International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-04, Issue-04, July 2018

Combination of Feature Selection and Extraction technique for Feature set Generation over Supervised Dataset

[1]Ms.Priti P. Shirke, [2]Prof. J. V. Shinde

[1]M. E. Student, [2]Asst. Professor, Late G. N. Sapkal College Of Engineering,Nashik, India.

[1] *pritishirke2@gmail.com*, [2]*jv.shinde@rediffmail.com*

Abstract- **This For high dimensional dataset processing Dimensionality reduction is an important task. For dimensionality reduction feature selection and feature extraction are two important techniques. These techniques are applied independently over dataset. The proposed work aims to provide a solution for dimensionality reduction based on these two techniques simultaneously. This technique works on supervised dataset. A Minimum Projection error Minimum Redundancy MPeMR-supervised technique is proposed. System will find compound feature using feature extraction process using Linear Discriminant Analysis-LDA by combining 2 features. After compound feature generation feature redundancy checking is applied using Correlation Coefficient technique. Using feature selection process redundant features are removed. This is an iterative process, executes until dimension count reduces to the desired user specific count. The system results will be compared with the existing technique in terms of classification accuracy.**

Keywords—classification, dimensionality reduction, feature selection, feature extraction, LDA, Correlation Coefficient

I. INTRODUCTION

The world wide usage of computer technologies creates bulk data. To process and preserve such data is challenging task. In order to meet this challenge some key technique is required to extract knowledge from such bulk data. Data mining technique is used for extracting useful information, patterns. A data with multiple attributes creates a challenge for processing. As number of attribute increases the storage space and processing time also increases. Irrelevant attributes may add noise in data. The noise in data hampers the accuracy of data mining algorithms. Dataset may also contain redundant attributes i.e. 2 or more attributes specifies the same thing. Due to such redundant attributes, the dimensionality of data increases and requires high processing time. To improve the efficiency of data mining algorithm dimensionality reduction is important aspect. Dimensionality reduction is the preprocessing step in which redundant and noisy attributes are removed. For dimensionality reduction approaches are used:

1: Feature selection
2: Feature extraction

Feature selection is a technique that identifies relevant attributes as a feature from dataset. In feature selection Feature subset is generated. The subset generation process follows2 approaches: one is forward selection in which, initially feature set is initialize to zero and then one by one feature is added to the feature set. The another approach is sequential backward selection in which, feature set is initialize with all the dimensions in dataset and iteratively one by one feature is removed. In feature extraction two or more attributes from a dataset are merged together to generate a new attribute. Feature extraction technique searches for minimum number of new attributes with the help of some transformations according to some performance measure. The study of these two technique is broadly classified in 2groups:

1: supervised learning
2: unsupervised learning

In supervised learning, a class attribute is present. The relevance of attribute is compared with the class attribute. Supervised dataset are generally used for classification algorithms in data mining. In case of unsupervised learning, no class attribute is provided. The relevance of attribute is calculated using the variance and separability. Unsupervised datasets are generally used in clustering algorithms of data mining. The evaluation criteria of generated feature set is varies with respect to the data format. For supervised dataset classification technique is used for accuracy measurement whereas clustering technique is used for unsupervised dataset to check the correctness. There are various techniques proposed for feature selection and feature extraction. In the following section these techniques are studied in details with their advantages and limitations.

II. REVIEW OF LITERATURE

A. *Feature Selection*:

The feature selection technique is classified in 3 groups:

 DOI : 10.18231/2454-9150.2018.0447

1. Filter:

The feature selection from a dataset is independent of any machine learning algorithm. For feature selection various statistical test are conducted using correlation technique. A fast clustering-based feature selection algorithm is proposed by Q. Song, J. Ni, and G. Wang[2]. This algorithm is executed in 2 steps. Initially it generates cluster of attributes using graph-theoretic clustering. In second steps most representative attributes are selected from clusters those are strongly related to class label.

Prabitra Mitra[3] proposes a feature similarity based feature selection technique. This is unsupervised feature selection technique. Maximum information compress index is the new feature similarity measure is proposed. This is a fast measure of pair wise similarity computation.

2. Wrapper:

Wrapper method selects the features based on some machine learning algorithms. This algorithm tries to satisfy the criteria mentioned by some classifiers such as KNN, Naive Bayes, etc. These methods are computationally expensive and hence not be applicable in real life application with large dataset.

3. Embedded Method:

This method combines the quality of filter and wrapper methods. This method achieves model fitting and feature selection simultaneously. LEAST ANGLE REGRESSION[4] and l2-1 normalized regression[5] are examples of embedded technique .Trace Ratio Criterion[6] is proposed for general graph-based feature selection. This technique uses Fisher score and Laplacian score to calculate the trace score of a feature. spectral feature selection technique is proposed by Zheng Zhao and LeiWang. This technique uses a regression model using sparse multi-output regression technique to minimize redundancy in dataset. These two techniques are supervised feature selection technique that measures the correlation with the class labels.

In [7] predominant correlation technique is proposed for feature selection. Fast Correlation-Based Filter algorithm is proposed in this technique. This algorithm is called as subset selection algorithm. Apart from the feature subset selection feature weighting is another technique [8]. In feature weighting weights are assigned to each attribute in dataset to define the importance of feature. The features can be selected on the basis of feature weights.

Relief[8] is a Instance based approach for feature selection. This algorithm uses two class problem solution. It converts multiclass classification problem to number of binary problems. It takes n instances from a dataset form 2 classes and feature selection process is executed iteratively.

Feature Extraction:

Feature extraction technique transforms the high dimensional dataset to low dimensional dataset by linear or non-linear transformation.

Principal Component Analysis and Linear Discriminant Analysis[9] are linear transformation techniques.

PCA technique uses orthogonal transformation of features from set of attributes to set of correlated variables. This technique is generally used for unsupervised dataset.

For supervised dataset feature extraction LDA technique is applied[10]. LDA reduces the dimensionality space with good class-separability in order avoid overfitting. This technique preserves class-discriminatory information.

sparse linear discriminant analysis (SLDA)[11] is also proposed in literature. This technique applies LDA to generate modified components with very few original features. Sparseness is included in LDA technique this improves the system efficiency.

Q. Gu, Z. Li and J. Han[12] proposes a technique that simultaneously applies feature selection and subspace learning technique. It uses L2,1-norm on the projection matrix. Using this matrix features are transformed and selected simultaneously.

The two techniques for dimensionality reduction are Feature selection and feature extraction. These two techniques are studied independently. Sreevani[1] proposes a technique that bridges the gap between these two techniques. These techniques are implemented together to generate the combined feature set. This technique uses LDA for two features merging and correlation confident for feature selection. The generated results gives better accuracy than the existing techniques.

III. PROBLEM FORMULATION

Feature selection and extraction are two important aspects in dimensionality reduction. In most of the existing strategies these techniques are studied independently. Combining these two techniques gives better result. The existing work uses LDA for feature extraction to combine 2 features at a time and correlation coefficient selects single attribute form a pair of two attributes. For combining two attributes and for attribute selection a threshold value is required. This Value varies with respect to the dataset. A dataset independent technique isrequired to generate feature subset using feature extraction and selection process.

 DOI : 10.18231/2454-9150.2018.0447

IV. SYSTEM DESCRIPTION

Following figure 1 represents the architecture of the system.

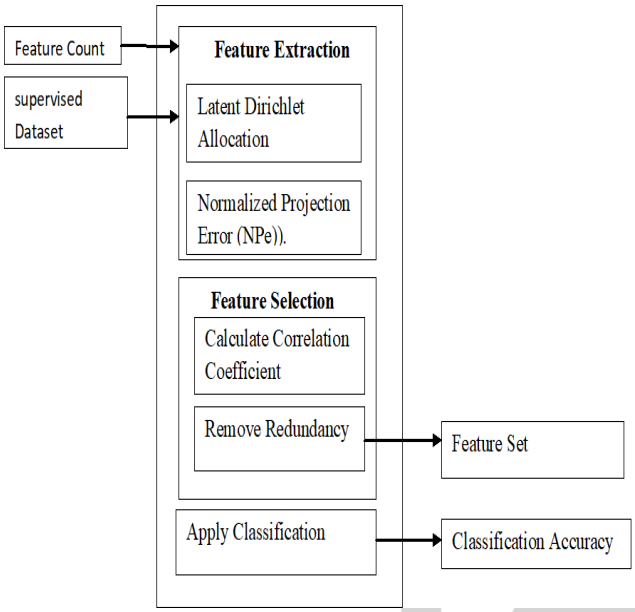

Fig 1:SMPeMR System architecture

The system works on supervised dataset for dimensionality reduction. System uses feature selection and feature extraction strategy. System follows the iterative process to remove redundancy in the dataset.For supervised feature extraction technique system uses LDA. Using LDA two features are combined. The Normalized Projection Error NPe of features is calculated. The compound feature with minimum NPe generates better result. The NPe is calculated as:

$$NPe = \frac{\lambda m}{\sum_{i=1}^{k} \lambda i}$$

Where,

λi : eigen values

λm : a minimum error introduced while projecting the data

System generates combination of pairs for all attributes in dataset and finds NPe for all attributes. The Attribute pair with minimum NPe is selected for feature extraction process.

For feature extraction system uses correlation coefficient. Using correlation coefficient the redundancy in feature set is checked. The redundant attributes are highly correlated. The value of correlation coefficient varied in the range of -1 to 1. If the value of correlation coefficient is -1 or 1 then the features are highly correlated. The correlation checking is the pair-wise comparison of attributes.

The correlation coefficient of 2 attributes X and Y is calculated as:

$$P = \frac{\sum_i (Xi-Xm)(Yi-Ym)}{\sqrt{\sum_i (Xi-Xm)^2}\sqrt{(Yi-Ym)^2}}$$

Where Xi is attribute value

Xm is attribute mean and

N is total number of values in Attribute X

One of the attribute in the highly correlated pair is removed to reduce the redundancy in dataset. The variance of attribute is measured to compare the weak feature. The weak feature has low variance. A low variance attribute is selected from highly correlated pairs and it is removed from the dataset.

The variance V of an attribute X is calculated as:

$$Var = \frac{\sum_i (Xi-Xm)^2}{K-1}$$

Where xi is the value of attribute X of i-th instance

K is total number of instances in Dataset

The correlation coefficient is calculated for all the pairs in the dataset and highly correlated pair is selected for feature selection process.

To select the minimum valued pair for feature extraction and maximum valued pair for feature selection, no dataset specific constants are required for attribute selection process.

The process of feature selection and extraction is executed continuously until the required numbers of features are extracted from the system.

V. ALGORITHMS

Algorithm : Minimum Projection error - Minimum Redundancy (MPeMR) For supervised Dataset

Input: D: Dataset, F: Feature Count
Output: Fset : Feature Set,
Aknn: Classification accuracy-KNN ,
Asvm:Classification accuracy-SVM
Processing:
Fset = features from D
If Fset. Size = Fcount
Apply SVM Classification
Apply KNN classification
Return (fset, aknn,asvm)
Select feature groups (2,2) from dataset Fset
Apply LDA
Calcualte NPe
(Ai,Aj) :Find Minimum NPe pair
Merge features (Ai,Aj)
Update Fset
cc: Calculate : Correlation Coefficient for all pairs
(Ax,Ay): Find maximum cc valued pair
Find variance of attribute (Ax,Ay)
Find minimum variance attribute
Remove feature from Fset
Go to step 2

2. Linear Discriminant Analysis: LDA

Input: Dataset D, Attribute group (d1,d2,..,dk)
Output: Compound feature

 DOI : 10.18231/2454-9150.2018.0447

Processing:
Compute the k-dimensional mean vectors for attribute group from a dataset.
Compute the scatter matrices
Compute the eigenvectors and eigen-values .
Sort the eigenvectors in descending order of eigenvalues
Choose eigenvectors with the largest eigenvalues e
Use this e×k eigenvector matrix to transform the samples onto the new subspace.

VI. IMPLEMENTATION

a) Experimental Setup: The system is implemented in java using jdk1.8. To store dataset weka– arff file format is used. The original dataset is initially converted in arff format for further processing.
b) Datasets: Datasets are downloaded from UCI[13] repository.

Table 1: Dataset Information

Sr. No.	Dataset	Number of Instances	Number of Attributes
1.	Movement Libras	360	91
2.	Yeast	1484	10
3.	Coil 20	5822	86

c)Performance Metric:

Time : The feature selection time is captured to compare 2–feature grouping and 3-feature grouping time efficiency.
Accuracy: The accuracy of the system is evaluated using KNN and SVM classifier.

VII. RESULTS

Accuracy Analysis: The following table 2 shows the accuracy evaluation for 3 different datasets using KNN and SVM classification. KNN1, KNN3 and KNN5 techniques are used for accuracy evaluation. From the accuracy analysis, Proposed system has higher accuracy than the existing system. Constant values T1 and T2 are required for the existing system[1].

Table 1: Accuracy Evaluation

Dataset	Feature count	Technique	T1	T2	Knn1	Knn3	Knn5	SVM
Libras Movement	21	MPeMR [1]	0.7	0.55	59.44	55	61.66	58.05
		MPeMR - Supervised	-	-	84.72	78.33	74.44	66.38
Coil20	130	MPeMR [1]	0.045	0.9	90.08	92.83	93.5	94.02
		MPeMR - Supervised	-	-	90.87	92.97	93.5	94.02
Yeast	7	MPeMR [1]	0.4	0.6	49.79	49.86	53.03	56.67
		MPeMR - Supervised	-	-	52.49	56.26	56.19	56.09

Following Figure 2 represents the graphical representation of accuracy evaluation for KNN classifier for 3 different datasets.

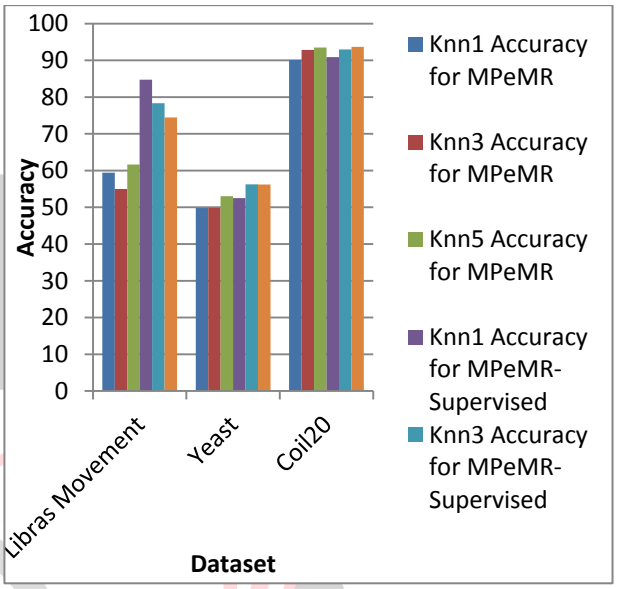

Fig 2: KNN Accuracy Evaluation

Following Fig 3 represents the accuracy evaluation for SVM classifier for 3 dataset.

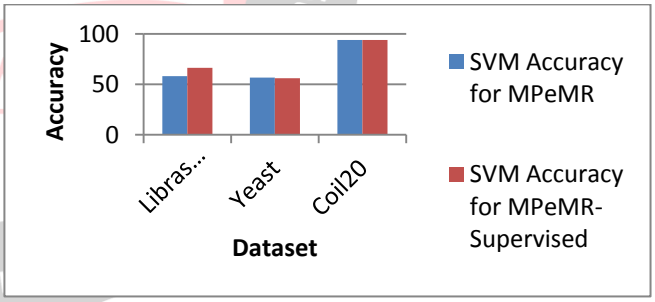

Fig 3: SVM Accuracy Evaluation

b)Analysis: The proposed system finds NPe and CC for each pair of dataset and hence it requires higher time than existing system.

Dataset	MPeMR Time (In sec) [1]	MPeMR-Supervised Time (in Sec)
Libras Movement	1.038	2.584
Yeast	0.262	0.781
Coil20	3.562	6.457

Table 2: Performance Evaluation

 DOI : 10.18231/2454-9150.2018.0447

Following figure 4 represents the time analysis for proposed system with respect to the existing MPeMR[1] System.

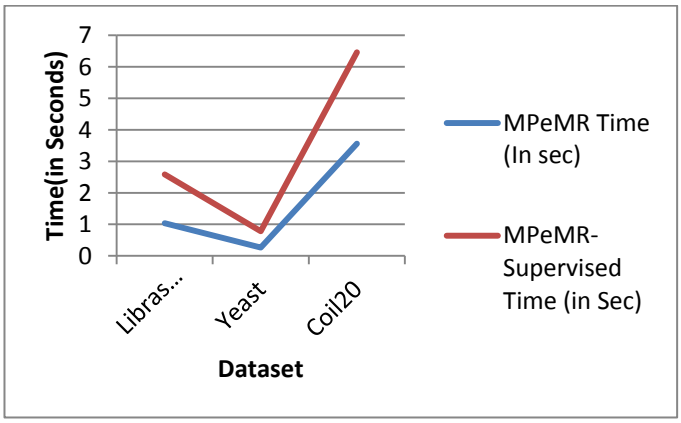

Fig 4: Time Analysis

VIII. CONCLUSION

Feature selection and feature extraction are two important approaches for feature set generation. In existing system these techniques are implemented independently. To achieve better accuracy in feature set generation and to remove redundancy in dataset, the proposed system generates feature set using feature selection and feature extraction technique. The system works with supervised dataset. For feature extraction LDA technique is used. LDA technique merges 2 features from dataset based on the NPe value. For feature selection correlation coefficient technique is used. The NPe and correlation coefficient is calculated for each attribute pair in dataset. To select the minimum valued pair for feature extraction and maximum valued pair for feature selection, no dataset specific constants are required for attribute selection process.

After feature set generation the classification techniques such as KNN and SVM is applied to evaluate the feature set classification accuracy. In future system can be implemented for unsupervised dataset.

REFERENCES

[1] Sreevani and C.A. Murthy, "Bridging Feature Selection and Extraction: Compound Feature Generation," IEEE Transactions on Knowledge and Data Engineering, Vol. 29, No. 4, pp. 757 - 770 , April 2017.

[2] Q. Song, J. Ni, and G. Wang, “A fast clustering-based feature subset selection algorithm for high-dimensional data,” Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 1, pp. 1–14, 2013.

[3] P. Mitra, C. Murthy, and S. K. Pal, “Unsupervised feature selection using feature similarity,” IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 3, pp. 301–312, 2002.

[4] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani et al., “Least angle regression,” The Annals of statistics, vol. 32, no. 2, pp. 407–499, 2004.

[5] J. Liu, S. Ji, and J. Ye, “Multi-task feature learning via efficient l 2, 1-norm minimization,” in Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, 2009, pp. 339–348.

[6] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, “Trace ratio criterion for feature selection.” in AAAI, vol. 2, 2008, pp. 671–676.

[7] L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” in ICML, vol. 3, 2003, pp. 856–863.

[8] M. Robnik-Šikonja and I. Kononenko, “Theoretical and empirical analysis of relieff and rrelieff,” Machine learning, vol. 53, no. 1-2, pp. 23–69, 2003.

[9] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” Journal of computational and graphical statistics, vol. 15, no. 2, pp. 265–286, 2006.

[10] K. Fukunaga, Introduction to statistical pattern recognition. Academic press, 2013.

[11] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, “Sparse discriminant analysis,” Technometrics, 2012.

[12] Q. Gu, Z. Li, and J. Han, “Joint feature selection and subspace learning,” in In Proceedings of the International Joint

[13] http://archive.ics.uci.edu/ml/index.html

 DOI : 10.18231/2454-9150.2018.0447