Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction

Akshay K Biju, Devika Sivakumar, Jeswell Mathew, Muhammd Azzaam,
Reuben Suju Varghese, Sachin Thomas, Sanjana Reji Kallingal, Thapas P pramod

***Abstract*—**A dataset contains a huge number of input features in various cases, which makes the predictive modeling task more complicated. To overcome the difficulty to visualize or make predictions for the training dataset with a high number of features, dimensionality reduction techniques are required to be used. In this paper, a fast hybrid dimensionality reduction method for classification is proposed.Multi-strategy based feature selection is used to filter out irrelevant features while the grouped feature extraction is used to remove redundancy among features. Firstly in order to compress the high dimensional dataset , the intrinsic dimensionality of the data set is estimated by the maximum likelihood estimation method. Feature selection using Fisher Score and Information Gain are used to remove irrelevant features.Following which clusters are formed based on the redundancy among the selected features. In every cluster,Principal Component Analysis (PCA) based feature extraction is carried out to remove redundant information.The runtime results of different methods show that the proposed hybrid method reduces dimensionality of raw data sets ,is consistently much faster than the other three in almost all of the sets used and has excellent efficiency and competitive performance compared with contrastive methods.

1.INTRODUCTION

This has become an era of big data where information has transcended the limits of the imaginable. Tons of data with far bigger dimensions require processing in order to extract useful information. For that very purpose, the need to reduce the dimensionality and remove redundant data in order to reduce the time and effort required for processing comes into need. Dimensionality reduction can filter out a lot of noises and make it easier for data processing, knowledge mining, and pattern classification. While extraction can reduce the size of data that needs processing, it paves way to a loss of a lot of relevant data.

To avoid this issue and to come up with an efficient way to reduce the dimensionality of data sets, in this project, we have tried to implement the theory explained in the "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction"[1]. The general method for dimensionality reduction could lead to a lot of information redundancy. Many irrelevant features that are correlated to each other will also be integrated into the group, resulting in the existence of irrelevant features in the final subset. As a solution to this, a series of multi-stage hybrid dimension reduction models were proposed to remove irrelevant, redundant, and noisy features. So using a combination of clustering strategies and hybrid operations, we have created a fast hybrid dimensionality reduction framework using the steps mentioned below:

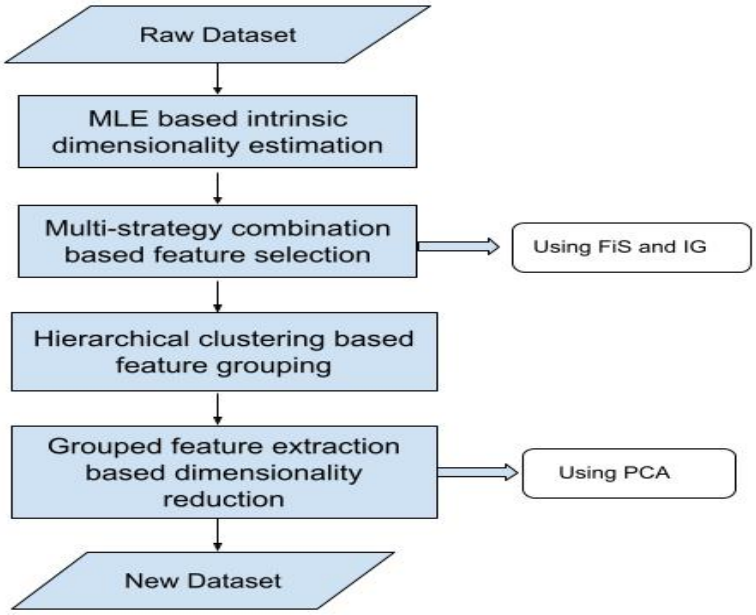

Fig. 1: Framework of the proposed methodology

- MLE based intrinsic dimensionality reduction.
- Multi-strategy combination based feature selection
- Hierarchical clustering based feature grouping
- Grouped feature extraction based dimensionality reduction

We have managed to implement a model as close to what is theorized on the paper cited before using the python programming language.

A detailed explanation of the steps involved on the proposed framework along with the result analysis will be shown below.

2.FRAMEWORK

2.1 Intrinsic Dimensionality Estimation

The first step in performing fast hybrid dimensionality reduction is calculating the raw dataset's intrinsic dimensionality using MLE (Maximum Likelihood Estimation) method. After this,to eliminate the peripheral features a multi-strategy approach of a combination of two different feature selection methods are used i.e- Fisher Score and Information Gain. To further remove the redundancy, the selected features are grouped into clusters on the basis of amount of redundant information in each. Then, in every cluster PCA (Principal Component Analysis) based feature extraction is implemented. Thus the new dataset is produced with the irrelevant features removed.

In Intrinsic Dimensionality Estimation, what we essentially do is take high-dimensional data and calculate the minimum number of free parameters required to describe it fully and understand the internal structure of it. Intrinsic dimensionality meth-

ods can be divided into 3- Projection-based Geometry-based Probability based Here we use a probability based estimation method - MLE where the maximum likelihood function is used to construct the distribution hypothesis of the data and obtain the most efficient results.

2.2 Multi-Strategy Combination Based Feature Selection

Coming to the Multi-Strategy Combination Based Feature Selection, a subset of relevant features based on defined criteria is selected and maintains the important features from the raw dataset.For this a combination of feature selection strategies based on Fisher Score(FiS) and Information Gain(IG) filters the noise information. In FiS, selected features are the ones with large between-class distance and low within-class. IG is based on information entropy to measure the correlation of features and the class label.

Given a data set F = F1,....., FD containing D features, FS1 = F11,.....,F1D — feature sequence ranked by FiS and FS2 = F21 ,....,F2D — feature sequence ranked by IG. To filter low scored features that are there in both feature sequences, we apply the union approach on the lowest C

FS = F — $\{\text{C\%}\{\text{FS1}\} \cup C\%\{FS2\}\}$

2.3 Hierarchical Clustering

The next step is Hierarchical Clustering of the features where we use maximal information compression index as the standard for measuring redundancy between two features. Let be the covariance matrix of two variables x and y, the maximal information compression index can be defined as (x, y) = the smallest eigenvalue of , i.e., $2\delta(x, y) = Sx + Sy\text{-}\sqrt{\left((Sx + Sy)^2 \text{-}4SxSy\,(1\text{-}\rho^2)\right)}$

where Sx=var(x), Sy=var(y), $\rho = cov\,(x, y)\,/\sqrt{SxSy}$.

$\lambda = 0$

Using this as the criterion, hierarchical clustering is used to integrate features that have high redundancy to the same cluster in accordance with the matrix containing between all the feature pairs.

2.4 PCA Based Reduction

Finally, we arrive at the last step of the reduction procedure - PCA based reduction which has variance in data as its criterion.Higher variance corresponds to a larger amount of information.The computation of PCA includes singular value decomposition and projection transformation.The original high dimensional data is mapped to a linear subspace formed by a few number principal components with relatively larger eigenvalues. Lastly,the correlation among the original dimensions is eliminated and the dimension of the data is reduced.

3.IMPLEMENTATION DETAILS

According to the given framework, we have implemented the model in python that is as close to the theorized model in the cited paper. The framework, which consists of 4 main features used to optimize the final step of dimension reduction was implemented separately. Later, each feature was combined to create a seamless model that is optimized for Dimension Reduction.

3.1 Implementing Intrinsic Dimensionality Estimation

We implemented the Maximum Likelihood Estimator which was a Probability based Estimation of Intrinsic Dimension.

3.2 Implementing Feature Selection

Then we implemented the Feature Selection with Multi-Strategy Combination. The different strategies we combined for this process are Information Gain and Fisher Score. To calculate the Information Gain, we use an important concept called Entropy. We calculate Entropy by using the formula:
Firstly, we were required to calculate the Entropy of the entire Dataset as a whole. We then calculate the Entropy of the individual features in the dataset and store them in an array. Finally, Information Gain of the features is calculated by subtracting the Entropy of each feature from the Entropy of the entire dataset.
To calculate Fisher Score of each feature, we use the formula:
Si=$\sum$nj(μij-μi) 2 ($\sum$nj $\rho 2ij$
where μij and ρij are the mean and the variance of the i-th feature in the j-th class, respectively,nj is the number of instances in the j-th class and μi is the mean of the i-th feature.

