Performance of Naïve Bayes, C4.5 and KNN using Breast Cancer, Iris and Hypothyroid Datasets

K. Pazhani Kumar, R. Raja Aswathi

***Abstract:** Data mining usually specifies the discovery of specific pattern or analysis of data from a large dataset. Classification is one of an efficient data mining technique, in which class the data are classified are already predefined using the existing datasets. The classification of medical records in terms of its symptoms using computerized method and storing the predicted information in the digital format is of great importance in the diagnosis of various diseases in the medical field. In this paper, finding the algorithm with highest accuracy range is concentrated so that a cost-effective algorithm can be found. Here the data mining classification algorithms are compared with their accuracy of finding exact data according to the diagnosis report and their execution rate to identify how fast the records are classified. The classification technique based algorithms used in this study are the Naive Bayes Classifier, the C4.5 tree classifier and the K-Nearest Neighbor (KNN) to predict which algorithm is the best suited for classifying any kind of medical dataset. Here the datasets such as Breast Cancer, Iris and Hypothyroid are used to predict which of the three algorithms is suitable for classifying the datasets with highest accuracy of finding the records of patients with the particular health problems. The experimental results represented in the form of table and graph shows the performance and the importance of Naïve Bayes, C4.5 and K-Nearest Neighbor algorithms. From the performance outcome of the three algorithms the C4.5 algorithm is a lot better than the Naïve Bayes and the K-Nearest Neighbor algorithm.*

***Keywords:** Data mining, Classification, Naive Bayes Classifier, C4.5 Algorithm, K-Nearest Neighbor Algorithm.*

I. INTRODUCTION

The word “mining” means the analysis of a large number of base materials which have a long process with the literature of other Disciplines such as artificial intelligence, statistics and database [1][4]. Of all the existing techniques, the distinguishes perceptions of data mining is the development of data mining techniques are applied to the database application on the large scale that turns on the application of large-scale data which can provide a lot of new challenges that could ultimately bring new methodologies [2][3].

Revised Manuscript Received on January 06, 2020.

Dr. K. Pazhanikumar, Assistant Professor, Department of Computer Science, S. T. Hindu College, Nagercoil, Affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.
Email: skpk73@gmail.com.

R. Raja Aswathi, Department of Computer Science, S. T. Hindu College, Nagercoil, Affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu India. Email:rajaaswathi@outlook.com

Data mining techniques are used in various field of science for an efficient prediction of results using the data mining algorithms. Some of the data mining methodologies that are most common in practice are classification, clustering, association rules, regression, prediction, sequential pattern analysis and others. The classification technique in data mining uses some of the mathematical concepts like the decision tree, probability, neural networks, distance metrics, linear programming and statistics.

Developing treatments and better understanding the characteristics of the diseases is almost exclusively based on the clinical and biological research [5]. Algorithm based research can provide new opportunities to develop data from the traditional research approaches. By using data mining technique the most challenging diseases outcome can be analyzed using the experimental historical data.

II. CLASSIFICATION TECHNIQUES

A. Naive Bayes Classifier

Naive Bayes classifier classifies the occurrence of values in a dataset using their probability value and also by counting the values combinations and their associated frequency. The naive theorem is combined with an attribute condition where it is assumed to be independent [6]. The Bayes theorem calculates the posterior probability P(Ci|x) from P(Ci), P(x) and P(x|Ci) as follows:

$$\mathbf{P(Ci|x)=P(x|Ci)*P(Ci) / P(x)}$$

- P(Ci|x)- Posterior Probability
- P(Ci)- Prior Probability
- P(x|Ci)- Conditional Probability
- P(x)- Predictor Prior Probability

Where Ci represents the classes and x represents the vector of attribute value. Naive Bayes uses the below steps to classify the given attributes:

Algorithm:

Input:

```
D    // Training Dataset
X    // Predictor Variable
```

Output:

C // Class of predicting values

Naïve Bayes Algorithm:

Step-1: The training data that are provided are read.

