Feature selection using PCA and Information Gain

Abbad Maliyekkal *Roll No:B190372CS*
Abhinav Ajithkumar *Roll no:B190461CS*
Antony Martin *Roll no:B190514CS*
Mohammed hashim *Roll no:B190253CS*
Nihal Muhammed Ashraf *Roll no:B190721CS*
Abhiram Nair MR *Roll no:B190079CS*
Shamil Sulaiman Karadan *Roll no:B190507CS*
Scaria Antony *Roll no:B190379CS*
Vishnu Ajay *Roll no:B190368CS*

Abstract— **The datasets having high dimensionality are difficult to classify as they contain lot of attributes having redundant and irrelevant information.Irrelevant and redundant attributes have also had a negative impact on the performance and accuracy of existing classification algorithms. The existing algorithms lack performance due to this high dimensionality in datasets, thus this model selects only relevant features of a datasets that gives maximum information about the dataset and thus the classification algorithms improve their performance after feature selection.Feature Selection and classification have previously been widely applied in various areas like business, medical and media fields. Some existing work uses all attributes for classification, some of which are insignificant for the task, thereby leading to poor performance.The hybrid model is then applied to support classification using machine learning techniques e.g. the Naïve Bayes technique.This experiment is carried out to increase the performance and efficiency of classification algorithms by dimensionality reduction and selecting only principle features that relevant for classification of the dataset.Thus the proposed hybrid model selects relevant feature sets before the data is pushed for classification.This model provides better performance as measured by accuracy,precision and recall.**

I. INTRODUCTION

Classification is the process of grouping of data points into classes which they belongs to, it is important to classify data because we need to know the category to which the data belongs.These classification algorithms in data mining helps us to organise and classify different types of datasets,they may be either complex or simple. The classification algorithms involved in this process can be easily modified to improve the performance of classification and quality of data. There are many challenges for classification. High dimensional data is one among them. There can be some features that are redundant and not of much use which affects complexity and performance of algorithms used for classification.

Curse of Dimensionality describes the explosive nature of increasing data dimensions and this results in increased computational complexity required in the processing,analysis and classification of data. So the solution is to reduce such features or avoid them during classification, that is select features or attributes that are relevant and not redundant.

Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset,dimensionality reduction achieves the goal by creating new attributes and checking for their combination in contrast to this feature selection methods does the job by including and excluding features that are present in the data without modifying them. Feature selection is the process of disregarding the input variables of a dataset when developing a classified model.It is convinient to decrease the irrelevant features inorder to reduce the computational efforts and time in modelling and this also ensures a better performance for the obtained model.

This document is a model for Feature selection on datasets to reduce its high dimensionality so that only relevant information is taken for classification of the dataset. This paper proposes a hybrid model for selecting relevant features and classifying data that results in dimensionality reduction of the data, reduced training time and provides improved performance of classification using the selected features which are relevant and contains maximum information about the data. It is anchored on principal components and Information Gain. This hybrid model uses a filtering technique that incorporates Principal Component Analysis (PCA) and Information gain for determining the most relevant features. It ultimately aims at selecting the best set of features from the first data which will give good classification.

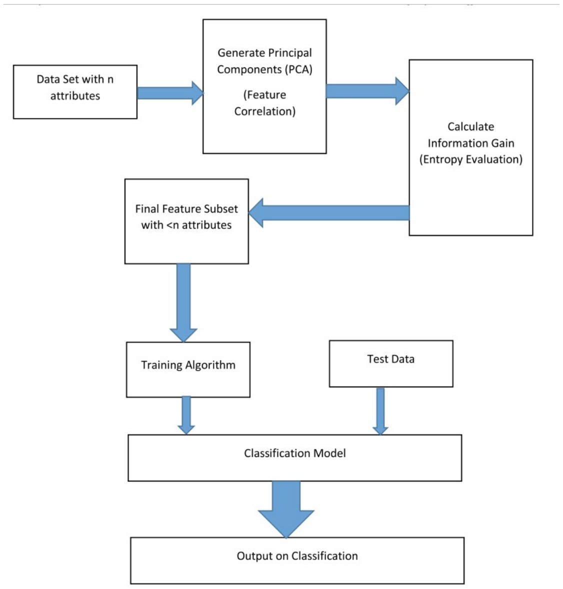
Data Set with n attributes
Generate Principal Components (PCA) (Feature Correlation)
Calculate Information Gain (Entropy Evaluation)
Final Feature Subset with <n attributes
Training Algorithm
Test Data
Classification Model
Output on Classification

Fig. 1. Hybrid Model

II. ABBREVIATIONS AND ACRONYMS

A. Abbreviations

Cov-Covariance

B. Acronyms

PCA- Principal Component Analysis
IG-Information Gain
SVM-Support vector Machine

III. INPUT AND OUTPUT

Input : X = (x1,x2,x3. . . . .,xn).
The input is a n dimensional training dataset with n features. Here x1,x2,x3. . . .xn are n features. This may contain redundant and irrelevant features.

Output : Y = (y1,y2,y3,..........ym). where m less than or equal to n.
The output is a m (m$\leq$n) dimensional feature set. Output does not contain any irrelevant or redundant features.Thus the original dataset X is reduced to a lower dimensional dataset having only the principal components.

IV. EQUATIONS

A. PCA

PCA is used for dimensionality reduction. It includes the following steps:
Calculate the mean of X:

$$(x)' = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

This standardizes the values and it will not result in a biased outcome and we can calculate the covariance from the above mean

$$Cov(x) = \frac{1}{N}\sum_{n=1}^{N}(xi - xi')(xi - xi')T$$

Using the above covariance matrix we can find the eigen values and eigen vectors which can be used for the spectral decomposition of the same, the eigen values are arranged in descending order and only the first required values are taken.

Y = (y1, y2, y3, ..., yP) This is the dataset which we require for classification which contains only principal components and is d-dimensional which is less than the original dataset.
Now the datasets dimensionality is reduced, now it contains only principal components.

B. INFORMATION GAIN(IG)

Let A be a attribute, C be the class,the expected information before,

$$H(C) = -\sum p(c)log_2 p(c)$$

Next find the expected information after the observing the attribute

$$H(C/A) = \sum p(a)\sum p(a)P(c|a)log_2 p(c)$$

where a is one of the attribute and c is one of the class label

Final procedure is to calculate the information gain. It is the difference between H(C) and H(C/A)

$$IG(A/C) = H(C) - -H(C/A)$$

V. ALGORITHMS

A. PCA

- Standardization of initial variables
  Xi=(Xi-mean(X))/std.deviation(X)
- Find the covariance matrix to get the correlations between dimensions
  cov=covariance(X)
- Find the principle components using the eigenvalues and eigenvectors of the covariance matrix
  eigenvalues(cov),eigenvectors(cov)
- Sort the obtained eigenvalues to eliminate the irrelevant principal components
- Project the data along the principal components axes which has minimum information loss

B. IG

Here we calculate information gain of each principal components and after that we choose those components whose information gain value is greater than the threshold value.

- The expected information or entropy to classify each row in dataset is Hc.
- For each principal componet:
  Calculate the information needed or entropy after observation of that principal component, that is calculate H(C/A), where C is set of all class labels and A is observed attribute
  Information gained is IG(A/C)=Hc-H(C/A).
- Now we select final set of attributes using the information gain value and the threshold value

VI. TIME COMPLEXITY OF PCA-IG ALGORITHM

The time complexity of the PCA-IG filter model is O(n3).In the above mentioned algorithm the spectral decomposition of the covariance matrix and the eigenvectors and eigenvalues has a time complexity of O(n3) all other operations in the algorithm has lesser time complexity than O(n3).

Therefore the overall time complexity is:
T(n)=O(n3)

VII. DATA SET

The dataset used for this mini project is breast cancer data set.This dataset contains thirty attributes and is used to indicate if cancer is present or not each record or tuple represents medical state or condition of a patient.The dataset is different from the one which is given in the research paper which is breast cancer dataset having only nine attributes.

VIII. RESULTS

A. Performance when all features are used

The performance exihibited by all the classification algorithms where poor when all the features of a dataset are used.The implemented results of the breast cancer dataset shows that when all the features are used the Naive Bayes algorithm gives higher accuracy as compared to decision tree and SVM.The performance of classification is however not optimal for the classification algorithms hence we need to selects the relevant features to the performance of these algorithms.As there are so many features that are not significant for classification,the time and efficiency of classification algorithms are poor thus feature selection helps in improving time and efficiency of classification algorithms.

Algorithm	Accuracy	Precision	Recall	Time
Naive Bayes	92.98%	0.93	0.93	2.00
SVM	92.98%	0.93	0.92	8.00
Decison tree	91.23%	0.92	0.91	9.00

B. Performance after feature selection

In this mini project feature selection was done using PCA and IG thus resulting a PCA-IG hybrid model which selects only the relevant features which contain the maximum information.When the PCA-IG model was implemented results shows that the performance of classification algorithms(Naive Bayes,SVM,Decision Tree) increased as compared to classification when there is no feature selection which is evident from the obtained graph(Fig.2).This graph shows the accuracy of classification algorithms after feature selection is done on the dataset.

Algorithm	Accuracy	Precision	Recall	Time
Naive Bayes	94.74%	0.95	0.95	3.90
SVM	92.98%	0.93	0.93	6.00
Decison tree	92.10%	0.92	0.93	5.00

