing is to provide a non-zero probability to unseen data and improve the accuracy of probability estimator [5].

2.3.1. Probability of Jelinek-Mercer (JM) Smoothing

$\boldsymbol{P(x \mid C_i)}$ is calculated by Jelinek-Mercer smoothing,

$$P(x \mid C_i) = (1-\lambda)\ P(x \mid C_i) + \lambda\ P(x \mid C) \quad (4)$$

Where, $\boldsymbol{P(x \mid C_i)}$ = probability of smoothened recorda
λ = balancing parameter between 0 and 1, and $\boldsymbol{P(x \mid C)}$ = estimated attribute in class $\boldsymbol{C}$.

2.3.2 Naive Bayes Classifier with smoothed Probability

To classify an attribute d, Naive Bayes (NB) is used and finds the class $\boldsymbol{C_i}$ that maximizes [3] [4]:

$$C_{nb} = P(C_i)\ P(d \mid C_i)$$

Where, $P(d \mid Ci) = \prod_{k=1}^{|d|} p(xk|ci)$ (5)

3. STATE OF THE PROBLEM

There are many kinds of health disease prediction system with various methods. Heart disease is very important and it might need to help immediately but they are not available as need as due to many reasons. People cannot identify and measure their symptoms, cannot carry medicines anywhere and without consulting by doctors. This system solves the problems and predicts whether a patient has heart disease or not.

3.1 Sample Data and Result

Predictable attribute: User can see the classifier result by choosing one of two options:

1. No - no heart disease; Yes - has heart disease
2. Represents the possibility of heart disease: No, Low, Average, High, Very high.

Input attributes:

- Age
- Sex
- Chest Pain Type
- Blood Pressure
- Serum Cholesterol
- Fasting Blood Sugar
- Rest ECG
- Thalach
- Exang
- Oldpeak
- CA
- Thal

To predict the Health Disease, this system used some symptoms of patient. For example: age, sex, blood pressure and blood sugar, chest pain, ECG graph data etc., It will be implemented in Android based application which takes medical test's parameter as an input. It can predict state of heart disease either Yes/No or No/Low/Average/High/Very High that user want to and depend on the predicted result,

system can allow to view suggestions and address of doctors and clinic center. It can be used as a training tool to train nurses and medical students to diagnose patients with heart disease [6] [8].

4. PROPOSE SYSTEM

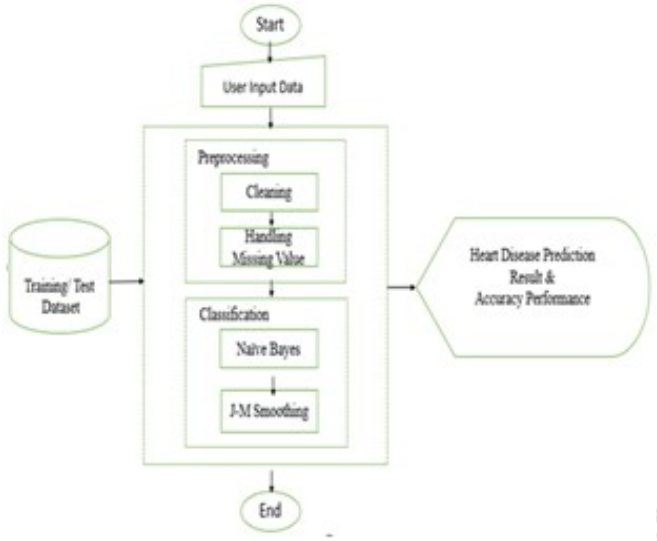

Fig1: System Flow Diagram for Heart Disease Prediction System

5. EXPERIMENT

5.1 Performance Measures

1. Performance Evaluation Metric

That focus on the predictive capability of a model rather than how fast it takes to classify or build models, scalability, etc.

2. Most widely-used two-class metric

	Predicted class		
		Class=Yes	Class=No
Actual class	Class=Yes	a(TP)	b(FN)
	Class=No	c(FP)	d(TN)

In the classification with two-classes, positive and negative, a single prediction has four possibilities.

1. The True Positive rate (TP) and True Negative rate (TN) are correct classifications.
2. A False Positive (FP) occurs when the outcome is incorrectly predicted as positive when it is actually negative.
3. A False Negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive.

Since the class label prediction is of multi-class, the result on the test set will be displayed as a two-dimensional confusion matrix with a row and column for each class. Each matrix element shows the number of test cases for which the actual class is the row and the predicted class is the column [10].

Accuracy - It refers to the total number of records that are correctly classified by the classifier.

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Precision - is the fraction of retrieved instances that are relevant.

$$Precision = \frac{TP}{TP+F} \quad (7)$$

Recall - is the fraction of relevant instances that are retrieved [7].

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

5.5.1. Sample Calculation with Naïve Bayes and Jelinek-mercer smoothing

ForNo(0)
P(0)*P(57|0)*P(0|0)*P(2|0)*P(130|0)*P(236|0)*P(0|0)*P(2|0)*P(174|0)*P(0|0)*P(0|0)*P(2|0)*P(1|0)*P(3|0)
=0.32*0.6781818182*0.6920634921*0.5224489796*0.2875*0.5298892989*0.4385093168*0.311111111*0.6617511521*0.675*0.337254902*0.3025641026*0.7351955307=164/303
=2.95789E-0.5 (Simple Naïve Bayes)
=4.29913E-0.7 (Naïve Bayes with Jelinek-mercer smoothing)

