INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

EFFICIENT UNSUPERVISED AND SUPERVISED LEARNING ALGORITHM USING LOCAL LINEAR EMBEDDING SYSTEM

Vinita Vaishnav[1]*, Ravi Kateeyare[2], Jitendra Singh Chouhan[3], Babita Dehriya[4], Kishalay Vyas[5]
Department of Electrical Engineering
[1,3]Central India Institute of Technology, Indore
[2, 4 ,5] RKDF School of Engineering, Indore

ABSTRACT

Many fields of science depend on exploratory data analysis and data redundancy. The demand to analyze large amounts of multivariate data raises the central problem of dimensionality reduction. The Spectral feature selection identifies relevant features by measuring their capacity of preserving sample similarity. It offers a potent framework for both supervised and unsupervised feature selection, it has been shown to be in force in many real-world applications. The redundant features can receive substantial adverse effects on learning performance, it is necessary to address this limitation for spectral feature selected. In this paper I aim to carry out the spectral feature selection algorithm to handle feature redundancy, taking a hybrid model. In this report to conduct theoretical analysis of the attributes of its optimal solutions, develop a correlation-based method for relevance and removing redundancy analysis, and carry on an empirical survey of its efficiency and effectiveness comparing with representative methods.

Keywords: Unsupervised learning and supervised learning, feature selection, relevance, redundancy, high dimensionality, clustering, feature selection.

1. INTRODUCTION

Managing high-dimensional data represent one of the most challenging problems for learning. Passed on the vast number of features, learning algorithms can over fit data and become less comprehensible. Feature selection is one effective way to reduce dimensionality by removing irrelevant and redundant features. Now day's researchers designed spectral feature selection algorithms to identify relevant features through evaluating features capability on preserving sample similarity. Given m features, and a similarity matrix S of the samples, the estimate of spectral feature selection is to pick out characteristics that adjust well with the leading eigenvectors of S. Since the leading eigenvectors of S contain structure information of sample distribution and group similar samples into compact clusters features aligning better to them will have stronger capability on preserving sample similarity.

These algorithms demonstrated excellent performance in both supervised and unsupervised learning. However, since the algorithms evaluate features individually, they cannot handle redundant features. Redundant features increase dimensionality unnecessarily and worsen learning performance when facing shortage of data. It is also shown empirically that removing redundant features can result in significant performance improvement

In this study, we address the limitation of existing spectral feature selection algorithms in handling redundant features and offer a novel spectral feature selection algorithm of an embedded model, which values the utility of a set of features jointly and can efficiently remove redundant features. The algorithm is derived from a formulation based on multi-output regression and feature selection is accomplished by applying L2, 1-norm constraint on the answers.We examine its capability on redundancy removal and study the properties of its optimal solutions, which paves the path for an efficient path following solver. By exploiting the necessary and sufficient conditions for the optimal solutions, our solver can automatically adjust its parameters to generate a solution path for selecting a specific number of features efficiently. We conduct extensive empirical study on the proposed algorithm in both supervised and unsupervised learning to demonstrate that it can select relevant features with low redundancy.

In the classic supervised learning process, a training set of labeled fixed-length feature instances. An instance is typically described as an assignment of values f = (f1, ..., fN) to a set of features F = (F1, ...,FN) and one of one possible classes c1, ...,cl to the class label C. The task is to induce a hypothesis that accurately predicts the labels of novel instances. The learning of the classifier is inherently determined by the feature-values. In theoretically more features should provide more discriminating power, but in practice, with a limited amount of training data, excessive features will not only significantly slow down the learning process, but also cause the classifier to over-fit the training data as irrelevant or redundant features may confuse the learning algorithm.

Let G be some subset of F and fG be the value vector of G. In general, the goal of feature selection can be formalized as selecting a minimum subset G such that P(C | G= fG) is equal or as close as possible to P(C | F = f), where P(C | G = fG) is the probability distribution of different classes given the feature values in G and P(C | F = f) is the original distribution given the feature values in F.

Optimal subset: Let features F1, ...,F5 be Boolean. The target concept isC=g(F1, F2) where g is a Boolean function. With F2 = F3 and F4 = F5, there are only eight possible instances.In order to determine the target concept, F1 is indispensable; one of F2 and F3 can be disposed of (note that C can also be determined by g(F1, F3)), but we must have one of them; both F4 and F5 can be discarded. Either {F1, F2} or {F1, F3} is an optimal subset. The goal of feature selection is to find either of them.

2. FEATURE RELEVANCE AND FEATURE REDUNDANCY

2.1 Feature Relevance

Based on a review of previous definitions of feature relevance, John, Kohavi, and Pfleger classified features into three disjoint categories, namely, strongly relevant, weakly relevant, and irrelevant features . Let F be a full set of features, Fi a feature, and Si = F −{Fi}.

2.2 Feature Redundancy

The feature redundancies are normally in terms of feature correlation. It is widely admitted that two features are redundant to each other if their values are completely correlated. In reality, it may not be so straightforward to determine feature redundancy when a feature is correlated with a lot of characteristics. We now formally define feature redundancy in order to formulate an approach to explicitly name and get rid of redundant features.

Definition (Markov blanket) Given a feature Fi, let Mi _ F (Fi /2Mi), Mi is said to be a Markov blanket for Fi iff P(F −Mi−{Fi}, C | Fi, Mi) = P(F −Mi−{Fi}, C | Mi) .

The Markov blanket condition requires that I subsume not only the information that Fi has about C, but also almost all of the other characteristics. It is pointed out in that an optimal subset can be held by a backward elimination procedure, known as Markov blanket filtering: let G be the current set of features (G = F in the beginning), at any stage, if there exists a Markov blanket for fit within the current G, Fi is removed from G. It is proved that this process guarantees a feature removed in an earlier phase will still find a Markov blanket in any later phase, that is, removing a feature in a later phase will not render the previously removed features necessary to be included in the optimal subset. According to previous definitions of feature relevance, we can also prove that strongly relevant features cannot find any Markov blanket. Since irrelevant features should be removed anyway, we exclude them from our definition of redundant features.

Definition (Redundant feature) Let G be the current set of features, a feature is superfluous and therefore should be removed from G if it is weakly relevant and takes in a Markov blanket me within G.From the property of Markov blanket, it is easy to understand that a redundant feature removed earlier remains redundant when other features are removed. Figure 1 depicts the relationships between definitions of feature relevance and redundancy introduced so far. It shows that an entire feature set can be conceptually divided into four basic disjoint parts: irrelevant features (I), redundant features, weakly relevant but non-redundant features (III), and strongly relevant features (IV). An optimal subset essentially contains all the features in parts III and IV. It is worthy to point out that although parts II and III are disjoint, different partitions of them can result from the process of Markov blanket filtering.

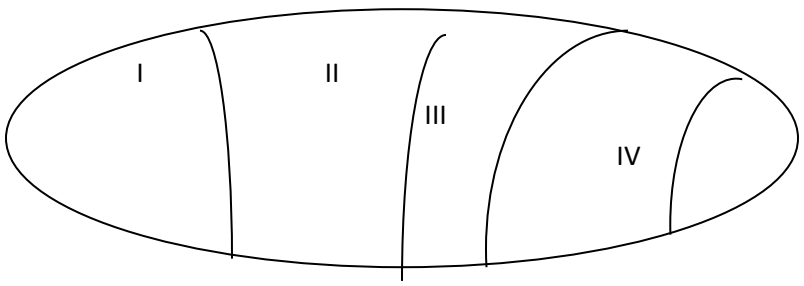

Figure 1: A feature relevance and redundancy.

I: Irrelevant features II: Weakly relevant and redundant features
II: Weakly relevant and non redundant features
IV: Strongly relevant features
III + IV: Optimal subset

3. EFFICIENT FEATURE SELECTION VIA RELEVANCE AND REDUNDANCY ANALYSIS

We now review two major advances in dealing with feature relevance and redundancy, analyze their limitations for high-dimensional data, and then propose a fresh framework of efficient feature selection based on relevance and redundancy analysis.

3.1 Existing Approaches to Dealing with Relevance and Redundancy

As noted before, there survive two major approaches in feature selection: individual evaluation and subset evaluation. Individualized evaluation, too known as feature weighting/ranking assesses individual features and assigns them weights according to their levels of relevance. A subset of features is often chosen from the crest of a ranking list, which approximates the set of relevant features (II, III, and IV in Figure 1). With its linear time complexity in terms of dimensionality N, this access is efficient for high-dimensional information. However, it is incapable of removing redundant features because redundant features likely have similar rankings. As long as features are deemed relevant to the class, they will all be selected even though many of them are highly correlated to each other. For high-dimensional data which may contain a large number of redundant features, this approach may produce results far from optimal.

Many feature selection methods take the subset evaluation approach which handles feature redundancy with feature relevance. The Subset generation produces candidate feature subsets based on a certain search strategy. Although there exist various heuristic search strategies such as greedy sequential search, best-first search, and genetic algorithm most of them still incur time complexity O(N2), which prevents them from scaling well to data sets containing tens of thousands of features.

3.2 A New Framework of Efficient Feature Selection

The state-of feature selection methods have to rely on the approach of subset evaluation which implicitly handles feature redundancy with feature relevance. These methods can create more serious results than methods without handling feature redundancy, only the high computational cost of the subset search makes them ineffective for high-dimensional information. Therefore, in our solution, we propose a new framework of feature selection which avoids implicitly handling feature redundancy and turns to efficient elimination of redundant features via explicitly handling feature redundancy.

