CONTENT BASED IMAGE RETRIEVAL USING MACHINE LEARNING TECHNIQUES

Saurin Shah

(Poly ID: 0440976)

Table of Contents

1. Introduction

1.1 Abstract

In this report we consider the problem of Content Based Image Retrieval(CBIR). It is one of the most important areas of research nowadays as the image database associated to any document online or otherwise is increasing. Hence, it becomes that much more important to come up with effective solutions for data mining of images. The application of CBIR can be found in many fields right from search engines to Forensic Investigation. It also has various medical application. Medical imaging has been for quite some time one of the most important areas of research, to be able to apply data mining to medical data can prove to be a very efficient and optimum use of medical data. We use various machine learning algorithms to classify the images and certain of their spatial features as attributes.

1.2 Problem Introduction

The objective of using machine learning algorithms for CBIR is to reduce uncertainty and minimize redundancy and give broader scope of classification. The classification technique used is the root of the problem. This paper is an alternative solution for the paper *Grayscale Image Retrieval using DCT on Row mean, Column mean and Combination* [1]. In the given paper they have taken the DCT of Row mean, Column mean and used method of Euclidean distance for extraction of image from the database. However, the efficiency of the system, time and performance wise can be improved with use of machine learning techniques. The main aim for this paper is to fully utilize the richness of the data and provide better classification results. Euclidean distance has some limitations which can be countered well with the learning of the data.

The test dataset is of 519 images spread across 6 categories and was adjusted from the original database to fit the scale of this particular project. The algorithm given [1] is tested for the modified database and the result stated correspond to it. The categories are,

i. Dinosaur (plain white background)
ii. Elephant (scenic background)

iii. Rose

iv. Scenic Background

v. Airplanes

vi. Sunsets

Sample Images from the database

1.3 Motivation

The main motivation behind this report is the countless possibilities of the application. The result even without the using of learning techniques is commendable. The data and the search scope are perfect environment to use learning techniques.

2.Design and Implementation

Proposed CBIR Techniques[1]

Image retrieval mainly has two steps Feature Extraction and Query Execution. Mainly four different techniques are used here for image retrieval, which are listed below.

i. DCT Row Mean,
ii. DCT Column Mean and
iii. DFT Row Mean.
iv. DFT Column Mean

Here first the row mean and column mean of an image are found and then discrete cosine transform and Fourier transform is applied on them to get feature vectors of image for respective technique of image retrieval.

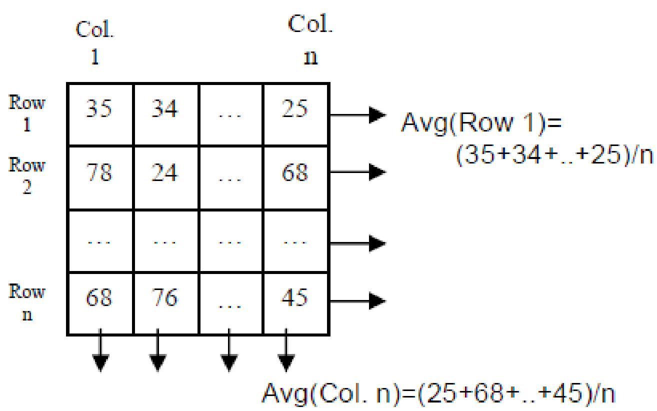


Fig. 1 Sample Image Template (with size nxn)

A. Obtaining Row Mean and Column Mean.

The row mean vector is the set of averages of the intensity values of the respective rows. The column mean vector is the set of averages of the intensity values of the respective columns (1)(2).

If fig.1 is representing the sample image with 4 rows and 4 columns, the row and column mean vectors for this image will be as given below.