Step-2: Calculate the probability value of data from corresponding same category.

Step-3: Find P(Ci), the total number of classes in the dataset.

Step-4: After finding the Prior Probability, the Conditional Probability value P(x|Ci) for each number of classes that are provided are calculated.

Step-5: Using the Conditional Probability, Prior Probability and Predictor Prior Probability, the Posterior Probability can be predicted.

Step-6: Then choose the maximum probability value P(Ci|x) which is the predicted class of new data.

The naive Bayes classifier operates on a strong independence assumption [7]. Despite of its limitations, naive Bayes was shown to be optimal for some important classes of concepts that have a high degree of feature dependencies, such as disjunctive and conjunctive concepts [8].

B. C4.5 Algorithm

Decision trees are one of the most popular methods for classification in various data mining applications [9][10] and assist the process of decision making [11]. C4.5 builds decision tree from the provided dataset in the same manner as the ID3 approach using the concept of entropy and information gain.

$$\mathbf{H}\ (p_1, p_2 \ldots p_s) = \sum_{i=1}^{s}\left(pi \log\left(\frac{1}{pi}\right)\right)$$

The term p_i denotes the probabilities (p_1 ,p_2, p_3,....,p_s) of the attribute value in the database, D [14]. The information gain is calculated from the following formula,

$$\mathbf{Gain\ (D,\ S) = H\ (D)}\ -\sum_{i=1}^{s} P(Di)H(Di)$$

The C4.5 algorithm builds the decision tree from the given set of training data and splits the training attributes using its information gain. The attributes with highest information gain ratio is chosen as the splitting attribute and the one that is used to make the decision and the second split attribute is the one with the next highest split. The process continues until all the attributes are splitted.

Algorithm:

Input:

D // Training Dataset

Output:

T // Returns decision Tree

C4.5 Algorithm:

```
Start
    Initialize an empty tree;
    Calculate entropy and gain ratio of attributes
If D is empty then
    Return a failure value
Else
    Create a node for attributes
    // Attribute with highest gain ratio value until all are splitted
End If
End
```

This algorithm is efficient in the following ways:

1. The missing data in any records can be predicted from the attribute value of other similar records in the similar dataset.
2. Can deal with ranges of continuous data.
3. Uses the pruning strategies like the subtree replacement and subtree raising.
4. Then generates the result in the form of rules or decision tree from the best splitting attribute.

C. K-Nearest Neighbor (KNN)

In pattern recognition, the KNN algorithm is an efficient method that is used for classifying objects based on closest training samples in the feature space. KNN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification [12][13]. The instances with particular range are combined within one class that are already predefined.

The distance measure used in this algorithm for KNN is the Euclidean distance measure and is predicted using the following method as given below:

$$\mathbf{D(x,y)}=\sqrt{\sum_{i=1}^{n}(xi-yi)}^{\,2}$$

The term d(x, y) in the above formula is the distance between two point x and y in any space. The Euclidean distance is one of the commonly used distance metrics to compute the distance between vectors. This method is more appropriate if the data is not standardized.

The KNN algorithm works as follows:

Algorithm:

Input:

D // Training Dataset

Output:

C // Classes to which the data are assigned

KNN Algorithm:

Step-1: Read the training dataset.

Step-2: Find the value of ‘k’

Step-3: Then calculate the distance between the given input instance and the training data.

Step-4: Now assign the predicted instance to the class with similar distance measure that is predicted already.

III. RESULT AND DISCUSSION

A. Dataset

The datasets used for training in this paper are the breast cancer dataset, iris dataset and hypothyroid dataset. These three datasets are used as training data for the three classification algorithms, Naive Bayes, C4.5 and the distance based KNN. Here the accuracy of each algorithm is compared with one another for each dataset to find which the most accurate one among them. By finding the best one the large amount of data that are being used in medical field and other fields of science can analyzed accurately in a short period of time.

