# Test Essential Assumption Knowledge on Naïve-Bayes Tree

Naveen Kumar Lekkalapudi
West Virginia University
nalekkalapudi@mix.wvu.edu

## ABSTRACT

A simple Naïve-Bayes (NB) classifier competes with several sophisticated and complex predictors. If an NB-classifier is given data that is similar to the data to be predicted, the precision and accuracy can be improved considerably. In this paper, a Naïve-Bayes tree is used to improve the efficiency of the classifier by testing assumed knowledge that similar data will lead to better accuracy, precision and recall.

## General Terms

Algorithms, Management, Measurement, Performance, Design, Reliability.

*Results*:

## Keywords

NB-tree, NB, TEAK, precision, recall, accuracy.

## 1. INTRODUCTION

Naïve-Bayes algorithms were shown to be surprisingly accurate on many classification tasks. The accuracy of a Naïve Bayes algorithm scale up well when NB-Trees are used [2]. Essential assumptions can be made which are based on analogy. Similar data always yields same class results. This assumption can be tested on the NB-Tree to calculate accuracy, precision and recall and compare it with just NB classifier to compare the performance of both methods. It is a known fact that one classifier cannot run effectively on all data sets and the performance values depends on data that classifier is tested on. Several data sets are tested on the method and results are documented as follows.

## 2. RELATED WORK

There is currently heavy research going on, bits that are used in preparing this paper are listed below:

### 2.1 TEAK

Menzies, et al. 2011 [1] have worked extensively on testing essential assumptions on analogy based effort estimation. The immediate neighbors of test data offer stable conclusions about that dataset [1]. The assumption is tested by generating a binary tree of clusters of effort data and comparing the variance of super-trees vs smaller sub-trees. Several conclusions are made from the tests and they are ought to be satisfactory. It was concluded that the estimation based on analogy could significantly improve prediction by dynamic selection of nearest neighbors, using only data from regions of small variance.

### 2.2 NB-Tree

Koavi, 2011. [2] has scaled up the accuracy of NB classifiers using a decision tree hybrid. The algorithm used is similar to recursive partitioning schemes, with the leaf nodes as the Naïve-Bayes categorizers predicting a single class instead of nodes. The accuracy is determined to be increased when NB-Tree classifiers are used.

### 2.3 Others

Friedman et all. 1996. Showed learning from a tree augmented Nb which is essentially an tree restricted Bayes classifiers. The approach used is said to be restrictive[2].

## 3. ANALYSIS

In order to build a recursive NB-Tree the data is initially projected onto a 2-dimensional plane using PCA.

### 3.1 Distance Measure

Orthogonally transforming the given datasets require relative distances or differences between each row of data in a dataset. The following successful mechanism is used in the algorithm for calculating orthogonal distances between rows:

1. Numerics are normalized (so all numerics have equal influence on the distance);
    a. Standard normalization: convert 0..1 as follows;
    b. x = (x - min)/(max - min)
2. Distance between numeric variables = (x-y)^2
3. Distance between non-numerics = 0 if x==y otherwise 1.
4. Missing attribute values are assumed to be maximally different from the value present.
5. If both values are missing, then distance = 1.

### 3.2 XY Projection

To project the data onto the xy-plane, the following algorithm is used which is bases that the cosine value rests on the x-axis and the Pythagoras value rests on the y-axis.

```
row = any(data)
east = furthest(row)
west = furthest(east)
c = dist(east,west)
for each row r in data:
        a = dist(r,east)
        b = dist(r,west)
        x = (a^2 + c^2 - b^2)/2c
        y = (a^2 + x^2)^0.5
```

By obtaining the xy values of each row, the data is projected on the xy-plane. A recursive partitioning algorithm is used to generate leaves of tree. These leaves are the smaller sets of data which are similar to each other.

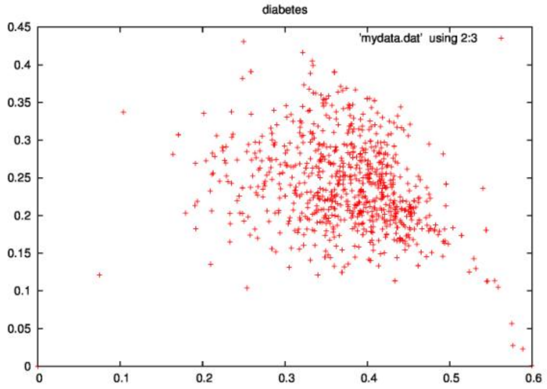

**Figure 1: Diabetes dataset on xy plane**

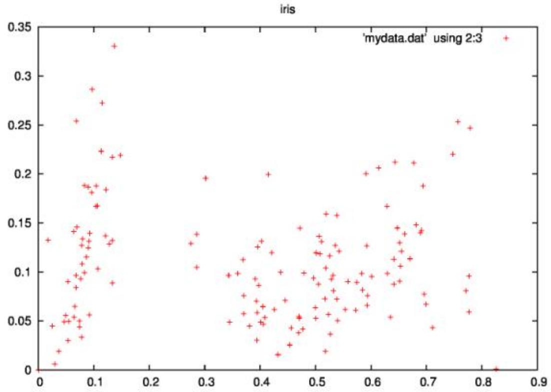

**Figure 2: Iris dataset on xy plane**

In the testing phase, the test row is projected on to the trained xy-plane and the Euclidean distances are calculated for each leaf from the xy dimensions of the test row. The leaf with least distance is considered to be the closest and NB algorithm is applied on the obtained data.

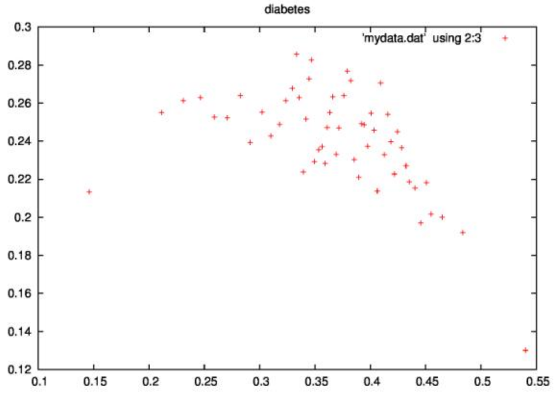

**Figure 3: Leafs of diabetes dataset**

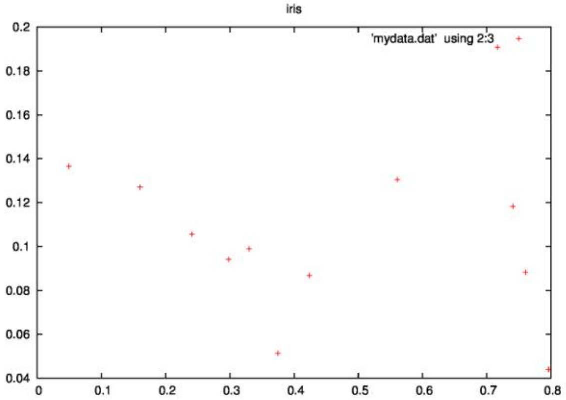

**Figure 4: Leaves of Iris dataset**

## 3.3 NB Classifier

To classify the data, few assumptions are made. The likelihood calculates the likelihood of a row belonging to a class. The numerical values in the dataset follow normalization to scale between the values of 0..1. The count of terms in the dataset is divided by total number of rows in dataset to obtain their normalized value. Logarithmic additions of these variables are done to calculate the likelihood of each row.

## 3.4 Performance Estimators

### *3.4.1 Accuracy*

Accuracy is defined as how close the predicted value is to the actual value. It can be calculated as the factor of correctly predicted values to the total number of predictions.

*Accuracy = Correct Predictions / Total Predictions.*

### *3.4.2 Precision*

Precision is the fraction of retrieved instances that are relevant. It is calculated as a fraction of true positives to total number of positive predictions.

*Precision = Tp / (Tp + Fp)*

### *3.4.3 Recall*

Recall is the fraction of relevant instances that are retrieved. It is calculated as a fraction of predicts to number of instances to be predicted.

*Recall = Fp / (Tp + Fn)*

There is a problem with accuracy when low frequency class values occur in a dataset. In order to effectively estimate the performance of the predictor, precision and recall are used.

# 4. RESULTS

The graphs below display the results of TEAK method applied on a Naïve-Bayes Tree. The accuracy depends on the type of dataset. If the data is more diverse then the predictor’s performance seems to be deteriorating. However accuracy of a NB can be successfully improved by use of TEAK with NBTree.

The following table lists the datasets that the hypothesis was tested on:

| Dataset | Instances | Size |
|---|---|---|
| Weather | 14 | 0.4k |
| Soybean | 682 | 1M |
| Iris | 149 | 2k |
| Image | 210 | 3k |
| Wine | 178 | 10k |
| Cmc | 1473 | 3k |
| Cylband | 550 | 10k |
| Diabetes | 768 | 2k |
| Covtype | 2757 | 75M |

**Figure 5: Datasets**

The following are graphs of accuracy generated:

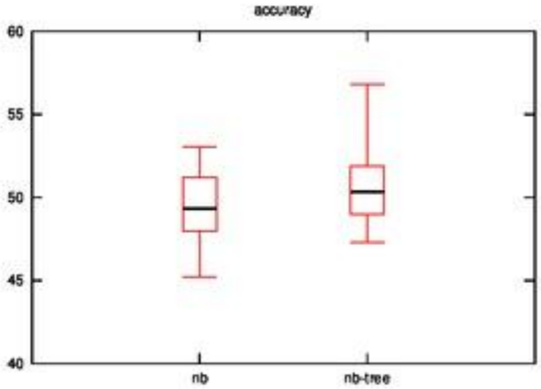

**Figure 6: Accuracy of cmc dataset**

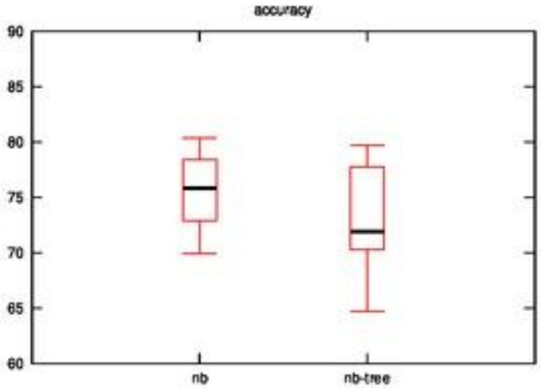

**Figure 7: Accuracy of Diabetes dataset**

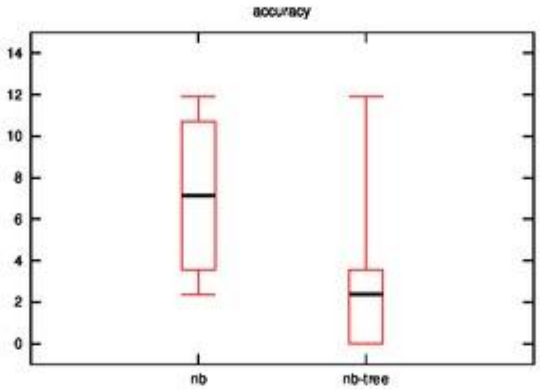

**Figure 8: Accuracy of image dataset**

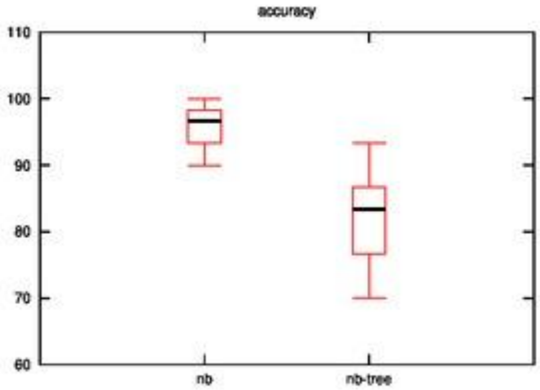

**Figure 9: Accuracy of Iris dataset**

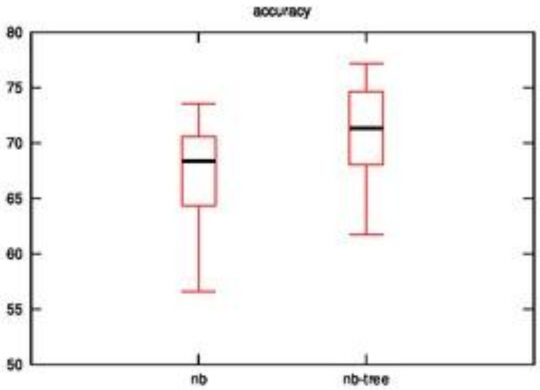

**Figure 10: Accuracy of Soybean dataset**

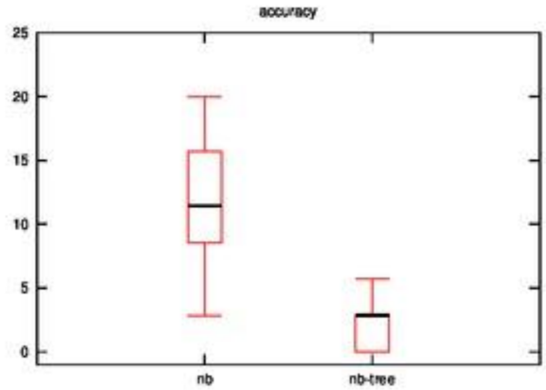

**Figure 11: Accuracy of Wine dataset**

The following are the set of graphs displaying the precision and accuracy values. The classes are color coded.

As we can observe, the points when NB-Tree is used are more scattered when compared to NB alone. This means that the NB-Tree is varying more as compared to NB and is not as efficient as we have assumed it to be.

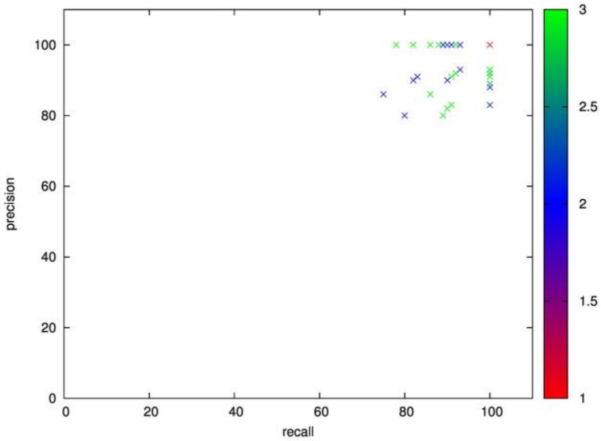

**Figure 12: Iris Precision and Recall**

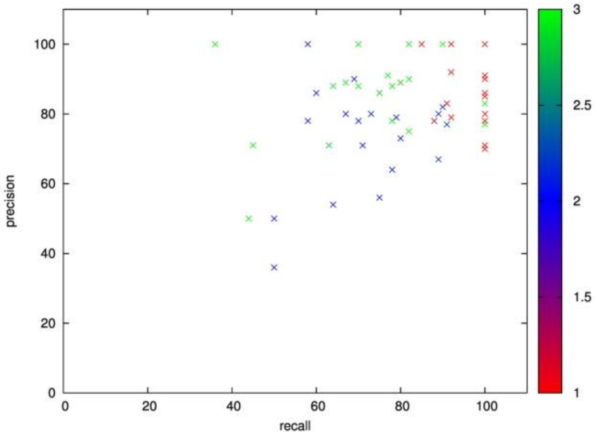

**Figure 13: Iris with NB-Tree Precision and Recall**

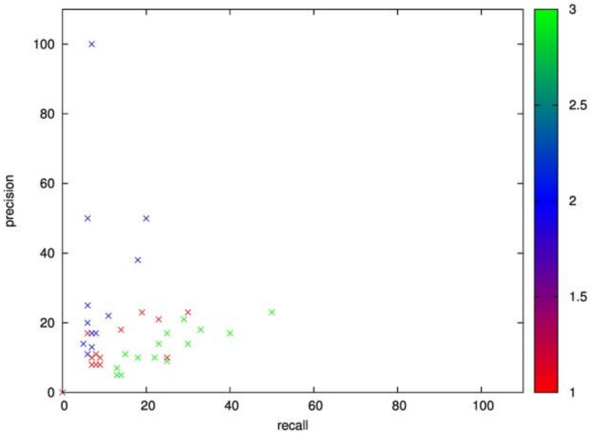

**Figure 14: Wine Precision and Recall**

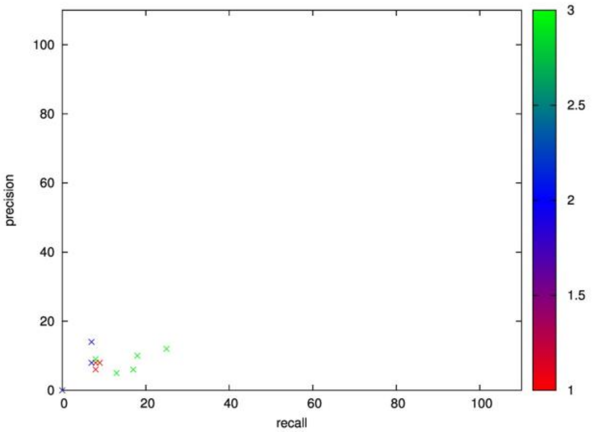

**Figure 15: Wine with NB-Tree Precision and Recall**

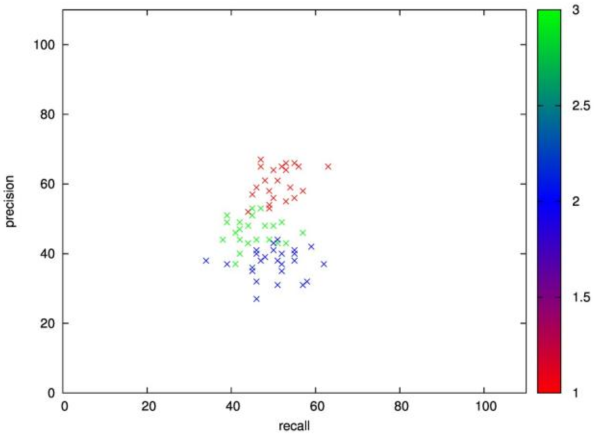

**Figure 16: Cmc Precision and Recall**

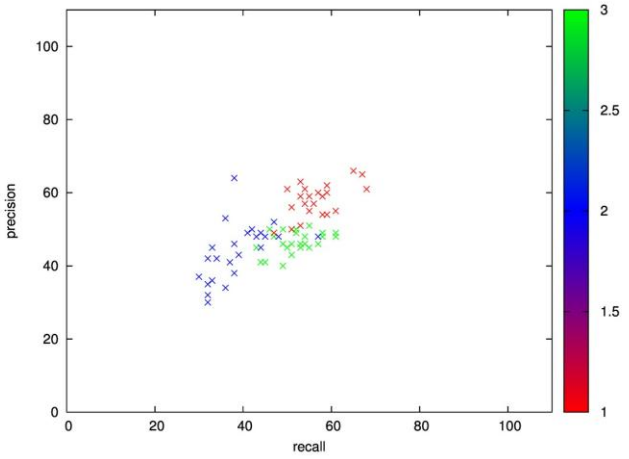

**Figure 17: Cmc with NB-Tree Precision and Recall**

## 5. THREATS TO VALIDITY

The hypothesis is tested on a limited number of datasets and the working of the algorithm might be a threat for the results acquired.

The method of calculating leaves first and then comparing them to test row to find the nearest leaf is also under scope. The leaves that are generated need not be similar to the test row. There might be a possibility that NB classifier on a node be more successful than on the leaf i.e. a point between the root and the leaf.

By the work of Menzies, et al at West Virginia University, it is proven that when datasets are looked at locally, the classifiers generate better performances.

## 6. FUTURE WORK

Currently the possibility of generating the best possible “node” from the tree of tables is considered over the concept of generating a list of leaves. This approach is currently under development and is generated by calculating distance of test row from each node instead of just leaf. The best possible node/leaf is fed into Naïve-Bayes classifier to predict the class.

Precision and Recall can be used to assess the performance of predictor instead of basing the theory over accuracy.

Also, the theory is to be tested on different datasets to prove its performance under different scenarios.

## 7. CONCLUSION

The NBTree proves to be effective for datasets that are concentrated over xy-plane rather than datasets that are spread across the plane. The performance of predictor can be improved by generating best possible node rather than leaf.

Preliminary results of precision and recall generate same results of accuracy, where predictor works well on cmc and soybean whereas NB dominates in the rest.

## 8. MORE INFORMATION

More information regarding the project and its results can be found at http://nave91.github.io/teak-nbtree/ and to contribute to the project, please refer to https://github.com/nave91/teak-nbtree.

## 9. REFERENCES

[1] Ekrem Kocagunelli, Tim Menzies, Ayse Bener, and W. Keung. 2011. Exploiting the Essential Assumptions of Analogy-based Effort Estimation. *IEEE transactions on Software Engineering*.

[2] Ron Kohavi. 2011. Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision-Tree Hybrid.

[3] Tim Menzies. 2013. https://menzies.us/cs573/