Using Information gain and Fisher score, we remove features that the combined Feature selection consider as redundant. The way we implemented that was, we took the features that have the lowest C% of Information Gain and features with lowest C% of Fisher Score and we performed a union set on them. Finally subtract this union set of features from the main dataset and get a new dataset with features reduced from implementing Feature Selection with Multi-Strategy Combination.

3.3 Implementing Hierarchical Clustering

Later, we implemented the Hierarchical Clustering of the Features to group redundant information with the maximum linear dependency. To identify features with high linear dependency, we use the help of a Covariance Matrix. A layout of a covariance matrix is given below

$$\operatorname{corr}(\mathbf{X}) = \begin{bmatrix} 1 & \frac{\mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)]}{\sigma(X_1)\sigma(X_2)} & \cdots & \frac{\mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)]}{\sigma(X_1)\sigma(X_n)} \\ \frac{\mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)]}{\sigma(X_2)\sigma(X_1)} & 1 & \cdots & \frac{\mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)]}{\sigma(X_2)\sigma(X_n)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)]}{\sigma(X_n)\sigma(X_1)} & \frac{\mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)]}{\sigma(X_n)\sigma(X_2)} & \cdots & 1 \end{bmatrix}.$$

Using this Covariance Matrix, we calculate the Maximal Information Compression Index which is the criterion we use to measure the redundancy between features.
If the value that we get from the Maximal Information Compression Index between 2 features is 0, we then group those features as they have high linear dependency between them. Hence, we get a reduced dataset after this step.

3.4 Implementing Dimensionality Reduction using PCA

Lastly, we have to implement the Dimension Reduction step onto our reduced dataset. The Dimension Reduction methodology we used is PCA(Principal Component Analysis). PCA measures the amount of information contained in the data. It reduces the high dimensional data and produces Principal components that individually can define the entire dataset. We then can choose the number of principal components required to describe the dataset. Higher number of principal components will describe the dataset better with

low information loss and high presevation of information. We input the dataset into the PCA function and are able to view principal components in a 2D scatterplot to describe the relationship between them.

Thus this is the implementation of the model that was theorized in the research paper assigned to us.

4.MODIFICATION TO PAPER

We have tried to implement the model as close as possible to the framework given in the paper. However we were not able to implement intrinsic dimensionality estimation using the maximum likelihood estimation method that was referenced in our research paper. To get over this roadblock we researched on other similar methods of dimensionality estimation and found a paper suggesting different methods of dimensionality estimation. Out of all the methods suggested, we decided to go forward with an ISOMAP dimensionality estimation method. With this changed method, we were also able to calculate the accuracy of our intrinsic estimation method. We feel that this change makes an improvement in our framework.

5.RESULTS

5.1 Intrinsic Estimation

Intrinsic Estimation of various datasets:

Dataset	Intrinsic Estimation
Diabetes	5
Breast	24
Heart	10
Inflation	5
WeatherAUS	17
Wine	8

5.2 Fisher Score/Information Gain

Initially the Information Gain and Fisher Scores of all attributes are computed seperately.Then to filter low scored features that are present in both feature sequences, we apply the union approach on the lowest threshold of the two sequences and filter out them from the original feature sets.

Entropy: 3.584962500721156

Entropy of each attribute is given below:

Attribute:	Information Gain
Price_beef_kilo:	0.9558541137054539
Price_rice_kilo:	02.597957369981772
Price_coffee_kilo:	0.7835874125082222
Inflation_rate:	3.5797412899532026
Price_rice_infl:	2.5665186251775243
Price_beef_infl:	1.0810858424957923
Price_coffee_infl:	0.7618721186738886

Information Gain of each column is given below:

Attribute:	Information Gain
Price_beef_kilo:	2.629108387015702
Price_rice_kilo:	0.9870051307393841
Price_coffee_kilo:	2.801375088212934
Inflation_rate:	0.005221210767953455
Price_rice_infl:	1.0184438755436318
Price_beef_infl:	2.503876658225364
Price_coffee_infl:	2.8230903820472673

The Fisher Scores of the attributes are given below:

Attribute:	Information Gain
Price_beef_kilo:	0.010431739169024241
Price_rice_kilo:	2.568284340463402
Price_coffee_kilo:	0.004147913723075363
Inflation_rate:	7.652614690194511e-06
Price_rice_infl:	3.258598926423663
Price_beef_infl:	0.09338845822948705
Price_coffee_infl:	0.009359676838718274

The calculated redundant values based on threshold values:

	Year	Month	Price_rice_kilo	Price_rice_infl	Price_beef_infl
0	1992	Feb	0.28	0.530000	4.780000
1	1992	Mar	0.28	0.530000	4.640000
2	1992	Apr	0.28	0.530000	4.510000
3	1992	May	0.27	0.510000	4.510000
4	1992	Jun	0.27	0.510000	4.420000
...	...	...	...	...	...
355	2021	Sep	0.40	0.390000	5.590000
356	2021	Oct	0.40	0.390000	5.630000
357	2021	Nov	0.40	0.390000	5.870000
358	2021	Dec	0.40	0.390000	5.870000
359	2022	Jan	0.43	0.476167	4.044006

5.3 Hierarchical clustering of the features

In the project, the Hierarchical Clustering was accomplished by a set of steps, starting with the formation of a covariance matrix. Using the newly formed covariance matrix, we generate the MIC array.

Calculated MIC Array

0	0.58870209	0.02149708	0.03476689
0. 58870209	0	0.2135821	0.03389722
0. 02149708	0. 2135821	0.	0.0052233
0. 03476689	0. 03389722	0. 0052233	0

Once the MIC array has been made, we can easily obtain the resulting redundant attributes using the threshold value.

5.4 Principle Component Analysis

The following scatter plot with two components is obtained after PCA. PC1 is the line in the K-dimensional variable space that best approximates the data in the least squares sense. The second principal component (PC2) is oriented such that it reflects the second largest source of variation in the data while being orthogonal to the first PC. PC2 also passes through the average point. The dimension of data is reduced, the correlation among the original dimensions is eliminated.

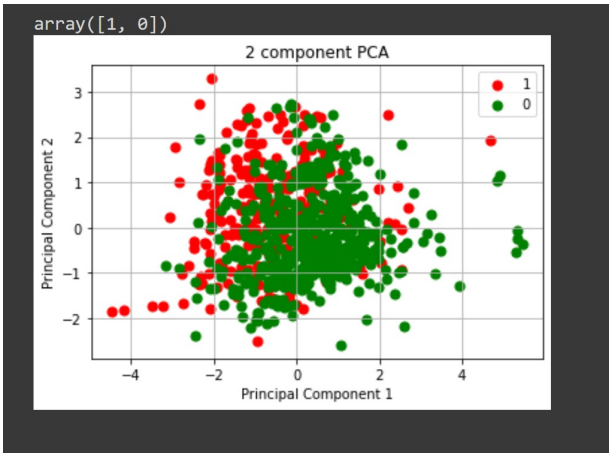

Fig. 2: Diabetes

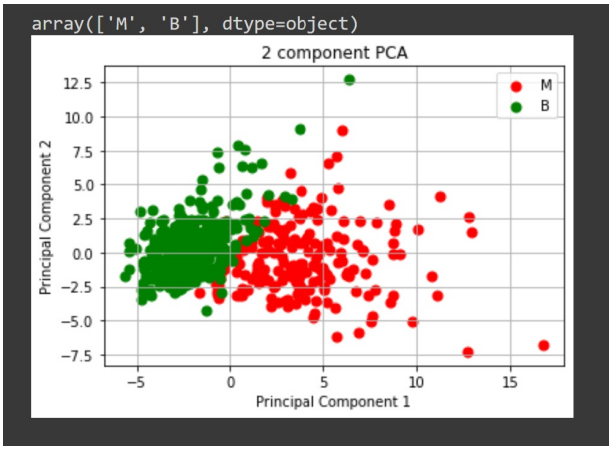

Fig. 3: Breasts

Accuracy tables for various datasets

	Dataset	Accuracy
0	diabetes	0.880
1	breast	0.879
2	wine	0.898
3	heart	-
4	weatherAUS	-
5	inflation	0.883

Comparative Analysis

Dataset	Features	Intrinsic Estimation	IG/FS	HC	PCA
Diabetes	8	5	6	5	5
Breast	30	24	28	27	27
Heart	13	10	7	6	6
Inflation	7	5	5	4	4
WeatherAUS	21	17	13	12	12
Wine	11	8	9	8	8

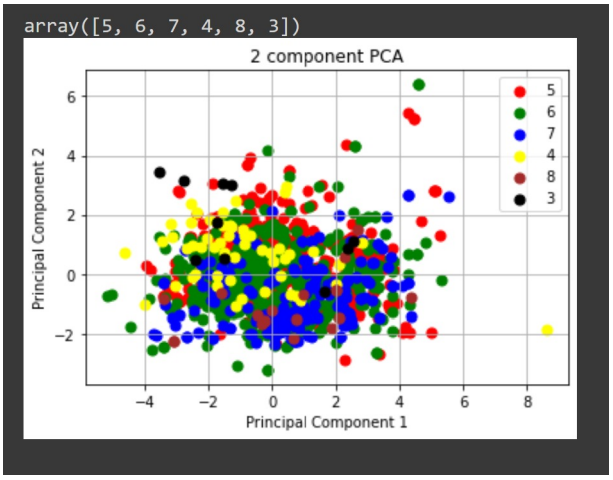

Fig. 4: Wine

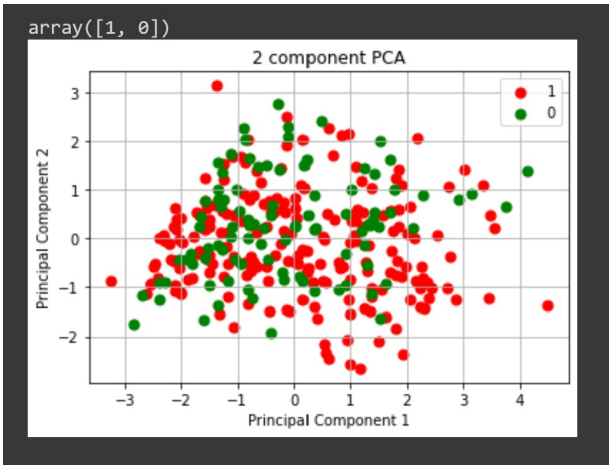

Fig. 5: Heart

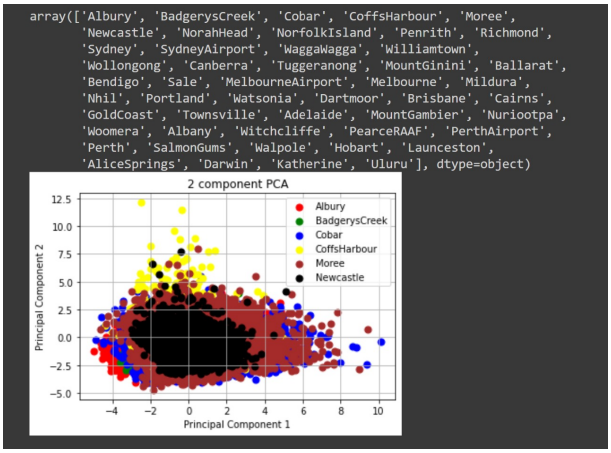

Fig. 6: WeatherAUS

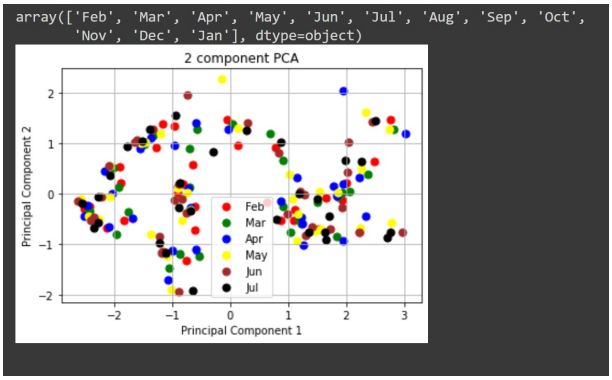

Fig. 7: Inflation(beef)

References

[1] Mengmeng Li et al. “Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction”. In: *Expert Systems with Applications* 150 (2020), p. 113277. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.113277.

URL: https://www.sciencedirect.com/science/article/pii/S0957417420301020.