1. Breast Cancer Dataset: This dataset contains 10 attributes with 286 instances.
2. Iris Dataset: This dataset uses 5 attributes with 150 instances.
3. Hypothyroid dataset: This dataset is provided with 30 attributes with 3772 instances.

B. Performance Evaluation Based on Datasets

The version 3.8.3 of the open source software WEKA is used to determine the efficiency of three algorithms using three medical datasets Breast Cancer, Iris and Hypothyroid.

i) Performance Evaluation Based on the Breast Cancer Dataset

The breast cancer dataset is used to classify the accuracy of the algorithm Naive Bayes where the number of instances that are correctly classified are 205 (71.678%) among 286 provided instances and the incorrectly classified instances are 81(28.3217%). The C4.5 algorithm classifies 216 (75.5245%) instances correctly and 70 (24.4755 %) instances incorrectly. The KNN classifies 207 (72.3776 %) instances correctly and 79 (27.6224 %) instances incorrectly. In this experiment the C4.5 algorithm has more accuracy than the other two algorithms.

When compare with other two algorithms the KNN is faster in terms of execution for this dataset. Here the execution time is calculated using seconds, the execution rate 0.00 denotes that the time taken for process is less than milliseconds.

Comparison between Classification Algorithms Using Breast Cancer Dataset with 286 Instances

Table.1.Represents the correctly and incorrectly classified datasets using Breast Cancer Dataset

Test Algorithm	Correctly Classified	Incorrectly Classified	Run time (seconds)
Naive Bayes	205(71.678%)	81(28.3217 %)	0.02
C4.5	216(75.5245 %)	70(24.4755 %)	0.05
KNN	207(72.3776 %)	79(27.6224 %)	0.00

The Table.1 shows the experimental results of correctly classified instances and incorrectly classified instances and the execution time needed to process that dataset using the Naive Bayes Classifier, C4.5 and K-Nearest Neighbor using the Breast Cancer dataset.

The following Fig.1 represents the performance of classification algorithms based on the Breast Cancer dataset.

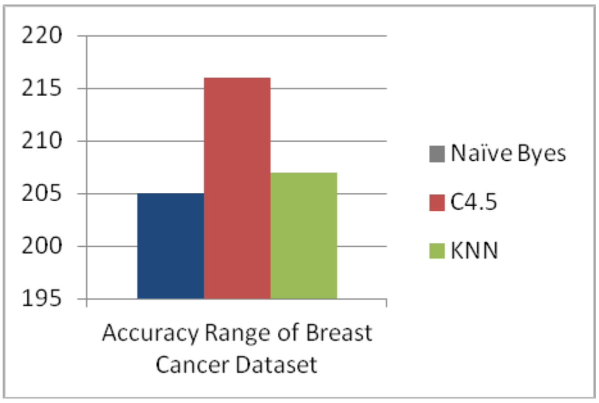

Fig.1 Bar chart representation of three algorithms accuracy using Breast Cancer Dataset

ii) Performance Evaluation Based on the Iris Dataset

Then Iris dataset is used to classify the accuracy of the algorithm Naive Bayes where the number of instances that are correctly classified are 144 (96 %) among 150 instances and the incorrectly classified instances are 6 (4 %). The C4.5 algorithm classifies 144 (96 %) instances correctly and 6 (4 %) instances incorrectly. The KNN classifies 143 (95.3333 %) instances correctly and 7 (4.6667 %) instances incorrectly. Here the Naive Bayes algorithm and the C4.5 algorithm have the highest and similar accuracy than K-Nearest Algorithm.

When compare with other two algorithms the KNN is faster in terms of execution. The execution time taken are represented using seconds, the execution rate 0.00 denotes that the time taken for process is less than milliseconds.

The below Table.2 shows the experimental results of correctly and incorrectly classified instances and the execution time for Naive Bayes, C4.5 and K-Nearest Neighbor using the Iris dataset.