ForLow(1)
P(57|1)*P(0|1)*P(2|1)*P(130|1)*P(236|1)*P(0|1)*P(2|1)*P(174|1)*P(0|1)*P(0|1)*P(2|1)*P(1|1)*P(3|1)
=0.32*0.1054545455*0.1365079365*0.1346938776*0.225*0.1977859779*0.2149068323*0.255555555*0.150230414*0.175*0.2196078431*0.2769230769*0.1374301676
=0.0000000003332112
=6.034829E-10 (Simple Naïve Bayes)
=1.02E-11 (Naïve Bayes with Jelinek-mercer smoothing)

ForAverage(2)
=P(57|2)*P(0|2)*P(2|2)*P(130|2)*P(236|2)*P(0|2)*P(2|2)*P(174|2)*P(0|2)*P(0|2)*P(2|2)*P(1|2)*P(3|2)
=0.12*0.0872727273*0.0571428571*0.1551020408*0.1625*0.1092250923*0.1155279503*0.1444444444*0.0764976959*0.0589285714*0.1869281046*0.2128205128*0.0536312849
=2.6441343E-13 =55/303
=3.14155E-14 (Simple Naïve Bayes)
=4.30178E-12 (Naïve Bayes with Jelinek-mercer smoothing)

ForHigh(3)
=P(57|3)*P(0|3)*P(2|3)*P(130|3)*P(236|3)*P(0|3)*P(2|3)*P(174|3)*P(0|3)*P(0|3)*P(2|3)*P(1|3)*P(3|3)
=0.12*0.0872727273*0.073015873*0.1142857143*0.1625*0.1092250923*0.1527950311*0.1444444444*0.067281106*0.0589285714*0.1738562092*0.1358974359*0.0480446927
=1.54070966E-13 ==35/303
=1.7797E-14 (Simple Naïve Bayes)
=3.26594E-12 (Naïve Bayes with Jelinek-mercer smoothing)

ForVeryHigh(4)
=P(57|4)*P(0|4)*P(2|4)*P(130|4)*P(236|4)*P(0|4)*P(2|4)*P(174|4)*P(0|4)*P(0|4)*P(2|4)*P(1|4)*P(3|4)
=0.12*0.0418181818*0.0412698413*0.0734693878*0.1625*0.0538745387*0.0782608696*0.1444444444*0.0442396313*0.0321428571*0.082529412*0.0717948718*0.025698324
=1.54070966E-13 =13/303
=5.01259E - 29 (Simple Naïve Bayes)
= 2.26922E-13 (Naïve Bayes with Jelinek-mercer smoothing)
=0.0000295789

The Result is 0. Predited Attribute is No.

5.5.2. Suggestion of the System

Do and Don't	Doctors and Clinic Centre	Address
#24 Hr Hotline phone no. (09259898661, 259898662)		
1. Take a walk 2. Stop smoking 3. Cut out or least cut down high fat fast foods 4. Avoid watching >2hrs of TV a day 5. Cut down on beverages and fruit juices with added sugar 6. Use mono-unsaturated and polyunsaturated fat 7. Use food with low salt and try to add less salt while cooking 8. Constant noise 9. Hormonal therapy 10. Air pollution 11. Anger 12. Avoid alcohol or drink in moderation preferably wine 13. Avoid food which have low nutrition value like sodas and candy	1. Dr. Than Than Kyaing (Chan Nyein Aung, Aryuthukha, Kankaw, Nandaw) 2. Dr. Kyaw Soe Win 3. (Chan Nyein Aung, ryuthukha) 4. Dr. Myint Ngwe (Chan Nyein Aung, Nyein) 5. Dr. Aung Thu (Nandaw) 6. Dr. Khin Maung Htwe (Nandaw, Aryuthukha) 7. Dr. Khin OoLwin (Aryuthukh) 8. Dr. Tun Shwe (Aryuthukha, Nyein)	**Chan Nyein Aung** (No.115, 74 Street, Between 27*28 Street, Mandalay, Ph. no:0272885, 72886, 72887) **Aryuthukha** (No.150, 74 Streets, Between 30*31 Street, Mandalay, and Ph. no: 0274482, 74483) **Kankaw** (34 Street, Between 78*79 Street, Mandalay, Ph. no: 0233258, 66880) **Nandaw** (71 Street, Between 28*29 Street, Mandalay, and Ph. no: 0260443, 60445, 60446) **Nyein** (No.333, 82 Street, Between 29*30 Street, Mandalay, and Ph. no: 0232050, 34795, 65460)

6. CONCLUSION

This system is the Heart Disease Prediction System by using Naive Bayesian and Jelinek-Mercer smoothing technique. Android HDPS accepts patients' dataset as system inputs and will be used 13 attributes of medical diagnosis. This system will be performed as an experiment on application of two data mining algorithms to predict the heart disease and to provide the comparison results of prediction and effectiveness of Simple Naive Bayes and enhanced Naive Bayes with J-M Smoothing. This system is to enhance the performance of Naïve Bayes Classifier in classifying patients' dataset with a modification of Jelinek-Mercer Smoothing and this system will also present and redirect the related suggestions for the patients.

REFERENCES

[1] Jiawei HanUniversity of Illinois at Urbana–ChampaignMicheline KamberJian PeiSimon Fraser University "Data MiningConcepts and Techniques Third Edition".

[2] Jerzy W. Grzymala-Busse University of Kansas "Handling Missing Attribute Values".

[3] http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf

[4] http://www.cs.cmu.edu/~10601b/slides/NBayes.pdf

[5] http://mlwiki.org/index.php/Smoothing_for_Language_Models

[6] https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118.

[7] https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall.

[8] https://www.medicinenet.com/symptoms_of_serious_diseases_and_health_problems/article.htm#15_signs_and_symptoms_of_a_heart_attack

[9] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques".

[10] http://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf