Noname manuscript No.
(will be inserted by the editor)

Graph-based supervised feature selection using correlation exponential

Gulshan Kumar · Gitesh Jain · Mrityunjoy Panday · Amit Kumar Das · Saptarsi Goswami

the date of receipt and acceptance should be inserted later

Abstract In this article, a graph-theoretic approach for supervised feature selection using matrix exponential of pairwise correlation value, has been illustrated. In machine learning, high dimensional data sets have a large number of irrelevant and redundant features. The sum of mean and standard deviation of exponential matrix has been set as the threshold for selecting relevant features. Principles of vertex cover and independent set have then been used to remove redundant features. In the next step, mutual information value has been used to select relevant features that were initially rejected. The results show that this approach has performed better than the benchmark algorithms when experimented on multiple standard data sets.

Keywords Feature selection · Feature redundancy · Graph-based visualization · Correlation exponentiation

Gulshan Kumar
Institute of Engineering and Management,
Kolkata, India
E-mail: gulshankumar454@gmail.com

Gitesh Jain
Institute of Engineering and Management,
Kolkata, India
E-mail: giteshjain844@gmail.com

Mrityunjoy Panday
E-mail: mrityunjoy@gmail.com

Amit Kumar Das
Institute of Engineering and Management,
Kolkata, India
E-mail: amitkumar.das@iemcal.com

Saptarsi Goswami
University of Calcutta,
Kolkata, India
E-mail: saptarsi007@gmail.com

1 Introduction

In the last few years, there has been a rapid increase in the number of high-dimensional data sets i.e data sets with a large number of features, generated from different domains like computational biology, geospatial technologies, text and social data mining, etc. The data from these domains are very large in size and have features in the range of thousands [3].

High-dimensional data sets are computationally expensive with respect to both space and time. To perform any task in finite time and limited storage space, we need optimization to reduce the number of features of the data set. All the features of a data set are not relevant [5]. Also, lot of features contribute the same information in predicting the results. These redundant features increase computational cost. The set of redundant features should be determined to reduce the overall size of the data sets. Few representative features should be selected from a set of redundant features. Hence, selecting a subset of features maximize utilization of resources and minimize execution time.

Feature selection is a combinatorial optimization problem [9] which aims at optimizing the number of features in the data sets without significant impact on output. An alternative is exhaustive search and selection of optimal feature subset, which is known to be NP-complete. During feature selection process, the complexity of the algorithm should also be taken care off. Optimizing feature subset has been found as NP-complete [1][2]. Using approximation algorithm instead of exhaustive search reduces the order of the search space to $O(N^2)$ or below [8].

2 Related Work

Researchers have adopted several techniques to solve the feature selection problem with a significant number of them using graph-theoretic approach. A survey paper on graph clustering [15] mentions the possibilities to convert feature vector data into graph format. Another work has implemented the concept of dominant set to do feature selection [17]. In the algorithm DSFFC [19], dense sub-graph and mutual information have been used to select optimal features. A minimum spanning tree (MST) based clustering algorithm was used to design an algorithm, named FAST (Fast clustering-based feature Selection algorithm)[16]. In Graph Cluster with Node Centrality (GCNC) algorithm [13], a graph-theoretic approach of feature selection using community detection algorithm was implemented.

Schaeffer[15] mentioned about the conversion of feature vector of a dataset into a graph with features representing vertices and similarity representing edges. In the approach used by Zhang and Hancock [17], first relevance matrix has been calculated using inter-feature mutual information, then dominant set clustering is done to cluster the feature vectors and lastly optimal feature subset is selected from each dominant set using multidimensional interaction information criterion. This work has been taken as a benchmark graph-based supervised feature selection algorithm to compare with the proposed algorithm. Another algorithm Ref39 used the variant of normalized mutual information (NMI) to obtain a subgraph with a minimum number of features. A two-step algorithm, FAST [16], in the first step, clusters the features into groups relatively independent of each other. In the second step, it selects the features most strongly representing each of the clusters.

