Performance Comparison of KNN, NB and DT Classifiers using Heart Disease Dataset

Yin Yin Htay
Faculty of Information Science, Computer University (Magway), Myanmar
yinyinhtayyyh2019@gmail.com

Ya Min
Department of Computer Science, Computer University (Lashio), Myanmar
yamin1977lso@gmail.com

Abstract

Today, the diagnosis of diseases is a vital and intricate job in medicine. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. In this situation, an automatic medical diagnosis system is beneficial by bringing all of them together. For this diagnosis system, many classifiers are essential and needed for disease classification. So, this system is proposed as the performance comparison system about classifiers to know which classifier is more effective than other. To compare the performance, this system classifies the heart disease dataset by using K-nearest neighbor(KNN), naive bayesian (NB) and decision tree (DT) classifiers.

Keywords: Comparison, KNN, NB, DT.

1. Introduction

Healthcare industry generates the large amount of data about patient, disease diagnosis etc. However, there is a lack of effective analysis tools to discover hidden relationships in data. Data mining provides a set of techniques to discover hidden patterns from data. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. A knowledge discovery process includes data cleaning, data integration, data selection, data mining, pattern evaluation and knowledge presentation. Data mining system can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used or the applications adapted. Data mining is the process of classification, association rule mining, clustering, etc. K-nearest neighbor(KNN), naive bayesian (NB) and decision tree (DT) classifiers are the most popular algorithms in the mining classification.

Major challenge facing Healthcare industry is quality of service. Quality of service implies diagnosing disease correctly and provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences which are unacceptable.

So, this system is proposed to predict whether the patient is having heart disease or not by using KNN, NB and DT classifiers that are data mining techniques. After classifying according to these three classifiers, this system compares the accuracy and processing time of each classifier. According to the comparison results, the healthcare industry can easily know which classifier is most effective for diagnosis system. In the remote areas like rural regions or country sides, the proposed system is also a user friendly, scalable and reliable system that can be implemented to imitate like human diagnosis expertise for treatment of heart disease.

2. Related Work

In 2016, M. Panwar, A. Acharyya and R. A. Shafik [1] presented a new methodology based on novel preprocessing techniques, and K-nearest neighbor classifier. The effectiveness of the proposed methodology is validated with the help of various quantitative metrics and a comparative analysis, with previously reported studies using the same UCI dataset focusing on pima-diabetes disease diagnosis.

In 2016, D. VijayaKumar and V. J. R. Krishniah [2] used decision tree classification model for diagnosis of three brain diseases namely ischemic stroke, hemorrhage and hematoma, and tumor. This system helped the physicians to identify the type of human brain hemorrhage and hematoma and the type of brain tumors for further treatment.

In 2017, N. R. Gorrepati and N. R.Uppala [3] compared the performances of different classifiers on diagnosis of the Erythemato-Squamous disease. The classifiers examined here are support vector machine (SVM), discriminant classifier, K-nearest neighbor and decision tree. They have performed their analysis with two well-known multiclass techniques.

3. K-Nearest Neighbor Classifier

When an unknown feature is introduced, the k-nearest neighbour (KNN) classifier finds k most similar training

features that are closet to the unknown feature [4]. The procedure of K-NN classifier is as follow [5]:

- Determine k. Calculate the similarity or distance between the testing data and all the training data.

$$d_{euc} = \sqrt{\sum_{j=1}^{N} | P_j \quad Q_j |^2} \qquad (1)$$

  where d_{euc} is the distance between the test and training data, P_j is the feature j of test data P, and Q_j is the feature j of training data Q.
- Sort the distance and determine k nearest neighbors based on the K^{th} minimum distance.
- Gather the categories based on majority vote.
- Determine categories based on majority vote.

4. Naive Bayesian Classifier

The processing steps of Naive bayesian (NB) classifier is as follows [6]:

1. Each data sample is represented by n-dimensional feature vector, X = (x_1, x_2… x_n) depicting n-measurements made on the sample from n- attributes, A_1, A_2,….., A_n.
2. Suppose that there are m classes, C_1, C_2, …, C_m. Given an unknown data sample X, Naïve Bayesian classifier assigns an unknown sample X to the class C_i if and only if
   $P(C_i \backslash X) > P(C_j \backslash X)$ for $1 \le j \le m, j \ne I$ (2)
   The class C_i for which $P(C_i \backslash X)$ that is maximized, called maximum posteriori hypothesis. By Bayes theorem,
   $P(C_i \backslash X) = P(X \backslash C_i)\, P(C_i)/ P(X)$ (3)
3. As P(X) is constant for all classes, only $P(X \backslash C_i)\, P(C_i)$ need to be maximized.
4. Given data sets with many attributes, the Naive assumption of class conditional independence is made. Thus,
   $P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i)$ (4)
   The probability $P(x_1 \backslash C_i)$, $P(x_2 \backslash C_i)$,… $P(x_n \backslash C_i)$ can be estimated from the data samples.
5. In order to classify an unknown sample X, $P(X \backslash C_i)\, P(C_i)$ is evaluated for each class C_i,. Sample X is then assigned to the class C_i if and only if
   $P(X \backslash C_i)\, P(C_i) > P(X \backslash C_j)\, P(C_j)$ (5)
   In other words, it is assigned to the class C_i for which $P(X \backslash C_i)\, P(C_i)$ is the maximum [5].

5. Decision Tree Classifier

In decision tree (DT) classifier, information gain measure (entropy) is used to select the test feature at each node in the decision tree. DT algorithm is as follows:
Algorithm : Generate_decision_tree.

- create a node *N*;
- if *samples* are all of the same class, *C* then
  return *N* as a leaf node labeled with class *C*;
- if *attribute-list* is empty then
  return *N* as a leaf node labeled with the most common class in *samples*;
- select *test-attribute*, the attribute among *attribute-list* with the highest information gain;
- label node *N* with *test-attribute*;
- for each known value a_i of *test-attribute*
  grow a branch from node *N* for the condition *test-attribute=*a_i;
- let s_i be the set of samples in *samples* for which *test-attribute=*a_i;
- if s_i is empty then
  attach a leaf labeled with the most common class in *samples*;
- else attach the node returned by Generate _decision _tree;

Information gain measure is used to select the test feature at each node in the tree with the highest information gain. Let s_i be the number of samples of S in class C_i. The information needed to classify a given sample is given by

$$I(s_1, s_2, \ldots, s_m) = \sum_{i=1}^{m} p_i \log(p_i) \qquad (6)$$

where p_i is the probability that an arbitrary sample belongs to C_i and is estimated by s_i / s. Let attribute A have v distinct values, $\{a_1, a_2, \ldots, a_v\}$. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}, \ldots, s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}) \qquad (7)$$

For a given subset S_i,

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = \sum_{i=1}^{m} p_{ij} \log(p_{ij}) \qquad (8)$$

Encoding information would be gained by branching on A:

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A) \qquad (9)$$

The feature with the highest information gain is chosen as the test feature for given set S [5].

6. Proposed System Design

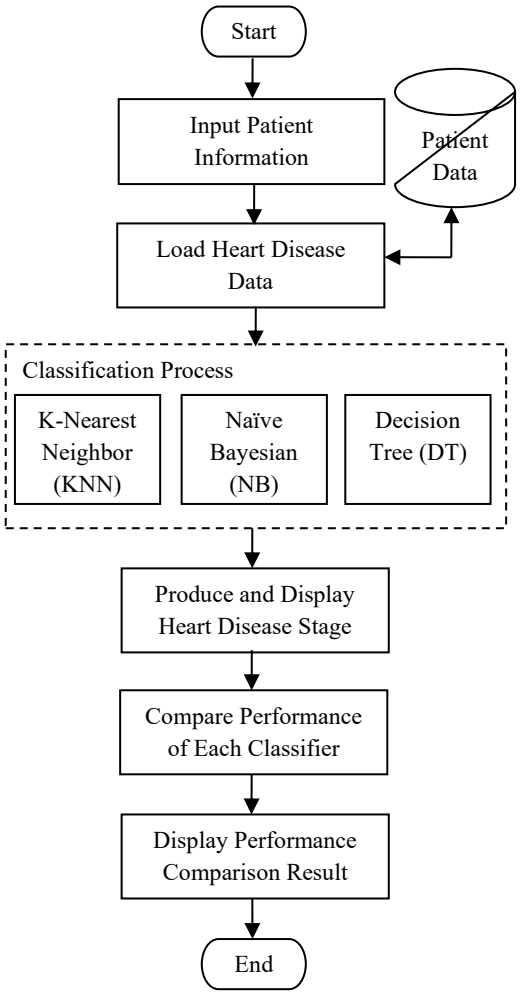

Figure 1. Proposed System Design

After classifying, this system produces and displays the heart disease stage. Then, this system measures the processing time of each classier and calculates the accuracy of each classifier by using Holdout method. Finally, this system displays the performance comparison result to the user.

6.1. Heart Disease Attributes

For classification, this system uses the heart disease dataset. In this dataset, there are 13 (symptoms) attributes and heart disease class. The heart disease data is retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Attributes and its information are shown in Table 1.

Table 1. Attributes and Its Information

ID	Attribute Name	Attribute Value
1	Age	20-80
2	Sex	1: Male 0: Female
3	Chest Pain Type	1.0 - 4.0
4	Trestbps (Blood Pressure)	0.0-1.0
5	Chol (Serum Cholesterol)	126.0 – 564.0
6	Fasting Blood Sugar	94.0 - 200.0
7	Restecg (resting electrographic results)	0.0 – 2.0
8	Thalach (maximum heart rate achieved)	71.0 – 202.0
9	Exang (exercise induced angina)	0.0 – 1.0
10	Oldpeak (ST depression induced by exercise relative to rest)	0.0 – 6.2
11	Slope (the slope of the peak exercise ST segment)	1.0 – 3.0
12	Ca (number of major vessels colored by floursopy)	0.0 – 3.0
13	Thal	3.0 – 7.0
14	Class	Level I Level II Level III Level IV

7. Explanation of the System

As a sample, this system uses 10 records that are obtained from 10 patients who suffer heart disease. Heart disease dataset includes five class levels that are Normal (N), Level I (I), Level II (II), Level III (III) and Level IV (IV). Sample heart disease dataset is shown in Table 2.

Table 2. Sample Heart Disease Dataset

Age	Sex	Chest Pain Type	Trestbps	Chol	Fasting Blood Sugar	Rest ECG	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Class
63	1	4	0	260	140	1	112	1	3	2	0	0	II
44	1	4	0	209	130	1	127	0	0	0	0	0	N
60	1	4	0	218	132	1	140	1	1.5	3	0	0	II
55	1	4	0	228	142	1	149	1	2.5	1	0	0	I

66	1	3	1	213	110	2	99	0	1.3	2	0	0	**N**
66	1	3	0	0	120	1	120	0	0.5	1	0	0	N
65	1	4	1	236	150	1	105	1	0	0	0	0	III
62	1	3	0	0	180	1	140	1	1.3	2	0	0	N
60	1	3	0	0	120	0	141	1	2	1	0	0	III
60	1	2	1	267	160	1	157	0	0.5	2	0	0	I

The patient input heart disease symptom into the system. The inputted information includes age (60), sex (1), chest pain type (4), trestbps (0), chol (260), fasting blood sugar (140), restECG (1), thalach (140), exang (1), oldpeak (1.5), slope (3), ca (0) and thal (0). Then, this system calculates and classifies the heart disease stage that the patient suffers.

7.1 KNN Classification Process

By using training and testing data, this system calculates the distance between each data. After calculating, this system chooses the most similar training class for the testing data. KNN classifier results are shown in Table 3.

Table 3. KNN Classifier Results

ID	Test and Training Data	Distance Result
1	Test and Training Data 1	24.5
2	Test and Training Data 2	95.5
3	Test and Training Data 3	50
4	Test and Training Data 4	27
5	Test and Training Data 5	113.2
6	Test and Training Data 6	299
7	Test and Training Data 7	47.5
8	Test and Training Data 8	220.2
9	Test and Training Data 9	282.5
10	Test and Training Data 1	40

According to the KNN classifier, this system produces the heart disease stage **"Level II"** to the patient.

7.2 DT Classification Process

To produce the decision rules, this system calculates the gain for each attribute. The attribute is chosen as root node if this attribute has highest gain result. In this sample. This system obtains the decision tree after finishing second iteration. According to the decision tree, this system produces the decision rules to produce the result. Decision tree is shown in Figure 2.

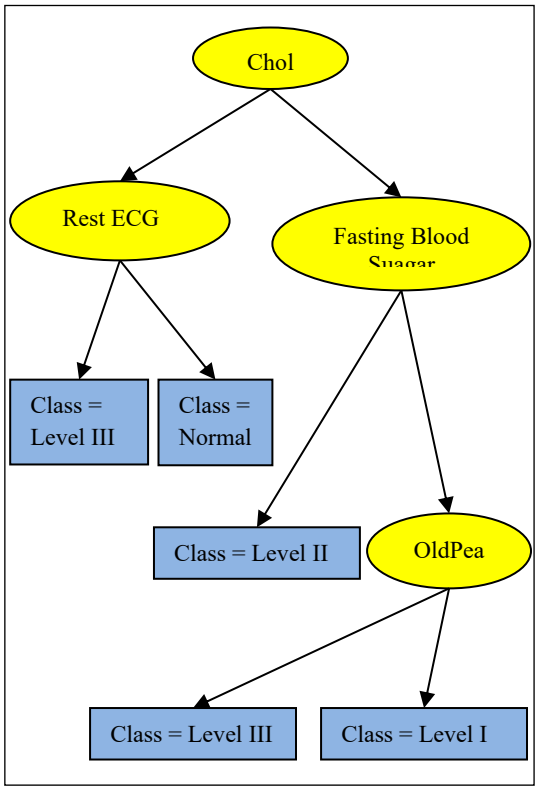

Figure 2. Decision Tree

Decision rules are generated from the decision tree. Rule 1 is {IF "Chol <= 163.1" AND "RestECG <= 1" THEN "Class= Level III"}. Rule 2 is {IF "Chol <= 163.1" AND "RestECG > 1" THEN "Class= Normal"}. Rule 3 is {IF "Chol > 163.1" AND "Fasting Blood Sugar <= 141" THEN "Class = Level II"}. Rule 4 is {IF "Chol > 163.1" AND "Fasting Blood Sugar > 141" AND "OldPeak <= 1.5" THEN "Class = Level III"}. Rule 5 is {IF "Chol > 163.1" AND "Fasting Blood Sugar > 141" AND "OldPeak > 1.5" THEN "Class = Level I"}. According to the Rule 3, the heart disease level that the patient suffered is **"Level II"**.

7.3 NB Classification Process

According to naive bayesian classifier, this system calculates probability of each attribute. NB classifier result is shown in Table 4.

Table 4. NB Classifier Results

ID	$P(X\|C_i)\ P(C_i)$	Probability Result
1	P(X|Class = I) P(Class = I)	0
2	P(X|Class = II) P(Class = II)	0.00312
3	P(X|Class = III) P(Class = III)	0
4	P(X|Class = N) P(Class = N)	0

The heart disease level that the patient suffered is **"Level II".**

8. Experimental Result of the System

To compare the performance of the system, this system uses the heart disease dataset from the UCI website. This system is tested 500 records. To compare the performance of each algorithm, this system calculates the accuracy and measures the processing time of each algorithm. For accuracy calculation of each classifier, this system uses the hold out method. Accuracy result of the system is shown in Table 5 and processing time result of the system is shown in Table 6.

Table 5. Accuracy Result of the System

Test Data	Accuracy (%)		
	Decision Tree (DT)	Naïve Bayesian (NB)	K-Nearest Neighbor (KNN)
150	97%	90%	91%
220	92%	91%	92%
270	89%	82%	85%
300	87%	84%	85%
330	83%	80%	81%

After calculating the performance, decision tree is the most precise classifier among other classifier. Then, KNN classifier is more precise than NB classifier. Sometimes, NB classifier can face problem about probability calculation of each attribute. But, the processing time of NB and KNN is less than the processing time of DT. Finally, DT is precise and takes more processing time than NB and KNN.

Table 6. Processing Time Result of the System

Test Data	Processing Time (ms)		
	Decision Tree (DT)	Naïve Bayesian (NB)	K-Nearest Neighbor (KNN)
150	30	15	14
220	38	16	15
270	42	16	15
300	46	19	17
330	47	20	18

9. Conclusion

This system is proposed an effective heart disease prediction system by using k-nearest neighbor, naïve bayesian and decision tree classifiers. This system is also implemented to show which classifier is more than another. This system points out the decision tree classier that is more precise than other classifier and KNN classifier that can quickly produce the heart disease result. Finally, this system is helpful for practice to confirm his/ her decision during heart disease prediction.

References

[1] M. Panwar, A. Acharyya and R. A. Shafik, "K-Nearest Neighbor Based Methodology for Accurate Diagnosis of Diabetes Mellitus", IEEE, 2016.

[2] D. VijayaKumar and V. J. R. Krishniah, "An Automated Framework for Stroke and Hemorrhage Detection using Decision Tree Classifier", IEEE, 2016.

[3] N. R. Gorrepati and N. R.Uppala, "Comparative Performance Analysis of Different Classifiers on Diagnosis of Erythmato-Squamous Diseases", International Conference on Innovations in Information, Embedded and Communication Systems,
IEEE, 2017.

[4] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, pp. 243-247, 2019.

[5] H. Jiawei and K. Micheline, "Data Mining Concepts and Techniques", Simon Fraser University, United States of America, 2001.

[6] M.S. Basarslan and I. D. Argun, "Classification of A Bank Data Set on Various Data Mining Platforms", IEEE, 2018.