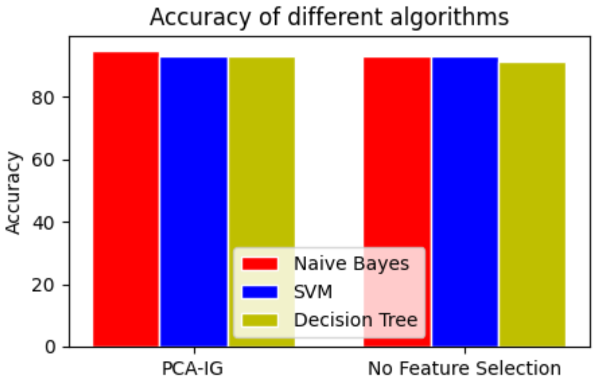

Fig. 2. Accuracy with and without feature selection

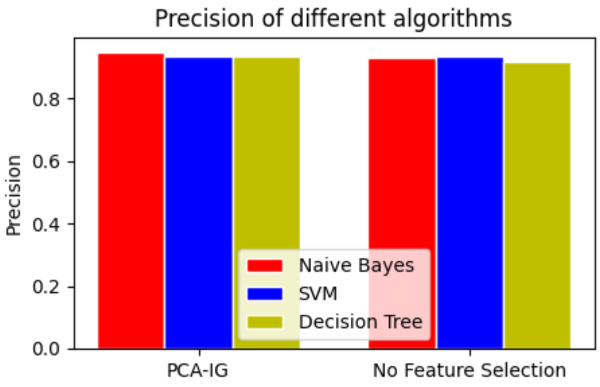

Fig. 3. Precision with and without feature selection

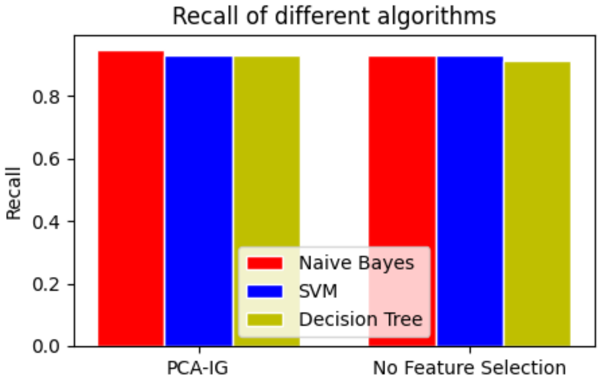

Fig. 4. Recall with and without feature selection

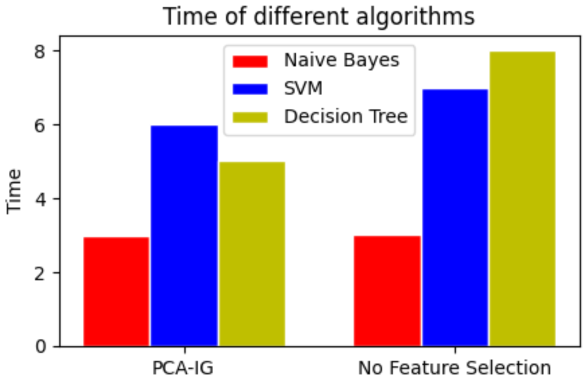

Fig. 5. Time with and without feature selection

IX. COMPARITIVE ANALYSIS

From the above obtained graphs it is clear that the performance measuring metrics like accuracy,precision and recall are more and time is less for PCA-IG model as compared to no feature selection model in most cases but some irregularities also exist in the obtained result.

Before feature selection when the naive bayes algorithm is applied the accuracy is 92.98 precentage but after applying the feature selection using PCA-IG model the accuracy with naive bayes as classification algorithm has raised to 94.74 percentage. Similarly for decison tree also accuracy raised from 91.23 to 92.10 percentage. But in case of SVM accuracy remained same.

In case of precision, by using PCA-IG model and naive bayes as classification algorithm precison increased from 0.93 to 0.95 but remained same for SVM and decision tree.

In case of Recall score all algorithm showed raise in value using PCA-IG model.For naive bayes it raised from 0.93 to 0.95, for SVM 0.92 to 0.93 and for Decison tree from).91 to 0.93.

In case of time naive bayes took more time after using PCA-IG model(3.9ms). previously it was 2ms.In case of SVM time reduced from 8 to 6ms and for Decision tree time reduced from 9 to 5 ms.

By using PCA-IG model naive bayes showed highest accuracy,precision,recall and lowest time among the three classification algorithm.

X. CONCLUSIONS

In this mini project, we implemented a hybrid model (PCA-IG) for feature selection. The experiment results show that feature selection technique improves the performance of the classification algorithms. PCA-IG model yields better results when compared to other techniques in performance,accuracy,computational cost,etc. The performance improved in all scenarios due to feature selection being done before classification. However, the percentage performance was higher when using Naive Bayes classifier both before and after carrying out feature selection. This also applied to the PCA-IG model compared to other feature selection techniques. Further study are often done to determine how this can be improved.

Thus we can conclude that the classification algorithms yield better results in terms of performance and accuracy when feature selection was done before classification. The Principal Component Analysis – Information Gain model that was proposed was ready to reduce dimensionality of data, improve performance and also significantly reduce the training time thus the requirements and aim of suggesting such a hybrid filtering model were met.

REFERENCES

[1] Erick Odhiambo Omuya a , George Onyango Okeyo b , Michael Waema Kimwele c a School of Engineering and Technology, Machakos University, Kenya b Carnegie Mellon University Africa c School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, Kenya