Most of work on filter based feature selection algorithms use pairwise metric e.g. variations of correlation coefficient or mutual information, for computing redundancy. However, in this paper, we compute the expected value by using sum over all paths and then use minimum vertex cover and maximum independent set to identify a subset of the potentially redundant features. It also uses a local property of redundancy to derive a property based on global relationship structure.

The remaining part of the paper is organized as follows:

- Section 3 discusses some preliminary concepts.
- The proposed approach is discussed in section 4.
- Methods and materials used are highlighted in section 5 which includes CFS, mRMR, and DSCA as benchmarks.
- Finally, results and analysis are presented in section 6.

3 Preliminary Concept

This section describes some fundamental concepts and information used.

3.1 Correlation

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ and may be referred to as the correlation coefficient or the Pearson correlation coefficient. The formula is:

$$\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}.$$

3.2 Expected Correlation

The Expectation value is calculated as the integral or weighted sum over all values. The Equivalent notion in graph is sum over all paths, which can be done using Matrix Exponential which converges due to normalization using Taylor Series expansion of e^A.

3.3 Mutual Information

Mutual information I(X; Y) between two random variables measures how much information about one can be extracted through the knowledge of the other [21]. Mathematically, we define this as the difference between the entropy of X and the entropy of X conditioned on knowing Y.

$$I(X;Y) = H(x) - H(X|Y)$$

3.4 Vertex Cover and Independent Set

Minimal vertex cover and maximal independent set are graph-theoretic principles to derive sub-graph of a graph. Hence, it can be used for identification of representative subset of features from a full feature set. Vertex cover of a graph is a set of vertices such that each edge of the graph is incident to at least one vertex of the set. Finding minimum vertex cover, i.e. the vertex cover with lowest number of vertices, is a NP-complete problem. However, there are algorithms to find a minimal vertex cover in polynomial time. An independent set is a set of vertices in a graph such that no two vertices belonging to the set are adjacent. Finding maximum independent set problem is an NP-hard optimization problem. However, a list of all maximal independent sets can be derived by algorithms in polynomial time.

4 Proposed Approach

In the proposed approach correlation has been used a global property instead of a local one. We attempt to represent each data set in consideration as an undirected graph. This has been done in following steps.

- Step 1 : Constructing base graph
  The vertices of the graph represent the features of the data set. The similarity between two vertices is represented by the edge between them. We find out the correlation matrix based on the pair-wise similarity between different features using Pearsons product-moment correlation coefficient. If correlation value is high, features are more similar.
- Step 2 : Exponential of base matrix
  The exponential of correlation matrix converts the local property of correlation between edges into global property. The sum of mean and standard deviation of the exponential matrix gives us a dynamic threshold. Features having correlation value higher than the threshold value form the feature graph. So we have a subset of the features which are having a high probability of redundancy and thus can be eliminated resulting to a comparatively lesser number of vertices.
- Step 3 : Identify representative features
  The principles of minimal vertex cover and maximal independent sets have been used to find out reduced number of vertices. If two vertices are similar, they have an edge between them in the feature association graph, either using minimal vertex cover or maximal independent set, we can take any one of those two vertices having higher degree of correlation. The selected vertex will act as the representative of its adjacent vertices. Thus from a group of similar vertices, we select a few and eliminate rest of the vertices which are the neighbours of the selected vertex. In this way, using minimal vertex cover or maximal independent set, we will get a sub-graph of original feature association graph having lesser number of features with least redundancy.
- Step 4 : Re-select relevant features
  It may happen that a feature is not related to any other feature. So it is not present in the Feature Graph because of low value of correlation coefficient with other features but contributes a lot in predicting the actual result. Thus we

can't ignore those features as they play a vital role in predicting the final output value. The property of Mutual Information has been used for this purpose. Suppose we have "m" features which are not the part of feature graph. The mutual information of these features with class labels have been found. The features has been ranked based on the calculated Mutual Information value and some percentage of these m features are selected from top as per as their ranking. In this way those relevant features which were initially ignored also gets selected.