Relevance definitions divide features into strongly relevant, weakly relevant, and irrelevant ones,redundancy definition further divides weakly relevant features into redundant and non-redundant ones. Our goal is to efficiently find the optimal subset (parts III and IV in Figure 1). We can achieve this goal through a new framework of feature selection composed of two steps:first, relevance analysis determines the subset of relevant features by removing irrelevant ones, and second, redundancy analysis determines and eliminates redundant features from relevant ones and thus produces the final subset. Its advantage over the traditional framework of subset evaluation lies in that by decoupling relevance and redundancy analysis, it circumvents subset search and allows a both efficient and effective way in finding a subset that approximates an optimal subset.

4. EFFICIENT RELEVANCE AND REDUNDANCY ANALYSIS

Using symmetrical uncertainty (SU) as the correlation measure, we are ready to develop an approximation method for both relevance and redundancy analysis.

Definition (C-correlation) The correlation between any feature Fi and the class C is called Ccorrelation, denoted by SUi,c .

Definition (F-correlation) The correlation between any pair of features Fi and Fj (i 6= j) is called F-correlation, denoted by SUi, j .

Definition (ApproximateMarkov blanket) For two relevant features Fi and Fj (i 6= j), Fj forms an approximate Markov blanket for Fi iff SUj,c _ SUi,c and SUi, j _ SUi,c .

The Markov blanket filtering, a backward elimination procedure based on a feature's Markov blanket in the current set, guarantees that a redundant feature removed in an earlier phase will still find a Markov blanket in any later phase when another redundant feature is removed. It is easy to verify that this is not the case for backward elimination based on a feature's approximate

Markov blanket in the current set. For instance, if Fj is the only feature that forms an approximate Markov blanket for Fi, and Fk forms an approximate Markov blanket for Fj, after removing Fi based on Fj, further removing Fj based on Fk will result in no approximate Markov blanket for Fi in the current set. However, we can avoid this situation by removing a feature only when it can find an approximate Markov blanket formed by a predominant feature, defined as follows.

Definition (Predominant feature) A relevant feature is predominant iff it does not have any approximate Markov blanket in the current set. Predominant features will not be removed at any stage. If a feature Fi is removed based on a predominant feature Fj in an earlier phase, it is guaranteed that it will still find an approximate Markov blanket (the same Fj) in any later phase when another feature is removed. To summarize, our method for redundancy analysis consists of (1) selecting a predominant feature, (2) removing all features for which it forms an approximate Markov blanket, and (3) iterating steps (1) and (2) until no more predominate features can be selected. An optimal subset can therefore be approximated by a set of predominant features.

4.1 FCBF ALGORITHM AND ANALYSIS

The approximation method for relevance and redundancy analysis presented before can be realized by an algorithm, named FCBF (Fast Correlation-Based Filter). It involves two connected steps.

(1) selecting a subset of relevant features

(2) selecting predominant features from relevant ones. for a data set S with N features and class C, the algorithm finds a set of predominant features Sbest . In the first step (lines 2-7), it calculates the SU value for each feature, selects relevant features into S0 list based on a predefined threshold d, and orders them in a descending order according to their SU values. In the second step (lines 8-18), it further processes the ordered list S0 list to select predominant features. A feature Fj that has already been determined to be a predominant feature can always be used to filter out other features for which Fj forms an approximate Markov blanket. Since the feature with the highest C-correlation does not have any approximate Markov blanket, it must be one of the predominant features. So the iteration starts from the first element in S0 list (line 8) and continues as follows. For all the remaining features (from the one right next to Fj to the last one in S0 list), if Fj happens to form an approximate Markov blanket

```
Input: S(S(F1,F2, ...,FN,C) // a training data set
       d // a predefined threshold
Output: Sbest // a selected subset
1 begin
2 for i = 1 to N do begin
3 calculate SUi,c for Fi;
4 if (SUi,c > d)
5 append Fi to S0
list ;
6 end;
7 order S0
list in descending SUi,c value;
8 Fj = getFirstElement(S0
list);
9 do begin
10 Fi = getNextElement(S0
list ,Fj);
11 if (Fi <> NULL)
12 do begin
13 if (SUi, j _ SUi,c)
14 remove Fi from S0
list ;
15 Fi = getNextElement(S0
list ,Fi);
16 end until (Fi == NULL);
```

17 Fj = getNextElement(S0
list ,Fj);
18 end until (Fj == NULL);
19 Sbest = S0
list ;
20 end;

FCBF Algorithm

for Fi (line 13), Fi will be removed from S0 list . After one round of filtering features based on Fj, the algorithm will take the remaining feature right next to Fj as the new reference (line 17) to repeat the filtering process. The algorithm stops until no more predominant features can be selected.

