Recent Trends in Intensive Computing
M. Rajesh et al. (Eds.)
© 2021 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/APC210180

Hybrid Parallel Feature Subset Selection for High Dimensional Datasets

Archana Shivdas Sumant [a, 1], Dr. Dipak V. Patil [b]
[a] *MET's Institue of Engineering Bhujbal Knowledge City, Nashik, India*
[b] *Gokhale Education Society's R.H. Sapat College of Engineering Management Studies and Research Nashik 422005, India*

Abstract. High dimensional data analytics is emerging research field in this digital world. The gene expression microarray data, remote sensor data, medical data, image, video data are some of the examples of high dimensional data. Feature subset selection is challenging task for such data. To achieve diversity and accuracy with high dimensional data is important aspect of this research. To reduce time complexity parallel stepwise feature subset selection approach is adopted for feature subset selection in this paper. Our aim is to reduce time complexity and enhancing the classification accuracy with minimum number of selected feature subset. With this approach 88.18% average accuracy is achieved.

Keywords. High dimensional data, parallel feature subset selection, stepwise selection, symmetric uncertainty, chi-squared.

1. Introduction

Data analytics on high dimensional data is a challenging task. As high dimensional data contains big number of features compared to number of samples in the datasets. If number of features as p and number of samples as n, then p>>n is the high dimensional data. All features are not equally important for extracting meaningful information from such dataset. It increases the time and space complexity as these data contains many redundant and irrelevant features. To avoid this problem ranking methods are used before applying algorithm. But ranking methods have disadvantage that does not considers feature dependency. Search methods plays important role here which selects optimal feature subset by considering feature dependencies.

The way toward distinguishing and evacuating unessential and excess features is known as feature subset selection (FSS). FSS boosts the algorithm to operate much quicker and accurate by reducing the dimensionality of data. FSS in other words known as variable selection, attribute selection or variable subset selection. Feature selection gives many advantages as it enhances expectation execution, understand-ability, versatility, and speculation capacity of the classifiers. It additionally diminishes computational complexity and storage, provides faster and more commercial model.

High Dimensional Data (HDD) poses different challenges on predictive algorithms. Let's say we have *n* samples and *p* features.

[1] Archana Shivdas Sumant, MET's Institue of Engineering Bhujbal Knowledge City, Nashik Email:archana.s.vaidya@gmail.com.

Here features are attributes, independent variables, explanatory variables. High dimensional data is data having n<<p and p are usually high in thousands or ten thousand. So dealing with these numbers of dimensions with high predictive accuracy is the challenge in these coming data era with lots of data generated is of high dimension. Two solutions are their one is dimensionality reduction and the second one is selecting a subset of features. There are several search strategies of different types, but no best algorithm for feature selection is found in general. Prior art [1] compare FS algorithms and conclude that there is not a single approach that outperforms all the others for all datasets. Therefore, it is necessary to continue providing the community with new feature selection alternatives as well as strategies to enhance the performance of the current ones.

Search Strategy in FSS contribute to reduce the time complexity and also to increase the accuracy. In this paper proposed hybrid parallel approach for high dimensional data is implemented.

2. Overview of Feature Subset Selection methods

Feature subset selection problem is stated as given the input data as N samples and M features. The objective of feature selection is to find a subspace of features from the M-dimensional observation space to reduced feature space X, that could be optimally separated the c classes.

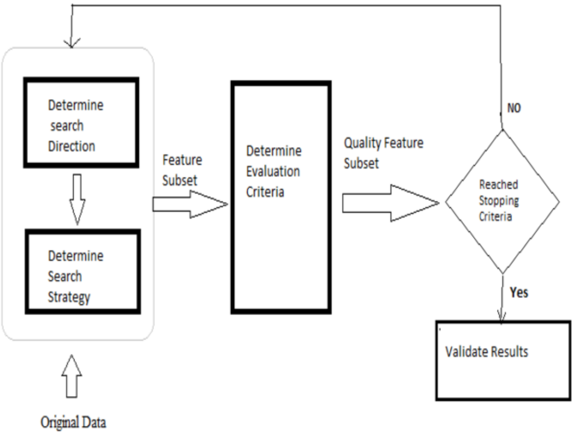
Determine search Direction
Determine Search Strategy
Original Data
Feature Subset
Determine Evaluation Criteria
Quality Feature Subset
NO
Reached Stopping Criteria
Yes
Validate Results

Figure 1. General process of feature subset selection

Figure 1 shows the process of feature subset selection includes search direction, search strategy and evaluation criteria. Feature selection aims to select a feature subset from the original set of variables from its relevance and redundancy.

The system in [2] classify the features into four categories: (i) completely inapt and noisy features, (ii) poorly relevant and surplus features, (iii) weakly relevant and non-redundant features, and (iv) powerfully relevant features. The best subset primarily contains every one of the features from the group (iii) and (iv).

Several main approaches for feature selection are distinguished in [3] as filter, wrapper, Hybrid and embedded methods. In recent years, new techniques are emerging, i.e., ensemble feature selection [4] and deep learning-based feature selection [5][6].

3. Proposed Parallel FSS Approach

Search strategies searches in feature space and selects relevant features by evaluating their performance. The proposed parallel approach is explained in algorithm parallel FSS using stepwise search for high dimensional data. In High dimensional dataset number of features are very large so in the first step features are ranked based on symmetric uncertainty and Chi- square as per equation 1 and 2. Features whose score is zero for these two measures are eliminated and rest of the features are selected for further processing. Symmetric uncertainty (SU) is normalized value measure of Mutual Information (MI) [7] calculated as in eq. (1)

$$SU(X,Y) = \frac{2*MI(X,Y)}{H(x)+H(y)} \tag{1}$$

Value 1 indicates strong relevance between X and Y, while 0 indicates X and Y are independent. SU measure is symmetric in nature therefore SU(X, Y) is same as that of SU(Y, X).

Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class. The initial hypothesis H_0 is the assumption that the two features are unrelated, and it is tested by chi squared formula:

$$\chi 2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-\mathrm{E_{ij}}\right)^2}{\mathrm{Eij}} \tag{2}$$

Eq. (2) gives chi-square score for each feature. O_{ij} is observed value E_{ij} is expected value. It determines significant relationship between two nominal feature vectors. In this feature to predictive class relationship is tested and score is calculated. A Stepwise selection [8] is a combination of forward and backward selections. It starts with no predictors, and then sequentially adds the most contributive predictors like forward selection. After adding each new variable, remove any variables that no longer provide an improvement in the model fit like backward selection.

4. Results Discussion and Implementation Details

The proposed approach is implemented with R 3.6.0 and validation with classifiers was done with Python 3.7.2. System configuration used here was an Intel i7 processor with 8 GB RAM. R language is used for feature selection and Python is used to validate results. Table 1 gives dataset details used in experiment. All datasets are downloaded from [9]. Here n is number of instances, p is number of features and C_k is number of classes. The KalR [10] package from R is used for implementation. Accuracy is used to measure system performance. To calculate accuracy of classifier cross fold validation technique is used. It divides dataset into train and test dataset. Here tenfold cross validation is used. The working of the system is explained with an algorithm 1.

$$\%Accuracy = \left(\frac{\text{Total number of correclty classified instances}}{\text{Total instances}}\right) * 100 \tag{3}$$

Accuracy is a measure of correctly classified instances in total number of instances. Equation (3) gives accuracy in percentage.

Algorithm 1: Parallel FSS using stepwise search for high dimensional data

Input: Dataset with (number of samples n and number of features p)

Output: Validated results with RF, SVM, KNN classifiers.

```
Start:
        for features i   1 to p do
        [scores]   Calculate(Symmetric Uncertainty)  with Equation (1)
        [scores]   Calculate(Chi-Squared score)  with Equation (2)
        end for
        q ← be the number of features with positive score from earlier stage
        Input select ranked subset with top scoring q features
        for features i   1 to q do
            for subset j 1 to m
                  selecting feature i and putting it in subset j
            end for
        end for
        do parallel for each subset stepwise feature selection
   Top selected features from each subset are combined to form final feature
   subset
   for each classifier calculate classification Accuracy based on final selected
   feature subset
        Random Forest(RF)
        Support Vector Machine (SVM)
        K- Nearest Neighbor (KNN)
    end for
  Return Accuracy for feature subset
  End
```

Table 1. Dataset details used in experiment

DN	Dataset Name	n	p	Ck
1	COLON	62	2000	2
2	Lung-cancer	203	12600	5
3	Ovarian	254	15154	2
4	CNS	60	7130	2
5	Leukemia	72	7129	2
6	Prostate	102	12600	2
7	DLBCL	47	4026	2

The description of the classifiers used is as follows.

Random Forest (RF) is an ensemble approach based on decision trees. Support Vector Machine (SVM) generates hyperplanes to separate samples belonging to different classes. For multi-class problems it converts problem as one versus rest i.e. dividing problem into multiple binary class problems. Here linear kernel is used for generating hyperplanes. K Nearest Neighbor (KNN) works based on proximity of test samples with neighbors instances. It is a supervised learning algorithm and works with neighbor samples. Here neighbor count is set to 5 and brute-force search algorithm is used.

Table 2. Proposed parallel stepwise search accuracy measured on high dimensional datasets

DN	1	2	3	4	5	6	7	Classifier wise Average
Dataset Name	COLON	Lung-cancer	Ovarian	CNS	Leukemia	PROSTATE	DLBCL	
SVM	94	88	99	74	90	88	90	89
RF	93	99	99	43	92	90	75	84.42
KNN	94	99	99	74	96	87	89	91.14
Average of All three Classifiers	93	95.33	99	63.66	92.66	88.33	84.66	**88.18**

Table 2 states accuracy measured with proposed method with parallel stepwise search. The highest performance has been observed on ovarian dataset as 99%. And the lowest performance has been observed for CNS dataset. The KNN classifier achieves the highest average accuracy as 91.14% while RF classifier achieves lowest average accuracy as 84.42 %. The average accuracy achieved is 88.18%. The time complexity analysis shows parallel approach is three times faster than sequential approach.

5. Conclusion

Feature subset selection is NP -hard problem and not a single solution generalizes the system for classification. So, there is need of solution to improve performance of existing system. Feature subset selection plays an important role in case of high dimensional dataset. The proposed parallel stepwise feature selection achieves average accuracy as 88.18%. In future ensemble approach can be adopted to increase system accuracy.

References

[1]. Guyon, A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.* **3**, 1157–1182 (2003)

[2]. Geem Z.W., " State-of-the-Art in the Structure of Harmony Search Algorithm," In: Geem Z.W. (eds) Recent Advances In Harmony Search Algorithm. Studies in Computational Intelligence, vol 270 2010, https://doi.org/10.1007/978-3-642-04317-8_1.

[3]. K. S. Lee and Z. W. Geem, A new meta-heuristic algorithm for continuous engineering optimization: Harmony search theory and practice, *Comput. Meth. Appl. Mech. Eng.*, vol. 194, nos. 36–38, pp. 3902–3933, Sep. 2005

[4]. M. B. Dowlatshahi, V. Derhami, and H. Nezamabadi-Pour, "Ensemble of filter-based rankers to guide an epsilon-greedy swarm optimizer for high-dimensional feature subset selection," Inf., vol. 8, no. 4, 2017, doi: 10.3390/info8040152

[5]. Yawen Xiao a , Jun Wu b , Zongli Lin c , Xiaodong Zhao b "A deep learning-based multi-model ensemble method for cancer prediction ",dx.doi.org/10.1016/j.cmpb.2017.09.005

[6]. Van-Sang Ha , Ha-Nam Nguyen, "Credit scoring with a feature selection approach based deep learning," DOI:10.1051 2016.

[7]. Hoque, N., Singh, M., & Bhattacharyya, D. K. (2017). EFS-MI: an ensemble feature selection method for classification. *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747-017-0060-x

[8]. Geem Z.W. (2010) State-of-the-Art in the Structure of Harmony Search Algorithm. In: Geem Z.W. (eds) Recent Advances In Harmony Search Algorithm. Studies in Computational Intelligence, vol 270. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04317-8_1

[9]. https://archive.ics.uci.edu/ml/index.php

[10]. https://rdrr.io/rforge/klaR/man/stepclass.html