Comparison between Classification Algorithms Using Iris Dataset with 150 Instances

Table.2.Represents the correctly and incorrectly classified datasets using Iris Dataset

Test Algorithm	Correctly Classified	Incorrectly classified	Run time (seconds)
Naive Bayes	144(96 %)	6 (4 %)	0.02
C4.5	144(96 %)	6 (4 %)	0.08
KNN	143(95.33 %)	7(4.6667 %)	0.00

The following Fig.2 represents the performance of classification algorithms based on the Iris dataset.

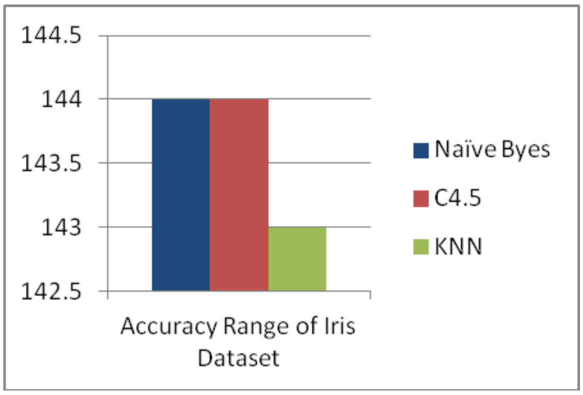

Fig.2 Bar chart representation of three algorithms accuracy using Iris Dataset

iii) Performance Evaluation Based on the Hypothyroid Dataset

The hypothyroid dataset is used to classify the accuracy of the algorithm Naive Bayes where the numbers of instances that are correctly classified are 3594 (95.281 %) among 3772 instances and the incorrectly classified instances are 178 (4.719 %). The C4.5 algorithm classifies 3756 (99.5758 %) instances correctly and 16 (0.4242 %) instances incorrectly. The KNN classifies 3452 (91.5164 %) instances correctly and 320 (8.4836 %) instances incorrectly.

In this comparison, the C4.5 algorithm has the highest accuracy than the other two algorithms. This dataset has the highest accuracy rage 99.5758 % than the breast cancer and iris datasets. It also shows that when the number of instances in the dataset increases, it is possible to produce the more accurate results.

When compare with other two algorithms the KNN is faster in terms of execution. Here the execution time is calculated using seconds, the execution rate 0.00 denotes that the time taken for process is less than milliseconds for KNN.

The below Table.3 shows the experimental results of correctly and incorrectly classified instances and the execution time for Naive Bayes, C4.5 and K-Nearest Neighbor using the hypothyroid dataset.

Comparison between Classification Algorithms Using Hypothyroid Dataset with 3772 Instances

Table.3.Represents the correctly and incorrectly classified datasets using Hypothyroid Dataset

Test Algorithm	Correctly Classified	Incorrectly Classified	Run time (seconds)
Naive Bayes	3594(95.281 %)	178(4.719 %)	0.03
C4.5	3756(99.57 %)	16(0.4242 %)	0.08
KNN	3452(91.516 4 %)	320(8.4836%)	0.00

The following Fig.3 represents the performance of classification algorithms based on the hypothyroid dataset.

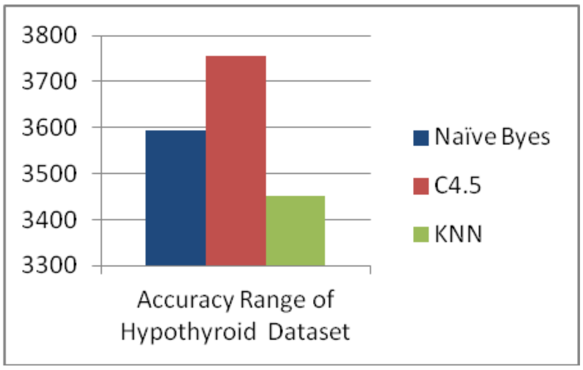

Fig.3 Bar chart representation of three algorithms accuracy using Hypothyroid Dataset

From this result it is concluded that the C4.5 algorithm has the highest accuracy than Naive Bayes and KNN for the hypothyroid dataset. But when compared to speed the C4.5 algorithm takes more time than the other two. And the K-Nearest Neighbor algorithm takes less amount of time with the accuracy level lesser than C4.5.