In our method approximates relevance and redundancy analysis by selecting all predominant features and removing the rest features. It uses both C- and F-correlations to determine feature redundancy and combines sequential forward selection with elimination so that it not only circumvents full pair-wise F-correlation analysis but also achieves higher efficiency than pure sequential forward selection or backward elimination. However, our method is suboptimal due to the way C- and F-correlations are used for relevance and redundancy analysis and the approximates that it uses. It is fairly straightforward to improve the optimality of the results by considering different combinations of features in evaluating feature relevance and redundancy, which in turn increases time complexity. Another way to improve result optimality is to find better heuristics in determining a feature's approximate Markov blanket.

5 EXPERIMENTAL SETUP

The efficiency of a feature selection algorithm can be immediately evaluated by its playing time over several data sets. As to effectiveness, a bare and direct evaluation criterion is how similar the selected subset and the optimal subset are, but it can simply be measured over synthetic data for which we know beforehand which features are irrelevant or superfluous. For real-world data, we oftentimes do not have such prior knowledge about the optimal subset, and then we apply the predictive accuracy on the chosen subset of features as an indirect step.

In conditions of the above measures, we limit our comparisons to the filter model as FCBF is a filter algorithm designed for high-dimensional information. We choose representative algorithms from both attacks. One algorithm, from individual evaluation is Relief which searches for nearest neighbors of instances of different grades and weights features according to how well they differentiate instances of different years. Another algorithm, from subset evaluation, is a variation of CFS denoted by CFS-SF (Sequential Forward). CFS exploits best-first search based on some correlation measure which measures the goodness of a subset by considering the individual predictive ability of each feature and the degree of correlation between them. Sequential forward selection is used in CFS-SF as initial experiments show CFS-SF runs much faster to create similar results than CFS. A third one, also from subset evaluation, is a variation of FOCUS denoted by FOCUS-SF. FOCUS exhaustively examines all subsets of characteristics, taking the minimal subset that separates classes as consistently as the full set can. It is prohibitively costly, even for data sets with moderate dimensionality. FOCUS-SF replaces exhaustive search in FOCUS with sequential forward selection. In our experiments, we heuristically set the relevance threshold g to be the SU value of the bin/logNcth ranked feature for each data set. To test how the selection of threshold affects the performance of FCBF, we also include in our comparisons the results of FCBF with g set to the default value 0. We use FCBF(log) to represent a version of FCBF with the former setting, and FCBF(0) with the latter setting in the rest of the paper.

6. CONCLUSIONS

In this report, we have identified the need for explicit redundancy analysis in feature selection, provided a formal definition of feature redundancy, and investigated the relationship between feature relevance and redundancy. We have offered a novel framework of efficient feature selection via relevance and redundancy analysis, and a correlation-based method which uses C-correlation in relevance analysis and both C- and F-correlations for redundancy analysis.

A novel feature selection algorithm FCBF is implemented and assessed through extensive experiments comparing with three representative feature selection algorithms. The feature selection results are further verified by two different learning algorithms. Our method demonstrates its efficiency and effectiveness of feature selection in supervised learning in areas where data contains many irrelevant and/or superfluous features.

Some future works are designed along the accompanying directions. First, since the symmetrical uncertainty measure only handles nominal or discrete values, our current method calls for continuous values be discretized, which affords the opportunity to investigate how different discretization methods affect the performance of FCBF. Second, it would be interesting to explore measures that can treat all types of values or ways of combining different amounts under our framework of relevance and redundancy analysis. Another direction is to investigate how our method can be extended to deal with regression problems in which the class contains continuous values. Moreover, additional effort is needed to experiment our method on genomic microarray data for informative gene selection and investigate how small samples affect the performance of feature selection.

REFERENCES

1.H. Almuallim and T. G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 69(1-2):279–305, 1994.
2.D. A. Bell and H. Wang. A formalism for relevance and its application in feature subset selection.Machine Learning, 41(2):175–195, 2000.
3.A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. Artificial Intelligence, 97:245–271, 1997.
4.M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering – a filter solution.In Proceedings of the Second International Conference on Data Mining, pages 115–122, 2002.
5.M. Dash and H. Liu. Feature selection for classification. Intelligent Data Analysis: An International Journal, 1(3):131–156, 1997.
6.M. Dash and H. Liu. Consistency-based search in feature selection. Artificial Intelligence, 151(1-2):155–176, 2003.
7.J. G. Dy, C. E. Brodley, A. C. Kak, L. S. Broderick, and A.M. Aisen. Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Transactions on Pattern Analysis andMachine Intelligence, 25(3):373–378, 2003.
8.U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pages 1022–1027, 1993.