The two subsets generated in steps 3 and 4 are combined and we get a feature set having lesser redundancy and also more relevant features, which is the desired goal of feature selection.

Algorithm : Feature Selection based on Correlation Exponential (FSCE)

Input: N-dimensional data set D_N, having original feature set O = f_1, f_2, ..., f_N.
Output: S_{OPT1}, S_{OPT2}(Optimal Feature Subsets)

Begin
// Part 1 : Computing the similarity matrix and drawing FSCE
1: $M_{corr} \leftarrow |(correlation(DN1))|$ //M_{corr} = correlation matrix
2: $M_{exp} \leftarrow$ expm(Mcorr) //M_{exp} = exponential of correlation matrix
3: Threshold = mean (M_{exp}) + sd (M_{exp}) // sd = standard deviation
4: For each M_{corr}[col-count, row-count]
5: If col-count == row-count
6: Set M_{corr}[col-count, row-count] = 0
7: If M_{corr}[col-count, row-count]>= Threshold
8: Set M_{corr}[col-count, row-count] = 1
9: Else
10: Set M_{corr}[col-count, row-count] = 0
11: End If
12: Next
13: $G_{FSCE} \leftarrow$ adjacency-graph (M_{corr})
14: $G_{NFSCE} \leftarrow$ column(D) vertex (G_{FSCE})
// Part 2 : Deriving the sub-graph of FSCE
15: $V_{VERTCOV} \leftarrow$ Vert-cover(G_{FSCE})
16: $V_{INDSET} \leftarrow$ Ind-set(G_{FSCE})
17: $S_{VC} \leftarrow$ D[$V_{VERTCOV}$]
18: $S_{IS} \leftarrow$ D[V_{INDSET}]
// Part 3 : Selecting the subset for non-FSCE features
19: $D_{NFSCE} \leftarrow$ D[vertex(G_{NFSCE})])
20: For each field in D_{NFSCE}
21: $F_{MI} \leftarrow$ Mutual-Info(field, class)
22: Rank (F_{MI})
23: $S_{NFSCE} \leftarrow$ top (β * count(F_{MI}) features)
24: Next
// Part 4 : Selecting the final feature subset
25. $S_{OPT1} \leftarrow S_{VC} + S_{NFSCE}$

26. $S_{OPT2} \leftarrow S_{IS} + S_{NFSCE}$
End

5 Methods and materials

5.1 Data sets used

For conducting the experiments, 16 standard datasets, details provided in table 1, from the machine learning repository of UCI (University of California, Irvine) [24] have been used. Out of the 16 datasets, 14 datasets have number of features less than 100 while the remaining 2 datasets have number of features greater than 100. This allows to have an understanding of the relative performance of the algorithms in low as well as high dimensional datasets. The datasets have been split in 70:30 ratio between training and test data respectively. The computing environment R has been used for the experiments. Many R libraries have also been used. All the datasets that have been used have class information which allow validation of classification results.

5.2 Competing Benchmark algorithms

The proposed FSCE algorithm has been compared with three benchmark algorithms for supervised feature selection. These algorithms are:

- Correlation-based feature selection (CFS)[12], a correlation based filter method which identifies relevant features as well as redundancy among relevant features
- Minimum Redundancy Maximum Relevance (mRMR) [23], uses mutual information values between individual feature and class as the primary parameter to identify feature subset having minimal redundancy and maximal relevance
- Dominant set clustering algorithm (DSCA) [17], based on graph-theoretic approach of dominant set clustering which also uses mutual information between features in the form of multi-dimensional interaction information as the feature selection criterion

5.3 Evaluation Criteria

The experiments have been done to evaluate the performance of the proposed algorithm compared to the benchmark algorithms with respect it is a measure for the efficiency of any algorithm. For accuracy, the measurement has been done as follows: Accuracy = (TP + TN)/(TP+TN+FP+FN) , where TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives respectively. For measuring the classification accuracy with the derived feature subset from each algorithm, Decision Trees have been used. Decision tree is a basic classification model which can be used to test the efficacy of any feature selection algorithm.

Data set	# of Features	# of classes	# of Instances
Apndcts	7	2	106
Btissue	9	6	106
Cleaveland	13	5	297
CTG	34	10	2126
Ecoli	7	8	336
Glass	9	6	214
ILPD	10	2	579
Madelon	500	2	2000
Mfeat	649	10	2000
Pgblk	10	5	5473
Sonar	60	2	208
Texture	40	11	5500
Vehicle	18	4	846
Wbdc	30	2	569
Wine	13	3	178
Wiscon	9	2	682

Table 1: UCI data sets analysed

6 Results and analysis

The experiments are conducted with $\beta = 0.1, 0.2, 0.5$ and 0.8 respectively. It was observed that at β =0.8 the performance is quite comparable with that obtained using state-of-the-art algorithms.

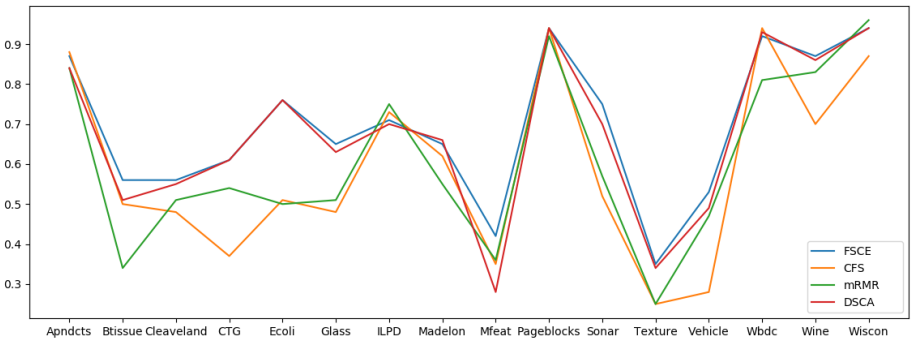

Fig. 1: Comparison of performance of algorithms for different data sets

In table 2, the best accuracy values achieved from FSCE at β =0.8 along with the accuracy values using CFS, mRMR and DSCA have been presented. For FSCE minimal vertex cover (MVC) and the other using maximal independent set (MIS), the approach resulting in the winning result has been captured in the table 2.

It is observed, from table 2 that:

- For 10 out of 16 datasets, FSCE is better in performance than the 3 benchmark algorithms CFS, mRMR and DSCA.

Dataset	FSCE	Winning approach	CFS	mRMR	DSCA
Apndcts	0.85	MVC	0.88	0.84	0.84
Btissue	0.52	MVC	0.5	0.34	0.51
Cleaveland	0.56	MIS	0.48	0.51	0.55
CTG	0.62	Both	0.37	0.54	0.61
Ecoli	0.77	MIS	0.51	0.5	0.76
Glass	0.65	MVC	0.48	0.51	0.63
ILPD	0.71	MIS	0.73	0.75	0.7
Madelon	0.65	MVC	0.62	0.55	0.66
Mfeat	0.38	MVC	0.35	0.36	0.28
Pgblk	0.95	MIS	0.94	0.92	0.94
Sonar	0.71	MIS	0.52	0.57	0.7
Texture	0.33	MIS	0.25	0.25	0.34
Vehicle	0.54	MIS	0.28	0.47	0.49
Wbdc	0.93	MVC	0.94	0.81	0.93
Wine	0.87	MIS	0.7	0.83	0.86
Wisconsin	0.95	MIS	0.87	0.96	0.94
Mean Accuracy	0.69	-	0.59	0.61	0.68

Table 2: Comparison of Accuracy Values

- Out of the remaining, in 5 datasets, the performance of FSCE is better than at least two of all features and the benchmark algorithms - CFS, mRMR and DSCA.
- Even in case of the remaining 1 dataset, the performance of FSCE is better than at least one of the benchmark algorithms - CFS, mRMR and DSCA.
- Mean accuracy over all datasets obtained from FSCE is better than that derived from the benchmark algorithms - CFS, mRMR and DSCA.
- In 6 out of 16 datasets, MVC is the winning approach while for 9 datasets MIS is the winning approach and 1 has the tie.

