International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Comparison of Classifiers Based on Case Studies

Simran D. Makhija[1]

[1]PESIT- Bangalore South Campus, Bangalore, Karnataka, India

Abstract: ***Machine Learning is a growing field which has its roots in statistics, probability, artificial intelligence and algorithms. Some business applications, such as stock market predictions, and academic applications, such as classifying objects in the sky, deal with huge volumes of data and machine learning techniques are being successfully applied to simplify the scientists' tasks of finding useful patterns. There is no classifier which can assure us the best classification results in all scenarios. The best choice of classifiers and parameters is the subject of this paper. The current work tabulates results of three classifiers: Naive Bayes, Support Vector Machines and Random Forests on different data sets and shows the contrast in the accuracy of each of them supported with reasons as to why one classifier works better than another classifier for a specific data set. This study would assist in choosing the best technique to improve results established on classification and simultaneously give examples to explain different classification techniques in depth.***

Keywords: Machine Learning, Classifiers, Naive Bayes, Support Vector Machines, Random Forests

1. Introduction

Machine Learning is a field that is concerned with the development and understanding of systems that can learn from data. It is inspired by classical approaches in Pattern Recognition and Artificial Intelligence but relies more on statistical techniques. When exposed to new data, computer programs are enabled to learn, grow, change, and develop by them. There has been an enormous growth in terms of the size and dimensionality of the data in the last few decades, making the task challenging. Large volumes of data pose significant challenges to inferring information and patterns; ML methods can be used to find underlying interesting patterns in large volumes of data so that tasks involving predictive analysis can be done in a structured manner.

A classifier performs the task of categorizing input data to a category using a mathematical function. The data set is split in an appropriate ratio into two sets, where one is used to build the model and other to validate it.

Different types of classification methods work differently on different data sets. Some techniques give a better accuracy for a data set with numeric attributes and some give a better result with nominal attributes. Evaluating the performance of a machine learning method is important to identify the strength and weakness of each classification algorithm. The choice of the classifier depends on the problem, the correlation and weight of features, nature of data set, distribution of data, etc.

An analysis of the classification results was done in order to differentiate the models from each other and obtain an idea on which of the three models perform better in different cases. Classification techniques can be compared taking into consideration the mean accuracy, speed or running time and scalability.

In this paper, the accuracy, deviation and speed of each classifier is computed and compared using 10 fold cross validation. These results are reliable because they're evaluated for each fold separately.

Classification may be used in identifying a planet as habitable or not, classifying waste as dry or wet, classification of stages or extent of occurrence of a particular disease, etc.

The structure of the paper is as follow: in section 2, there is a description of the classification methods used; in section 3, the discussion is based on the implementations, metrics to evaluate an algorithm, comparison, methods used to predict accuracy, evaluation of quality of output of classifier and the results on comparison of the models; and section 4 comprises of the conclusion.

2. 2. Description of classification methods

The models used for classification which has been implemented and going to be discussed in the paper are:

1. Gaussian Naive Bayes
2. Support Vector Machines (SVM)
3. Random Forests (RFC)

2.1 Naive Bayes:

Naive Bayes (NB) is a simple technique which involves a family of classifiers following a common principle: given the category label, the value of a particular attribute doesn't depend on the value of any other attribute.

Bayes' theorem:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

By definition:

i. $$P(A/B) = \frac{P(A \wedge B)}{P(B)}$$

ii. $$P(B/A) = \frac{P(A \wedge B)}{P(A)}$$

Dividing the two equations, we obtain Bayes' theorem.

Each term in Bayes' theorem has a conventional name:

- P(A) is the prior probability of A
- P(A|B) is the conditional/posterior probability of A, given B.
- P(B|A) is the conditional probability of B given A.
- P(B) acts as the normalizing constant.
- L(A|B) is the likelihood of A given B.

Thus,

Posterior= Likelihood * Prior / Normalizing constant

For example, a person is considered to be a male if he weighs 70kg, a height of 6 feet and foot size 10. These features don't depend on each other while stating the person to be a man.
The theorem is combined with a decision rule: Choose the hypothesis that is most probable to occur.

X being an article to be classified into a labeled group, then the theorem can be seen as giving the likelihood of belongingness to one of the groups C_1, C_2, C_3, etc by calculating $P(C_i/X)$. We assign X to the group with the highest conditional probability. We have:

$$P(Ci/X) = \frac{P(X/Ci)P(X)}{P(Ci)}$$

2.2 Support Vector Machines:

Support Vector Machines (SVM) is a kernel method based on analysis of a pattern which studies and finds relations and similarity in data sets. Using a user-specified feature map, the unprocessed data can be converted into vector form. Feature vectors and weights of the feature vectors are combined using a scalar product to model a predictor function to determine a score for making a prediction. On the other hand, kernel methods require only a user-specified kernel: a similarity measure over pairs of data points in raw format. Each prediction scans the entire data set rendering the computation time intensive for large data sets.

SVM is explicitly told to find the best separating line. One efficient approach to construct the plane as far as possible from both sets is to make the smallest convex sets that group all the data in each class (i.e. the convex hull) and find the nearest points in them. A line is then drawn connecting them by performing vector subtraction. It then declares the perpendicular bisector of the connecting line to be the best separating plane. Its focus is only on the points that are the most difficult to tell apart. [1]

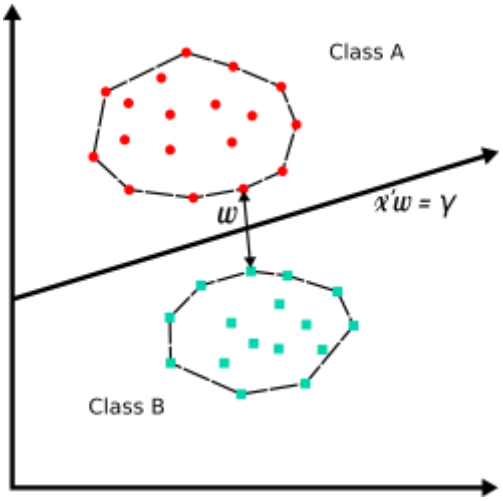

Figure 1: Convex hull method

x'w = 0 is the equation of the hyper plane, which is the dot product of the vectors.

2.3 Random Forests:

A decision tree is a structure where every node in it is a condition on a single attribute. The data set is branched into two so that similar outcome values end up in the same set, where the leaf node holds the class label. These tree predictors are collectively assembled to obtain the random forest classifier. The correlation between individual trees contributes to the strength of the classifier It proves that a group of "weak learners" can form a "strong learner". Random forests are stable as a slight change in the input data may affect individual trees, but the characteristics of the forest remain unchanged.

To classify a previously unobserved sample, the vector is traced down each tree in the forest. Each tree independently classifies the sample or votes for a certain class that it should belong to. The forest chooses the classification based on majority voting, as depicted in Fig. 2.

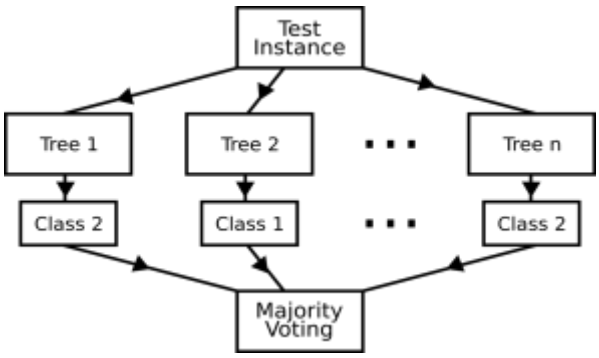

Figure 2: Majority voting

Some approaches use probabilistic prediction method instead of the majority vote for the model aggregation. The average of the predicted class probabilities of individual tree predictors in the forest is calculated to obtain predicted class probabilities of a test sample.

Random forest applies Bootstrap Aggregating (Bagging) to improve accuracy and make it more stable using a model

averaging approach. It is done by picking a random sample, constructing a tree to fit it, and repeating the procedure.

Impurity is the value based on which the most suitable split condition is chosen. For classification, either Gini impurity (a measure of misclassification) or information gain impurity is used. The impurity decrease from each feature can be averaged and the features are prioritized according to this measure. It measures the weight of each feature the model.

3.3. Results and Discussion

3.1 Performance Measures for Classification:

3.1.1 Confusion Matrix:

To obtain the quality of performance of a classification method, the confusion matrix is computed. The number of instances for which the class label is correctly predicted is represented by the main diagonal, while the number of samples for which the class label is incorrectly predicted is represented by the non-diagonal elements.

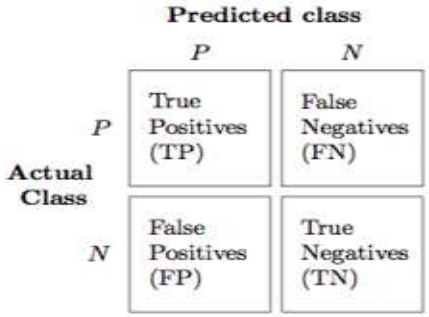

Figure 3: Confusion matrix format (Source: International Journal of Computer Applications Volume 55 -No.6, October 2012)

1. **True Positive (TP)**: cases when it's predicted to be true, and it's actually true.
2. **False Positive (FP)**: cases when it's predicted to be true, but it's actually false.
3. **True Negative (TN)**: cases when it's predicted to be false, and it's actually false.
4. **False Negative (FN)**: cases when it's predicted to be false, but it's actually true.
5. **Sensitivity**: Measure of how often the classifier predicts true when it's actually true. It is calculated by TP/(TP+FN).
6. **Specificity**: Measure of how often the classifier predicts false when it's actually false. It is calculated by TN/(TN+FP).
7. **Misclassification Rate**: An overall measure of how often the prediction is wrong. It is calculated by (FP+FN)/Total.
8. **Accuracy**: An overall measure of how often the prediction is correct. It is calculated by (TP+TN)/Total. [2]

3.2 Methods for estimating the accuracy of a method:

Holdout Method: It requires a test and training set, which are mutually exclusive. A larger training set would produce a better classifier, while a larger test set would provide a better estimate of the accuracy. A balance needs to be drawn and maintained by choosing an appropriate ratio to divide the sets. The sets should be disjoint to prevent the estimate from being biased and their union should comprise of the universal set so that the whole population is represented.

Random sampling: Process of repeating the holdout method several times and computing the mean of the accuracy of all the trials. This produces better error estimates at is done with a different combination of sets every repetition.

K-fold Cross-Validation Method: Cross-Validation (CV) test splits the training samples into many partitions. One of them is kept aside to test the model, while the remaining builds it. In K-fold CV method, k-1 folds are used to train the model and it's tested on the one not considered for training. This process is repeated with each fold considered for testing exactly once, and mean of them is obtained. This method determines the best estimate as all partitions have been used for evaluation.

3.3. Comparison of Results on Using CV Method for Default Parameters of Classifiers:

Case 1: Fertility Data Set on UCI Machine Learning Repository:

Based on sperm concentration, which in turn in depends on environmental and lifestyle factors, fertility is predicted. It has 100 samples with 9 features each. [3]

Information about the features used:

1. Season
2. Age
3. Diseases as a kid
4. Major trauma
5. Surgery
6. Fever in the last year
7. Alcohol
8. Smoking
9. Time spent sitting
10. Output: Diagnosis normal (N), altered (O)

Mean accuracy and standard deviation of the accuracy of classifiers on this data set:

NB: 0.830000 (0.161555)
SVM: 0.880000 (0.116619)
RFC: 0.860000 (0.101980)

Results prove SVM works best for this data set in comparison to the other two classifiers.

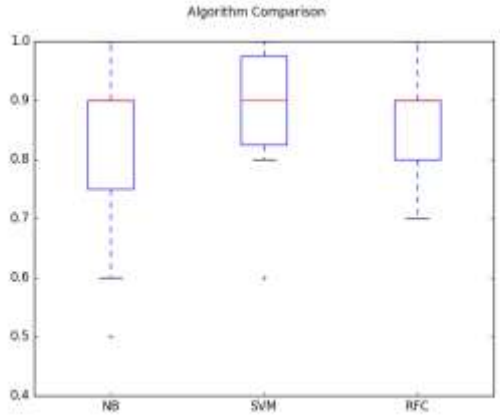

Figure 4: The range of the accuracy scores across each fold for each algorithm on Fertility data set.

Case 2: Pima Indians Diabetes data set from UCI Machine Learning Repository:

The diagnostic variable obtained is to determine if the patient has diabetes. It has 768 samples and 9 features. [3]

Information about the features used:

1. Pregnancy rate
2. Concentration of plasma glucose
3. BP
4. Triceps skin fold thickness
5. Serum insulin
6. BMI
7. Pedigree function
8. Age
9. Category label

Mean accuracy and standard deviation of the accuracy of classifiers on this data set:
NB: 0.755178 (0.042766)
SVM: 0.641025 (0.0721)
RFC: 0.733879 (0.039762)

Results prove Naive Bayes is the best classifier for this data set in comparison to the other two classifiers.

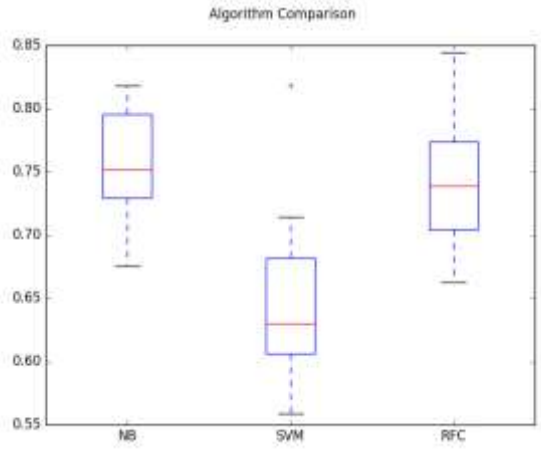

Figure 5: The range of the accuracy scores across each fold for each algorithm on Pima Indians Diabetes data set.

Case 3: Breast Cancer Wisconsin (Diagnostic) Data Set from UCI Machine Learning Repository:

The diagnostic, binary-valued variable investigated is to determine if the cancer is benign or malignant. It has 569 samples with 32 features each. [3]

Information about the features used:

1. Registration number
2. Diagnosis
For each cell nucleus:
a. radius
b. texture
c. perimeter
d. area
e. smoothness
f. compactness
g. concavity
h. point which are concave
i. symmetry
j. fractal dimension

Mean accuracy and standard deviation of the accuracy of classifiers on this data set:

NB: 0.838406 (0.114049)
SVM: 0.661159 (0.118503)
RFC: 0.954244 (0.043690)

Results prove RFC works best on this data set in comparison to the other two classifiers.

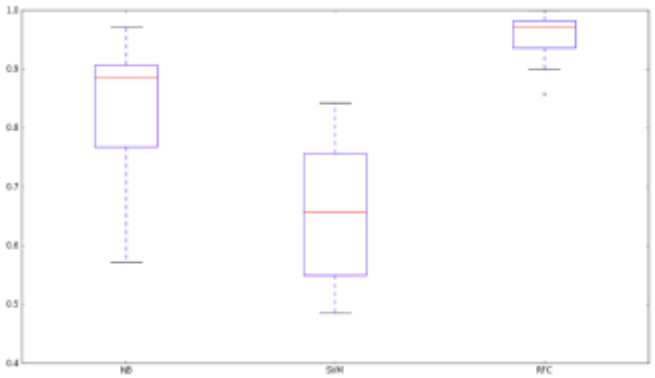

Figure 6: The range of the accuracy scores across each fold for each algorithm on Contraceptive choice method data set.

3.4 Using Random Sampling for Analysis

Contraceptive Method Choice Data Set from UCI Machine Learning repository:

The problem is contraceptive method duration prediction based on various characteristics. It has 1473 samples with 9 features each. [3]

Information about the features used:

1. Age
2. Education level of Wife
3. Education level of Husband

4. Number of children
5. Religion
6. Occupation of wife
7. Occupation of husband
8. Index of lifestyle
9. Amount of time spent on media
10. Contraceptive method used label

Mean accuracy of ten random samples from the data set:

NB: 47.15
SVM: 68.79
RFC: 94.00

Results prove RFC works best on this data set in comparison to the other two classifiers.

Results of classification tabulated:

Classifier	Accuracy	Sensitivity	Specificity
SVM	68.7908	0.9373	0.8744
RFC	94.0042	0.9940	0.9648
NB	47.1465	0.8340	0.5277

3.5 Differences in the Training and Testing Time on the Contraceptive Method Choice Data Set:

Classifier	Training time(in s)	Testing time(in s)
NB	0.0283989906311	0.00471186637878
SVM	6.84579801559	0.676080942154
RFC	0.144664049149	0.0148549079895

Results prove Naive Bayes' has the least training as well as testing time on this data set in comparison to the other two classifiers.

3.6 Understanding Classification Examining the Distribution of Data:

Data set plot has training data represented by bold colors and testing data with lighter colors shown in Fig. 7.

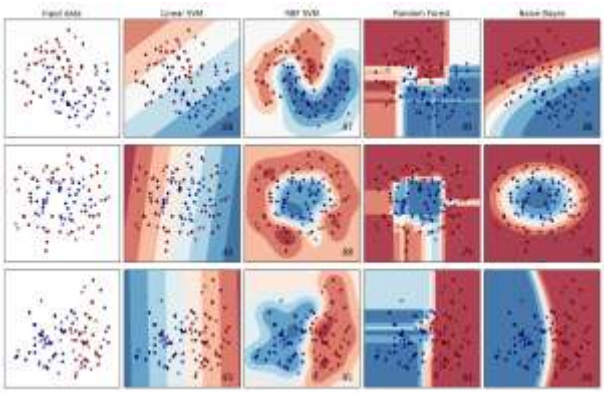

Figure 7: Distribution of entities

The intuition conveyed by this plot does not always apply to datasets in practical scenarios.

Data is more linearly separable in certain high-dimensional spaces. Thus classifiers like Linear SVM and Naive Bayes' might work better on general cases.

In order to obtain best results, one must examine, visualize and develop an intuitive idea about the data they're dealing with. Each feature must be understood individually and a correlation between them must be established. This can be done in several ways:

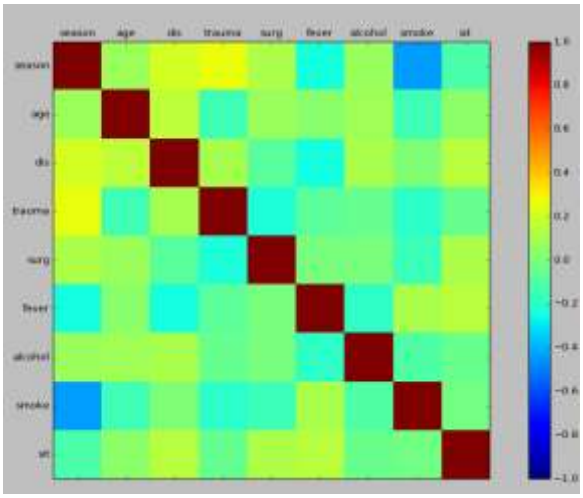

Figure 8: Correlation between features of Fertility dataset

Histograms: They give an intuitive idea of the distribution of each feature, by grouping them separately into columns.

Box-plots: They give an idea about the distribution of each feature, with the red line at the middle and box at the middle 50% of the data, to examine the dispersion of the data.

Correlation matrix: Correlation gives us an idea of how varying one feature variable influences another. Using correlation matrix, one can examine the correlation between each pair of features as shown in Fig. 8.

Scatter-plots: They show the connection between any two features as points in two dimensions as shown in Fig. 9.

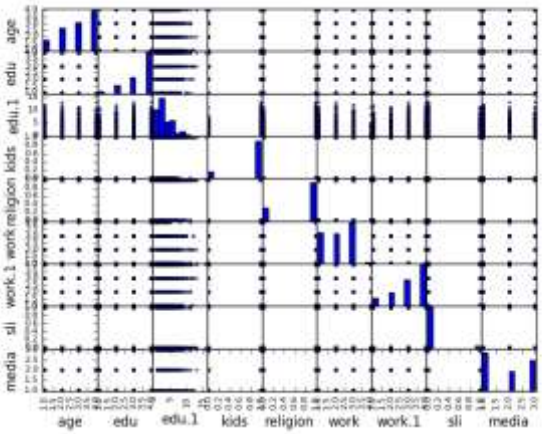

Figure 9: Scatter-plot matrix for Contraceptive Method choice

3.7 Comparing the Performance of Classifiers:

The size of the test set needs to be taken into consideration while comparing different classification models. If two models M_a and M_b work with an accuracy of 85% on 30 test records and 75% on 1000 test records, M_a can't be considered as a better model than M_b due to the vast difference in sizes of the test set. Thus, we must gauge the confidence interval of a given classifier performance.

Our aim is to find out if the errors are negligible statistically. Taking into consideration two classifier models M_1 and M_2 to test and train datasets D_1 and D_2, having n_1 and n_2 samples, with the error rates e1 and e_2 for M_1 on D_1 and M2 on D_2 respectively. The observed difference in error rate is denoted as $d = e_1-e_2$, then d is also normally distributed with mean dt and variance, σ^2_d. The variance of d can be computed as follows:

$$\sigma^2_d = \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}$$

Where each of the two terms on the right-hand side are the variances of error rates. [5]

4.4. Conclusion

Machine Learning methods are case-based learning. The rank and performance measures of classifiers are different for distinct data sets, and hence classification techniques need to be chosen effectively. Using data sets from UCI Machine Learning repository, tests have been performed and results with three different classification models for their default parameters have been displayed. The performance metrics, which give a detailed analysis of the behavior of the model, have been discussed in detail to highlight their importance. The performance can be improved by altering the configurations of parameters.

The conclusion obtained from these experiments proves that the behavior of a model depends on the nature, correlation, and distribution of the training feature samples. Therefore, choice of the best classification method can be done taking these factors into consideration. Other factors that can influence this decision are number of samples, the number of features, etc. In cases where training samples are limited, it is important to evaluate the working and learning of these models.

References

[1] Kristin P. Bennett and Erin J. Bredensteiner, "Duality and geometry in SVM Classifiers", in Proc. 17th International Conf. On Machine Learning, 2000.
[2] D.L. Gupta, A.K. Malviya and Satyendra Singh, "Performance analysis of classification tree learning algorithms", International Journal of Computer Applications (0975-8897) Volume 55 -No.6, October 2012.
[3] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: The University of California, School of Information and Computer Science.
[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[5] G.K. Gupta (2004), Introduction to Data mining with Case Studies.

Author Profile

Simran D. Makhija is currently pursuing Information Science and Engineering at PES Institute of Technology- Bangalore South Campus.