IV. CONCLUSION

From the predicted result, it is known that the accuracy of each algorithm is different with respect to the number of records in the datasets. The highest accuracy is obtained in the hypothyroid dataset because it is provided with many attributes and records, while comparing to the other two datasets. It is concluded that more the information we provide to the algorithm for processing the higher the accuracy we get. In terms of processing speed the KNN takes lesser time to process the datasets. The algorithm C4.5 stands out with highest accuracy range in classifying the medical datasets and it is the best algorithm in classifying the records in an efficient way than the Naïve Bayes and K-Nearest Neighbor algorithms. As for future work, more information about the datasets can be gathered using field survey and an improved C4.5 algorithm to handle multidimensional data.

REFERENCES

1. M. J. Berry, G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, New York: John Wiley & Sons, Inc, 1997.
2. D. T. Larose, Data Mining Methods and Models, Canada: A John Wiley & Sons, Inc, 2006.
3. A. Kumar, O. Singh, V. Rishiwal, R. K. Dwivedi, R. Kumar, “Association Rule Mining On Web Logs For Extracting Interesting Patterns Through Weka Tool,” International Journal of Advanced Technology In Engineering And Science, vol. 3, no. 1, pp. 134-140, 2015.
4. C. D., Discovering Knowledge in Data: An Introduction to Data Mining, Canada: John Wiley & Sons, 2014.
5. George Dimitoglou, JamesA. Adams, and Carol M. Jim, “Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability”, Journal of Computing, Volume 4, Issue 8, 2012.
6. Leni Marlina, “Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)”, Inte rnational Journal of Engineering Trends and Technology (IJETT) – Volume 38 Number 7- August 2016.
7. S.J.Russel, and Norvig, "Artificial Intelligence: A modern approach (International edition), Pearson US imports & PHIPES, Nov 2002.

Retrieval Number: C8795019320/2020©BEIESP
DOI: 10.35940/ijitee.C8795.019320

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

8. P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss", Machine Learning, 29:103–130, 1997.
9. H. I. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, (2005).
10. M. J. Berry and G. S. Linoff, "Data mining techniques: For marketing, sales, and customer support", John Wiley & Sons, Inc., (1997).
11. J. R. Quinlan, "Decision trees and decision-making", IEEE Transactionson Systems, Man and Cybernetics, vol. 20, no. 2, (1990), pp.339-346. Cover, T.M. (1968) "Rates of convergence for nearest neighbor procedures", In Proceedings of the Hawaii International Conference on System Sciences, Univ. Hawaii Press, Honolulu, 413–415.
12. Devroye, L. (1981) "On the equality of Cover and Hart in nearest neighbor discrimination", IEEE Trans. Pattern Anal. Mach. lntell. 3: 75-78.
13. Dunham, M. H. (2003): Data mining: Introductory and advanced topics.Pearson Education, 94-96.

AUTHORS PROFILE

Dr.K.Pazhanikumar has been working as an Assistant Professor of the Department of Computer Science and Research in S.T. Hindu College, Affiliated to Manonmaniam Sundaranar University for the last 23 years. He has completed his M.C.A in Madurai Kamaraj University, M.Phil Computer Science degree in Manonmaniam Sundaranar University, Tirunelveli and Ph.D (Computer Applications) in Manonmaniam Sundaranar University, Tirunelveli.

R. Raja Aswathi has completed her B.Sc (Computer Science) and M.C.A(Computer Applications) in Manonmaniam Sundaranar University, Tirunelveli.
.