7 Conclusion

In this paper instead of using pairwise correlation, sum over all paths is used for supervised feature selection, using matrix exponential, along with data driven thresholds. We found that this works better than benchmarks on certain data sets, showing the value of the expected correlation value, which is also intuitively better than using a simple pairwise value. Future work will explore the optimal threshold for feature selection.

References

1. Koch PN, Simpson TW, Allen JK, Mistree F (1999) Statistical approximations for multidisciplinary design optimization: the problem of size. J Aircr 36(1):275286 (1999)
2. Li G, Wang S-W, Rosenthal C, Rabitz H, High dimensional model representations generated from low dimensional data samples. I. mp-Cut-HDMR. J Math Chem 30(1):130 (2001b)
3. Dougherty, E.R., Hua, J., Tembe, W., Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition, 42, pp. 409-424 (2009)
4. Das, A. K., Goswami, S., Chakraborty, B., & Chakrabarti, A. (2016). A graph-theoretic approach for visualization of data set feature association. In ACSS.
5. Dash, M., & Liu, H. (1997). Feature selection for classification. Intell. Data Anal., 1, 131156.
6. Dash, M., & Liu, H. (2000). Feature selection for clustering. In PAKDD.

7. Deo, N. (1979). Graph Theory with Applications to Engineering and Computer Science. PHI.
8. John, G.H., Kohavi, R., Peger, K., Irrelevant Features and the Subset Selection Problem, ICML (1994)
9. John, G.H., Kohavi, R., Wrappers for Feature Subset Selection. Artif. Intell., 97, 273 - 324 (1997)
10. Goswami, S., Chakrabarti, A., & Chakraborty, B. (). An efficient feature selection technique for clustering based on a new measure of feature importance. Journal of Intelligent & Fuzzy Systems, (pp. 112).
11. Goswami, S., Das, A. K., Chakrabarti, A., & Chakraborty, B. (2017). A feature cluster taxonomy based feature selection technique. Expert Systems with Applications, 79, 7689.
12. Mark, A. Hall,Correlation-based Feature Selection for Machine Learning
13. Moradi, P., & Rostami, M. (2015a). A graph theoretic approach for unsupervised feature selection. Eng. Appl. of AI, 44, 3345.
14. Moradi, P., & Rostami, M. (2015b). Integration of graph clustering with ant colony optimization for feature selection. Knowl.-Based Syst., 84, 144161.
15. Schaeffer, S. E. (2007). Graph clustering. Computer Science Review, 1, 2764.
16. Song, Q., Ni, J., & Wang, G. (2013). A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Trans. Knowl. Data Eng., 25, 114.
17. Zhang, Z., & Hancock, E. R. (2011). A graph-based approach to feature selection. In GBRPR.
18. Zhang, Z., & Hancock, E. R. (2012). Hypergraph based information-theoretic feature selection. Pattern Recognition Letters, 33, 19911999.
19. Bandyopadhyay, S., Bhadra, T., Mitra, P., & Maulik, U. (2014). Integration of dense subgraph nding with feature clustering for unsupervised feature selection. Pattern Recognition Letters, 40, 104112.
20. Strehl, A., Ghosh, J., 2002. Cluster ensembles a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583617.
21. Cover, T.M., Thomas, J.A., 2012. Elements of Information Theory. John Wiley & Sons, New York, USA.
22. Estrada, E., & Rodrguez-Velzquez, J.A. (2005). Subgraph centrality in complex networks. Physical review. E, Statistical, nonlinear, and soft matter physics, 71 5 Pt 2, 056103.
23. Peng, H., Long, F., & Ding, C.H. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 1226-1238.
24. Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml.