ISSN 0975-2366
https://doi.org/10.31838/ijpr/2018.10.04.051

Research Article

Performance Analysis of Different Classification Algorithms used for Cancer Datasets

[1]SELVAM .S, [2]DR. P. MAYILVAHANAN
[1]Assistant Professor , Department of Computer Science, Govt Arts College, Dharmapuri-5, India1
email id : selvamsivan1981@gmail.com
[2]Professor, Department of Computer Application, Vels University [VISTAS], Chennai-117, India2
Email id : mayilkadir@yahoo.com
Received: 25.07.18, Revised: 25.08.18, Accepted: 25.09.18

ABSTRACT
Classification is a major technique in data mining and most used in different fields. Classification is a data mining function which assigns object in a collection to target categories or classes. Classification models predict categorical class labels. The aim of classification is to accurately predict the target class for each problem in the data. This paper collects different cancer dataset and three different classification algorithms applied to generate the accuracy of the algorithm and find the best algorithm..

Keywords: Classification, Random forest algorithm, K-NN, Naive Bayes.

INTRODUCTION

Data Mining is an interdisciplinary field merging ideas from statistics, machine learning, information science, visualization and other disciplines. It is a very useful approach to integrate information and theory for knowledge discovery from any informatics such business, medicine, commerce and Materials informatics and so on [2]. The impact of Data Mining and knowledge discovery has been evidenced by many successful research experimental results. Therefore, Data mining can be used to extract non-trivial, hidden, previously unknown, potential useful and ultimately understandable knowledge from massive cancer databases. Data Mining has two primary Models: Descriptive Model and predictive Model. Descriptive models describe or summarize the general characteristics or behavior of the data in the cancer database. Predictive models perform inference on the current data in order to make the prediction. Both of them are fundamentals to understand cancer behaviors. In general, in cancer informatics, Data mining can be used in the following task.

Classifier Algorithm

Some machine learning algorithms can be used for cancer class prediction, and objects classification models such as Support Vector Method (SVM) and Artificial Neural Network (ANN), Random Forest (RF), can be used to build up the Predict models.

Cluster Algorithm

As an exploratory data analysis tool, it can sort different objects or properties into groups in such a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. And, cluster analysis can be integrated with high-throughput experimentation for rapidly screening combinatorial data.

Association Mining

Association Mining is good at discovering patterns, and can be used to develop heuristic rules for object behavior based on large data sets [2].

Outlier Analysis

An outlier analysis is used to identify anomalies, especially to assess the uncertainty and accuracy of results, and distinguish between true discoveries and false-positive results.

Visualization

Reconstruction of material structure information based on materials data would help researchers to analyze the relationships between object and object properties.

Classification

Classification is a supervised learning. The classification method is used to predict categorical class labels like discrete or nominal [3]. It uses labels of the training data to classify new data. There are two main steps in classification first step is we have to construct classification model based on training data and second is usage of model, we have to test its accuracy before using our model. In this research work the classification algorithm like Random forest ,K-NN, Naïve Bayes are used for classification [3]. This paper uses five different cancer dataset with three classification algorithm and popular classification tool called Rapid Miner 5.0 .Rapid Miner is a collection of machine learning algorithms for data mining tasks. Rapid Miner contains tools for data pre-processing, classification, Regression, clustering, Association rules, and visualization. Random forest algorithm, K-NN and Naïve Baye is used for the classification using Rapid Miner tool.

Methodology

The objective of this paper is to find out the performance of the three different classification algorithms and find the accuracy of algorithm which is working better for future and all problem solvable capability for different data sets [3]. Classification is the most commonly applied data mining technique. The proposed work of this paper is first fetching data's from the five different cancer dataset from UCI Repository. To apply the different instance and then classify it by classification algorithms like Random Forest, K-NN and Naïve Bayes algorithm to find out the accuracy of the algorithm and performance of the three classification algorithm.

Random Forest

Random forests are an ensemble learning method. It is used to solve classification, regression problems and also other problems. Random forest is one of the accurate learning algorithms. The basic concept of the algorithm is to build many small decision-trees and then merging them to form a forest [4]. It is computationally easy and cheap process to build many such small and weak decision trees. So such decision trees can be formed in parallel and then it can be combined to form a single and strong forest. The algorithm for random forests uses the common technique of bootstrap bagging. Given a training set S = {(x1, y1),……, (xn, yn)}, bagging repeatedly (B times) selects a random sample from the training set and construct trees to fit these samples. This procedure leads to better performance that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated.

Random Forest pseudo code

Randomly select **"k"** attributes from total **"m"** attributes. Where **k << m**

Among the **"k"** attribute, calculate the node **"d"** using the best split point.

Split the node into **daughter nodes** using the **best split**.

Repeat **1 to 3** steps until "l" number of nodes has been reached.

Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

The beginning of random forest algorithm starts with randomly selecting **"k"** attributes out of total **"m"** attributes In the image, you can observe that we are randomly taking attributes and observations For each tree in the forest, select a bootstrap sample S(i) from the training set S. Then construct a decision tree for the selected sample. The pseudo-code for constructing the decision works as follows: for each node of the tree, select a very small subset of attributes **k** from **m** where **m** is the complete set of features. It is computationally expensive process to decide which attributes to be selected for the decision tree learning. But by narrowing the set of attributes, the learning process becomes very fast. Then to classify a new object, apply the input vector to each tree in the forest and each tree will give a particular class as output. The forest will choose the class with the most votes.

K-Nearest Neighbor (KNN)

KNN classifier is an instance-based learning Algorithm which is based on a distance function for pairs of observations, such as the Euclidean distance or Cosine. In this paradigm, KNN of a training data is computed first. Then the similarities of one sample from testing data to the k nearest neighbors are aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class. KNN is an non parametric lazy learning algorithm. That is a pretty concise statement. When you say a technique is non parametric, it means that it does not make any assumptions on the underlying data distribution. This is pretty useful, as in the real world, most of the practical data does not obey the typical theoretical assumptions made

K-NN Algorithm Pseudo code

Calculate $Z(X, X_i)$ if i =1, 2, ….., **n**; where **Z** denotes the Euclidean distance between the points.

Arrange the calculated **n** Euclidean distances in non-decreasing order.

Let **K** be a positive integer, take the firstK distances from this sorted list.

Find those K-points corresponding to these K-distances.

Let K_i denotes the number of points belonging to the i^{th} class among K points i.e. $K \geq 0$

If $K_i > K_j \ \forall \ i \neq j$ then put X in class i. K-NN fundamentally works on the belief that the data is connected in a feature space. Hence, all the points are considered in order, to find out the distance among the data points [5]. Euclidian distance or Hamming distance is used according to the data type of data classes used. In this a single value of K is given which is used to find the total number of nearest neighbors that determine the class label for unknown sample. If the value of K=1, then it is called as nearest neighbor classification.

Naive Bayes

Naive Bayes algorithm is the simple Statistical Classifier. It is called Naive as it assumes that all variables contribute towards classification and are mutually correlated. This assumption is called class conditional independence. It is also called Idiot's Bayes, Simple Bayes, and Independence Bayes. They can predict class membership probabilities, such as the probability that a given data item belongs to a particular class label. A Naive Bayes algorithm considers that the presence (or absence) of a particular attribute of a class is unrelated to the presence (or absence) of any other feature when the class variable is given. The Naive Bayes technique is based on Bayesian Theorem and it is used when the

dimensionality of the inputs is high. Bayes algorithm is based on Bayes Theorem and Bayes Theorem is stated as below [6]: Let X is a data sample whose class label is not known and let H be some hypothesis, such that the data sample X may belong to a specified class C. Bayes theorem is used for calculating the posterior probability P(C|X), from P(C), P(X), and P(X|C). Where

$$P(C|X) = \frac{P(X|C)\,P(C)}{P(X)}$$

$$P(C|X) = P(X_1|c) \cdot P(X_2|C) \ldots P(X_n|C).P(C)$$

P(C|X) is the posterior probability of target class.
P(C) is called the prior probability of class.
P(X|C) is the likelihood which is the probability of predictor of given class.
P(X) is the prior probability of predictor of class.

The Naive Bayesian classifier is a type of probabilistic classifier. This method uses Bayes' theorem and also assumes that each and every features in a class are highly independent, that is the appearance of a feature in a particular category is not connected with to the presence of any other feature. The naive Bayes algorithm done based on the prior probability and likelihood of a tuple to a class.

The working of naive Bayes algorithm is given as follows:

The training data samples are partitioned based on class labels. Each data partition id denoted using class labels Ci where i=1,2,...,m and each class will have set of tuples represented as Xj where j=1,2,...,n.

After training process, if an unknown tuple X is given for classification then the classifier will find the posterior probability of Ci for give tuple X and assign X to class Ci if and only if posterior probability of Ci given X is greater than posterior probability of Cj given X where 1<=j>=m and j not equal to i. The posterior probability of Ci for give tuple X can be calculated as,

Posterior probability (Ci given X)=(Likelihood of tuple X with class Ci * Class prior probability Ci)/ Prior probability of tuple X.

The Naive Bayesian classifier works with both continuous and discrete attributes and works well for real time problems. This method is very fast and highly scalable. The drawback of this technique is when a data set which has strong dependency among the attribute is considered then this method gives poor performance

Materials And Methods

We have used the popular open-source data mining tool for Rapid Miner (version 5.0) to analysis five different data sets have been used and the performance of a comprehensive set of classification algorithms (classifiers) has been analyzed. The analysis has been performed on a Dell Windows system with Intel® Core ™ i3 CPU, 2.40 GHz Processor and 4.00 GB RAM. The data sets have been chosen such that they differ in size, mainly in terms of the number of attributes.

Table1. Composition of Cancer Data Sets

S.No	Data Sets Name	Total No of Instance	Total No of Attributes	Total No of Classes
1	Prostate Cancer	100	8	2
2	Cervical cancer	237	11	2
3	Breast Cancer General	699	9	2
4	Breast cancer Diagnosis	569	30	2
5	Breast Cancer Prognosis	198	33	2

Data set 1

The first data set is a small sized prostate Cancer data used in this study. The data set contains 8 regular attributes apart from the class attribute with 100 instances and 2 classes.

Data set 2

The second data set is a small sized cervical cancer data set with The data set has a total of 11 regular attributes and 237 instances and 2 class.

Data set 3

The third data set is a medium size breast Cancer general data used in this study. The data set contains 9 regular attributes apart from the class attribute with 699 instances and 2 classes.

Data set 4

The fourth data set is a medium sized breast cervical cancer diagnosis data set. The data set has a total of

30 regular attributes and 569 instances and 2 classes.

Data set 5

The fourth data set is a medium sized breast cervical cancer Prognosis data set. The data set has a total of 33 regular attributes and 198 instances and 2 classes.

Experimental Result

We tested the three aforementioned algorithms using for the same numbers of folds for cross validation dataset1 and holdout method for dataset2, in order to discover whether they have a great effect on the result. Finally, we set the algorithms with default parameters are used 10-foldcross validation and holdout method for 2/3 training data and 1/3 test data. We show the classification accuracy and rates obtained with the three algorithms for the five cancer dataset.

Table 4.1 Analysis of Different Classification algorithms for Prostate Cancer Dataset1

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	78	22
2.	K-NN	74	26
3.	Naïve Bayes	78	22

The Table 4.1 shows the different classifiers such as Naïve Bayes and Random Forest it has correctly classified 78 % on data. KNN it has also correctly classified 74 % on data classified. When compare to these algorithms Navie Bayes , Random Forest so better than KNN algorithm on Prostate Cancer classification.

Table 4.2 Analysis of Different Classification algorithms for Cervical Cancer Dataset1

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	83.55	16.45
2.	K-NN	71.65	28.35
3.	Naïve Bayes	100	0

The Table 4.2 shows the different classifiers such as Naïve Bayes has correctly classified 100 % on data. It has also classified 83.55 % data on Random Forest algorithms and 71.75 on KNN Algorithm . When compare to these algorithms for Navie Bayes is so better performance than KNN ,Random Forest is Classified on Cervical Cancer Dataset.

Table 4.3 Analysis of Different Classification algorithms for Breast Cancer General Dataset1

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	95.57	4.43
2.	K-NN	95.56	4.44
3.	Naïve Bayes	95.85	4.15

The Table 4.3 shows the different classifiers such as Naïve Bayes has correctly classified 95.85 % on data. It has also classified 95.57 % data on Random Forest algorithms and 95.56 % on KNN Algorithm . When compare to these algorithms for Navie Bayes is so better performance than KNN, Random Forest is Classified on Breast Cancer General Dataset.

Table 4.4 Analysis of Different Classification algorithms for Breast Cancer Diagnosis Dataset1

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	92.61	7.39
2.	K-NN	91.22	8.78
3.	Naïve Bayes	93.15	6.85

The Table 4.4 shows the different classifiers such as Naïve Bayes has correctly classified 93.15 % on data. It has also classified 92.61 % data on Random Forest algorithms and 91.22 % on KNN Algorithm . When compare to these algorithms for Navie Bayes is so better performance than KNN ,Random Forest is Classified on Breast Cancer Diagnosis Dataset.

Table 4.5 Analysis of Different Classification algorithms for Breast Cancer Prognosis Dataset1

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	76.29	23.71
2.	K-NN	67.13	32.87
3.	Naïve Bayes	67.24	32.76

The Table 4.4 shows the different classifiers such as Random Forest has correctly classified 76.29 % on data. It has also classified 67.24 % data on Naive Bayes algorithms and 67.13 % on KNN Algorithm . When compare to these algorithms for Random Forest is so better performance than KNN ,Naive Bayes is Classified on Breast Cancer Prognosis Dataset.

Table 4.6 Analysis of Different Classification algorithms for Prostate Cancer Dataset2

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	57.14	42.86
2.	K-NN	62.86	37.14
3.	Naïve Bayes	85.71	14.29

The Table 4.6. shows the different classifiers such as Naive Bayes has correctly classified 85.71 % on data. It has also classified 62.86 % data on KNN algorithms and 57.14 % on Random Forest Algorithm . When compare to these algorithms for Naive Bayes is so better performance than KNN ,Random Forests is Classified on Prostate Cancer Dataset

Table 4.7 Analysis of Different Classification algorithms for Cervical Cancer Dataset2

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	82.5	17.5
2.	K-NN	65	35
3.	Naïve Bayes	97.5	2.5

The Table 4.7. shows the different classifiers such as Naive Bayes has correctly classified 97.5 % on data. It has also classified 65 % data on KNN algorithms and 82.5 % on Random Forest Algorithm . When compare to these algorithms for Naive Bayes is so better performance than KNN ,Random Forests is Classified on Cervical Cancer Dataset

Table 4.8 Analysis of Different Classification algorithms for Breast Cancer General Dataset2

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	98.71	1.29
2.	K-NN	98.28	1.72
3.	Naïve Bayes	98.28	1.72

The Table 4.8. shows the different classifiers such as Random Forest has correctly classified 98.71 % on data. It has also classified 98.28 % data on KNN algorithms and 98.28.5 % on Naive Bayes Algorithm . When compare to these algorithms for Random Forest is so better performance than KNN , Naive Bayes is Classified on Breast Cancer General Dataset

Table 4.9 Analysis of Different Classification algorithms for Breast Cancer Diagnosis Dataset2

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	93.12	6.88
2.	K-NN	92.06	7.94
3.	Naïve Bayes	95.24	4.76

The Table 4.9 . shows the different classifiers such as Naive Bayes has correctly classified 95.24 % on data. It has also classified 92.06 % data on KNN algorithms and 93.12 % on Random Forest Algorithm . When compare to these algorithms for Naive Bayes is so better performance than KNN

,Random Forests is Classified on Breast Cancer Diagnosis Dataset

Table 4.10 Analysis of Different Classification algorithms for Breast Cancer Prognosis Dataset2

S.No	Algorithm	Correctly Classified	Error on Classified
1	Random Forest	83.33	16.67
2.	K-NN	62.61	37.39
3.	Naïve Bayes	60.61	39.39

The Table 4.10. shows the different classifiers such as Random Forest has correctly classified 83.33 % on data. It has also classified 62.61 % data on KNN algorithms and 60.61 % on Naive Bayes Algorithm . When compare to these algorithms for Random Forest is so better performance than KNN , Naive Bayes is Classified on Breast Cancer Prognosis Dataset

Result

The results from the above 10 tables have been analyzed manually and they indicate that the classifiers work better when there is an increase in the number of attributes in the data set. But, none of the classifiers outperformed the others in terms of the accuracies. The classifiers Random Forest, KNN and Naïve Bayes have performed better on all 5 data sets. However, the performance of KNN should not be considered because of less accuracy for five data sets by the classifier to generate results. The Naïve Bayes algorithm around 95.85.% accuracy for the large data set whereas the classifier. The remaining classifiers Random Forest performed better on large data sets which are expected.

Conclusion

This study focuses on finding the right algorithm for classification of data that works better on diverse data sets. However, it is observed that the accuracies of the tools vary depending on the data set used. It should also be noted that classifiers of a particular group also did not perform with similar accuracies. Overall, the results indicate that the performance of a classifier depends on the data set, especially on the number of attributes used in the data set and one should not rely completely on a particular algorithm for their study. So, we recommend that users should try their data set on a set of classifiers and choose the best one.

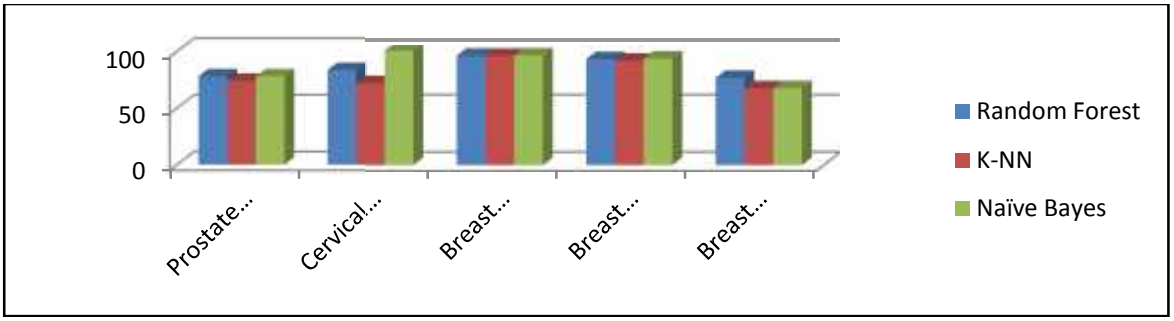

Figure 1: Cross Validation for different Classification Algorithm and different cancer datasets

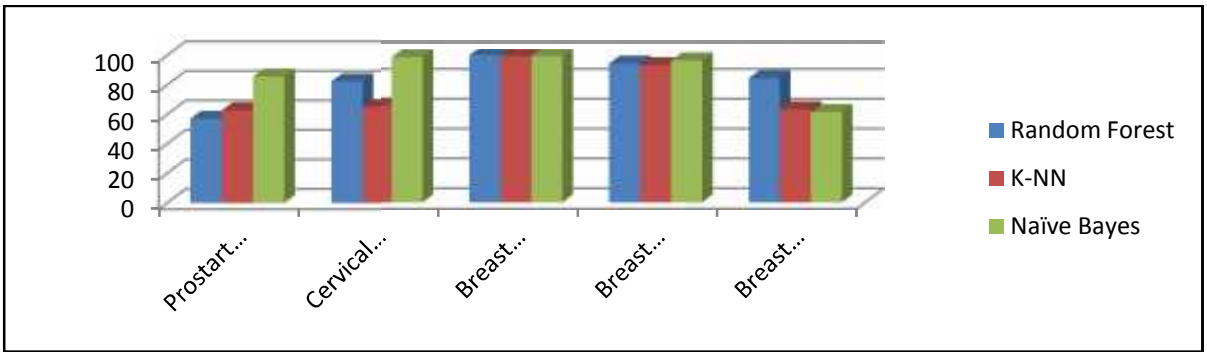

Figure 2: Holdout Validation for different Classification Algorithm and different cancer datasets

References

1. Raj Kumar and Dr. Rajesh Verma "Classification Algorithms for Data Mining:A Survey "International Journal of Innovations in Engineering and Technology (IJIET)
2. Doreswamy[1], Hemanth K S[2] "Hybrid data mininig techniques for Knowledge discovery from Engineering Materials sta sets"
3. Maria Nithi[1], Jeya Priya[2] "Performance Analysis of Different Classification Methods in Data Mining "International Conference on Advancements in Computing Technologies–ICACT
4. S.Ponmani, Roxanna Samuel, P.Vidhu Priya "Classification Algorithms in Data Mining – A Survey" *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 6, Issue 1, January 2017*
5. Sayali D. Jadhav [1], H. P. Channe[2]" Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064
6. Sagar S. Nikam "A Comparative Study of Classification Techniques in Data Mining Algorithms" Oriendal journal of computer science and technology ISSN: 0974-6471 April 2015 Vol 8 no(1)
7. Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy: "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press 1996
8. Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo Hunt EB, Marin J, Stone PJ (1966) Experiments in induction. Academic Press, New York
9. B. Kotsiantis · I. D. Zaharakis · P. E. Pintelas, "Machine learning: a review of classification and combining techniques", Springer Science10 November 2007.
10. Raj Kumar, Dr. Rajesh Verma," Classification Algorithms for Data Mining P: A Survey" IJIET Vol. 1 Issue August 2012, ISSN: 2319 – 1058.
11. Ms. Aparna Raj, Mrs. Bincy, Mrs. T.Mathu "Survey on Common Data Mining Classification Techniques", International Journal of Wisdom Based Computing,.[13] http://www.tutorialspoint.com/data_mining/dm_rbc.htm
12. Salvatore Ruggieri "EÆcient C4.5"[15] Ping Gu and Qi Zhou "Students performance prediction based on Improved C4.5 decision tree algorithm",
13. Rupali Bhardwaj , Sonia Vatta "Implementation of ID3 Algorithm", Volume 3 Issue 6, June 2013, ijarcsse.
14. Jiawei Han and Micheline Kamber (2001), *Data Mining: Concepts and Techniques: Book (Illustrated)*, Morgan Kaufmann Publishers.
15. J.R. Quinlan (2003), Induction in decision trees, *Journal of Machine Learning,* Vol.1,Issue1,pp 8– 106.
16. T.S. Lim, W.Y.Loh, Y.S.Shih (2000), A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, *Journal of Machine Learning,*Vol 40, pp203–228.