Row Mean Vector =[Avg(Row 1), Avg(Row 2), ….,Avg(Row n)](1)

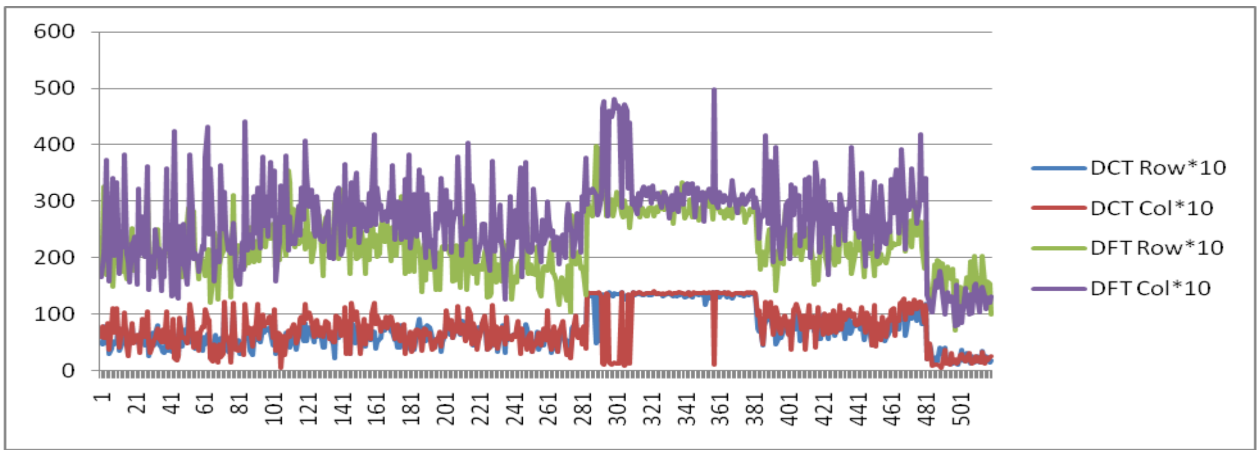
Column Mean Vector =[Avg(Col. 1), Avg(Col. 2), ….,Avg(Col. n)](2)

B. Feature Vectors Extraction

DCT And DFT(Fourier Transform)

Like any Fourier-related transform, discrete cosine transforms (DCTs) express a function or a signal in terms of a sum of sinusoid with different frequencies and amplitudes. Like the discrete Fourier transform (DFT), a DCT operates on a function at a finite number of discrete data points. The obvious distinction between a DCT and a DFT is that the former uses only cosine functions, while the latter uses both cosines and sines (in the form of complex exponentials). However, this visible difference is merely a consequence of a deeper distinction: a DCT implies different boundary conditions than the DFT or other related transforms[2].

The DCT and DFT can be applied to the row mean and column mean vectors of image to get DCT and DFT row mean and DCT and DFT column mean feature vectors respectively. For DCT combination, the DCT,DFT row mean and DCT,DFT column mean feature vectors are considered together. The generated DCT,DFT coefficients will be playing the role of feature vectors of the image which can further be used for image retrieval.



The above figure shows the result of DCT and DFT results. The Y axis is the value and X axis is the image ID. As stated before there is fair amount of correlation of data enabling it for application of learning techniques.

C. Classification

In order to give a comprehensive understanding of the learning algorithms used the classification was done over two dimensions. One, the classification techniques and second, training testing data split and cross validation. This accomplished many goals such as

i. Overall comparison of algorithms
ii. Better understanding of working of algorithm by comparing the performance for various size of test datasets and also cross validation.

The learning techniques used were

i. Bayes: Bayes Net and Naïve Bayes
ii. Functional: Logistic Regression and Perceptron (Multilayer)
iii. Tree: J48 without pruning and J48 with pruning(Minimum number of objects=15)

The various test cases used were

i. Training Set: Where the entire training data was used for testing.
ii. 66% Split: The training data to testing data split was 2:1.
iii. 80% Split: The training data to testing data split was 4:1.
iv. 5 Fold Cross Validation.
v. 10 Fold Cross Validation.

These tests were done in WEKA with standard setting unless mentioned otherwise.

3.Results and Analysis

Part 1: Overall Comparison

Average Precision achieved by the original method was 0.3381 and recall was 0.2293. The minimum Precision achieved (J48 with Pruning) is 0.8054 and minimum recall achieved is 0.7994 The large difference achieved is contributed by 2 key facts

i. Inclusion of another parameter of Fourier Transforms
ii. Consideration of all the parameters while classifying.

The original method used Euclidean distance of each of the parameters. This method was not as successful because it was not able to consider all the parameters together. It took Euclidean distance of DCT of rows, columns separately and compared the results. Using machine learning algorithms we get the advantage of considering all the parameters, they can be considered as nothing but different set of attributes. The biggest advantage of using machine learning algorithms is we can add as many parameters to improve our search. With each useful attribute added our classification becomes more precise. We have the power to include/exclude attributes as per our choice. More attribute just makes classification richer in quality.

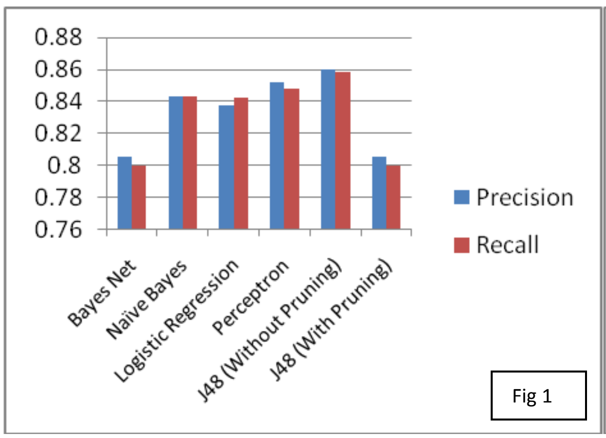


Fig 1

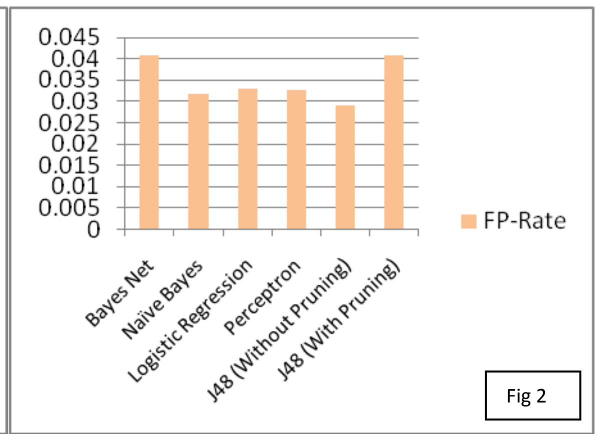


Fig 2

Fig 1 and Fig 2 are overall comparison of Precision, Recall and False Positive Rates. As evident from figure 1, that J48 algorithm without any pruning is supposedly the algorithm giving best precision and Recall. And the worst results are given by Bayes Net and J48 with pruning. It is very important to note here that these results do not directly reflect the performance. The important conclusion from the data above are

i. J48 without pruning gives the best results, but that can be on account of overfitting. The sample size is not very high (519 images). Hence, a decision tree made without pruning can give good result but it is not indicative of the performance. It is demonstrated in the next section.

ii. Naïve Bayes shows remarkable performance over Bayes Net and J48 (with pruning).and logistic regression to a certain extent. One of the main reason contributing to this fact is that in Naïve Bayes each distribution(parameter) is considered as an individual parameter without any dependence This assumption may seem far reaching but as demonstrated, Naïve Bayes performs really well.

iii. Comparison between Functional algorithms Logistic Regression and Perceptron algorithm is interesting. Overall it is noted that Perceptron performs better compared to Logistic Regression. Now, keep in mind that this result is an average over 5 different kinds of tests based on different size of sample data and cross validation. Logistic Regression is known to be a culprit for overestimation when it comes to small sample data, but the overestimation factor diminishes with increase in sample data. Perceptron also faces similar issues but is comparatively a stricter approach and works well in smaller datasets. Logistic regression tries to smooth out the process and allows various adjustments(noise) that Perceptron would not, and that is beneficial if we are looking at larger datasets.

iv. J48 with pruning may not give the best results but is definitely an important algorithm since it counters the problem of overfitting in J48 and gives better results for larger datasets. Again the FP rate of J48 just reiterates the fact of overfitting

v. Even though, False Positive rate can be directly reflected from the Precision and Recall it is interesting to observe. Bayes Net gives the highest FP rate, it is a reflection of interdependencies of various attributes. The rate is still not that high but is the maximum when compared to others. In our case, the dependencies of the attributes does exist to certain extent since all are in one way or the other spatial representation of the images. The performance of the Bayes Net would improve of course with more data but more importantly more attributes which are dependant on each other. For example, inclusion of wavelet transform which are in a way an extension of Fourier Transform would improve the performance of Bayes Net.

Part 2

Training Set

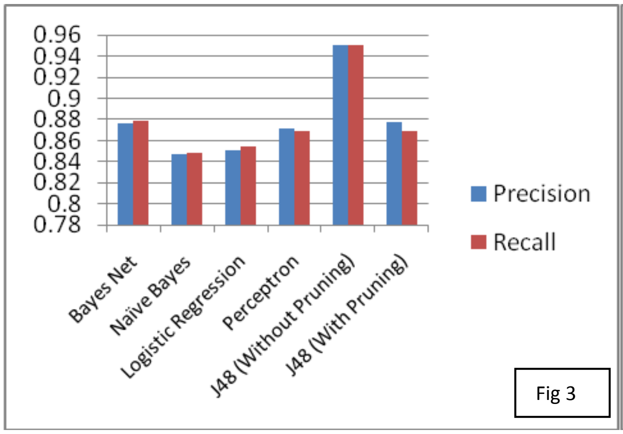


Fig 3

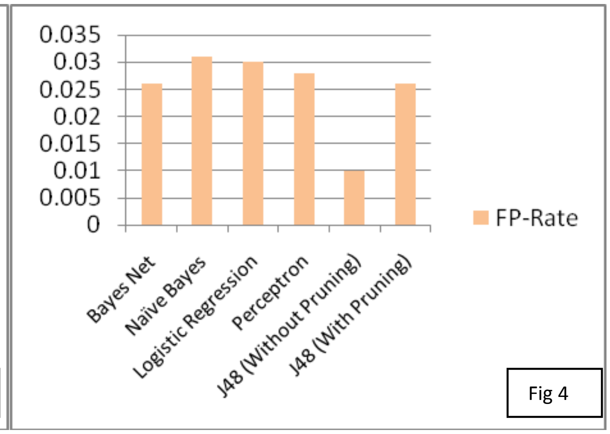


Fig 4

Fig 3 and Fig 4 are results of all the algorithm when the test data set was taken as the Training dataset itself. The observations are

i. J48 gives unreasonably high performance compared to other algorithm. It is the classic example of overfitting. Since, the algorithm was trained to accommodate the training data when we used the same data for testing the algorithm performed much better.

ii. Naïve Bayes gives the worst results, however it is just a matter of perception. It performed well but compared to others it didn't. The reason again is its independence on other variables giving it slightly loose classification compared to others. The performance of each algorithm had to be the best as the testing data was same as training data.

66% Split

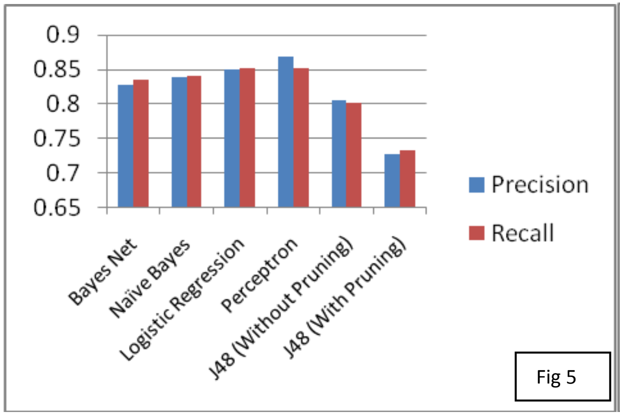


Fig 5

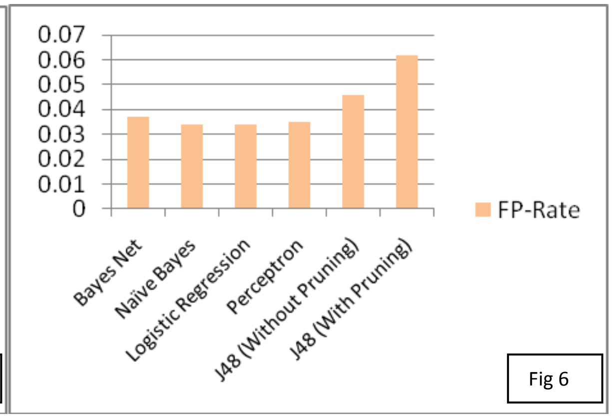


Fig 6

80 % Split

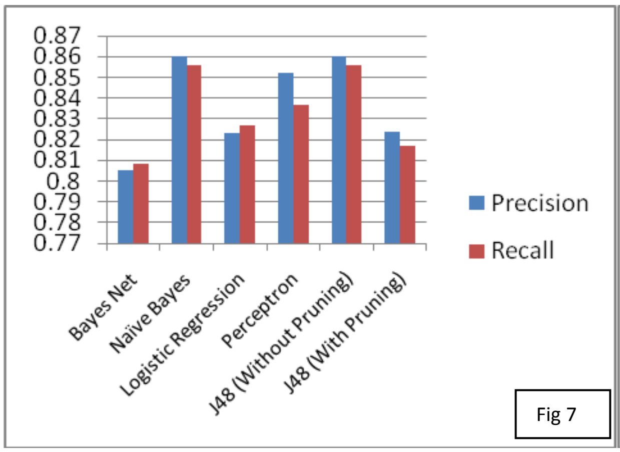


Fig 7

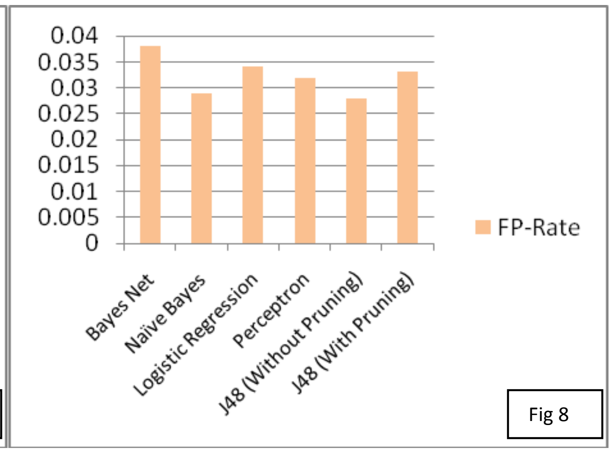


Fig 8

The results of 66% training to test data split and 80% training to test data are shown above. The important observation to note from above are

i. As the separation of training data and test data occurs Naïve Bayes suddenly becomes an excellent algorithm. This points towards the excellence of Naïve Bayes and also supports the fact of underlying assumption of Naïve Bayes.

ii. Performance of J48 also improves as inclusion of more data. The decision tree becomes more refined with more data and gives better results.

iii. Difference in performance of Logistic Regression and Perceptron algorithm increases with inclusion of more data, this can be attributed to the fact that with more data the Perceptron algorithm strives for convergence better compared to logistic regression.

5 Fold Cross Validation

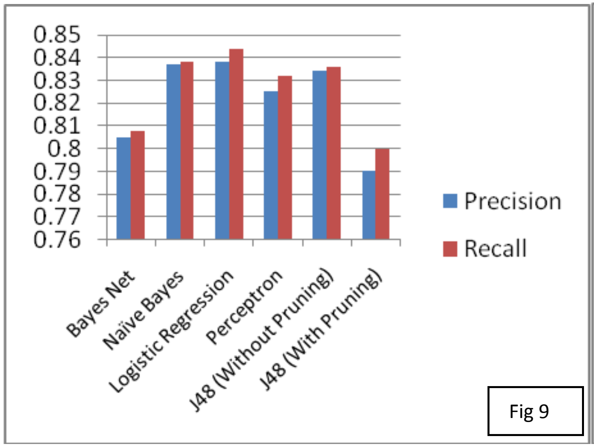


Fig 9

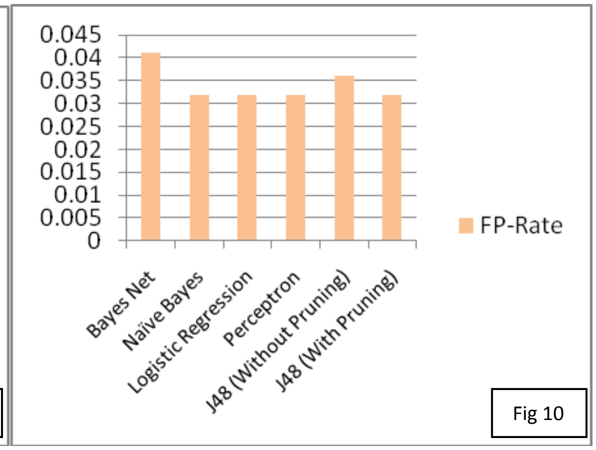


Fig 10

10 Fold Cross Validation

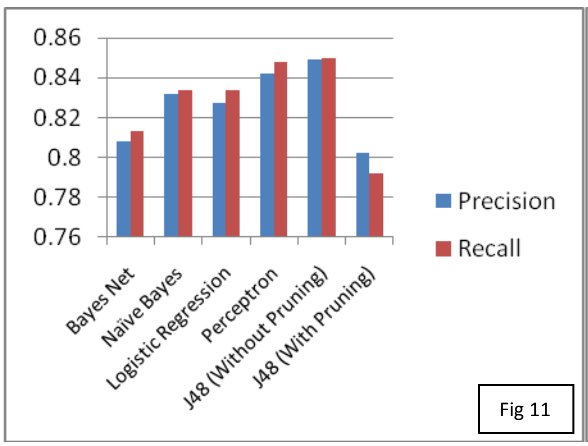


Fig 11

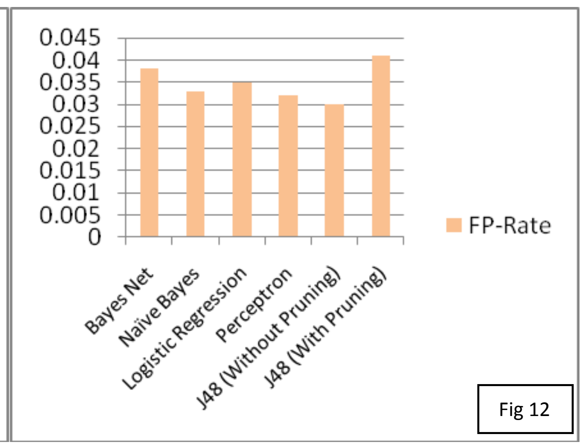


Fig 12

Cross Validation is nothing but measure of correctness of the algorithm. Note, that above cross validation is of unstratified data. The cross validation reaffirms our observations till now about J48, Naïve Bayes and Bayes Net. The surprise element in this testing is performance of logistic regression in 5 fold cross validation. It is indeed a surprise, as logistic regression ideally not supposed to work better than Perceptron in lesser amount of data. Perceptron is a special case of logistic regression. One of the explanations could be the randomness of the data in each fold or the kind of data not allowing a strict convergence. Again, J48 with pruning proves to be not that good an algorithm but as mentioned earlier it should be noted that pruning improves the classification as it allows the inclusion of noise that normally would distort the classification.

Part 3

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 95  1  0  1  1  1 |  a = 1
  2 78  1  8 10  0 |  b = 2
  0  0 93  1  1  4 |  c = 3
  3 12  3 26 17  0 |  d = 4
  0  4  0  8 88  0 |  e = 5
  0  0  3  0  0 58 |  f = 6
```

Above, is an example of confusion matrix of one of the results. We shall focus on only one as others reflect similar results. The observation from the confusion matrix are as follows.

i. Category 1,3 and 6 are classified quite well. Category 1 which is of dinosaurs with plain background, 3 which is of roses and category 6 which is of sunsets have spatial properties that make it easy for us to classify them.

ii. The error that is caused is mainly caused due to category 2 (Elephants in scenic backgrounds), category 4(scenic backgrounds, mountains) and category 5 (Airplanes in the sky).

iii. This actually is not the failure of the system but indeed a measure of correctness, for the background of sky in a picture of an airplane can also be the sky in the scenic background. The background of the forest in a photograph of an elephant can indeed be the scenic background of mountains.

iv. We can modify our algorithm in such a way that whenever we get a photograph whose classification is not certain we can return the possible list of categories that it can be a part of. Hence, if in our search give as input an elephant in the mountains, then it should give as result categories of elephant as well as the mountain.

v. Hence, the accuracy of all the algorithms can be further increased and our goal can be achieved.

4.Conclusion

As illustrated above it is clearly seen that machine learning techniques are for more effective as compared to Euclidean distance. The reason for this is flexibility the learning algorithms give in selecting the attributes. Euclidean distance can be done for a single parameter at a time whereas learning algorithms can have as many parameters as possible. In comparison of various machine learning algorithms J48 gives the best result. Perceptron also gives consistent result. But, it should be noted that convergence in case of Perceptron and overfitting in case of J48 can be prime reasons for this. This just indicates that for smaller databases these algorithms should be used. Naïve Bayes, gives a steady performance but is the most cost effective because of the assumption it makes. Performance of Bayes Net improves at a faster rate with inclusion of more data as compared to Naïve Bayes. Also, it should be noted that there are some anomalies in the result. This is just attributed to the fact that the test result are as much dependant on the algorithms as they are per the data. A dataset may favor an algorithm over the other. In that it is our job to determine which algorithm to take as per the dataset at hand. The wide comparison made was just to demonstrate the above fact. In our database if our size is small we should go for J48 or Perceptron if not Naïve Bayes, Bayes Net, etc. might be a good idea.

5. Future Work

i. Building a GUI: Building a GUI which would take as in input an image and display images of that particular category after it is classified. Display different categories if there is a conflict.

ii. Self classification for search engines: Develop an algorithm that would be similar to the model of a search engine. That is, modify the classification and set of retrieved images as per current searches and patterns. On addition of training data identify when new category data is input and classify it appropriately. That is grow the number of classes as and when new data is input.

iii. Better Attributes: Add new richer spatial attributes which make the classification better.

6. References

[1] Dr. H.B.Kekre, Sudeep Thepade, Archana Athawale, Anant K. Shah, PrathameshVerlekar, Suraj Shirke "*Grayscale Image Retrieval using DCT on Row mean,Column mean and Combination",* at Journal of Sci., Engg. & Tech. Mgt. Vol 2 (1), January 2010.

[2]Wikipedia.com.

Data: Data(Images) has been acquired from Mr. Anant Shah Masters student at Colorado State University(co-Author of [1]). The data designed for use of WEKA by me.

Code: Code is self written. That includes implementation of the algorithm mentioned in [1] and modifications for implementing this paper.