INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1320 - 1324
RECEIVED : 29/10/2016. PUBLISHED: 07/11/2016
November 7, 2016

EFFICIENT LEARNING RANK BASED METHOD FOR SEARCHING ANONYMOUS DATA

[1]S.Sharmila, [2]Thulasidass,
[1]PG Scholar,Dept Of Computer Science Engineering,Mailam Engineering College,Villupuram,
[2]Assistant Professor Of Computer Science Engineering,Mailam Engineering College,Villupuram.

Abstract:

Feature selection is applied to reduce the number of features in many applications where data has hundreds or thousands of features. Existing feature selection methods mainly focus on finding relevant features. In this paper, we show that feature relevance alone is insufficient for efficient feature selection of high-dimensional data. We define feature redundancy and propose to perform explicit redundancy analysis in feature selection. A new framework is introduced that decouples relevance analysis and redundancy analysis. We develop a correlation-based method for relevance and redundancy analysis, and conduct an empirical study of its efficiency and effectiveness comparing with representative methods.

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

1. INTRODUCTION

In classic supervised learning, one is given a training set of labeled fixed-length feature vectors (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 l possible classes c1,...,cl to the class label C. The task is to induce a hypothesis (classifier) that accurately predicts the labels of novel instances. The learning of the classifier is inherently determined by the feature-values. In theory, 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. Feature selection has been an active and fruitful field of research and development for decades in statistical pattern recognition (Mitra et al., 2002), machine learning (Liu et al., 2002b; RobnikSikonja and Kononenko, 2003), data mining (Kim et al., 2000; Dash et al., 2002) and statistics (Hastie et al., 2001; Miller, 2002). It has proven in both theory and practice effective in enhancing learning efficiency, increasing predictive accuracy, and reducing complexity of learned results (Almuallim and Dietterich, 1994; Koller and Sahami, 1996; Blum and Langley, 1997). 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 (Koller and Sahami, 1996). We call such a minimum subset an optimal subset, illustrated by the example below.

2. FRAMEWORK

From previous discussions, it is clear that in order to eliminate redundant features, the state-of-theart feature selection methods have to rely on the approach of subset evaluation which implicitly handles feature redundancy with feature relevance. These methods can produce better results than

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1320 - 1324
RECEIVED : 29/10/2016. PUBLISHED: 07/11/2016
November 7, 2016

methods without handling feature redundancy, but the high computational cost of the subset search makes them inefficient for high-dimensional data. 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.

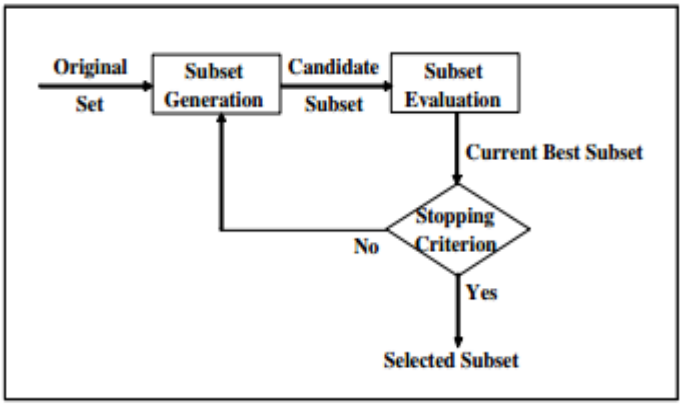

Fig.1. A traditional framework of feature selection

Our goal is to efficiently find the optimal subset . 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. It is sensible to use efficient heuristic methods to approximate the computation of relevant features and redundant features under our new framework for two reasons. On one hand, searching for an optimal subset based on the definitions of feature relevance and redundancy is combinatorial in nature. It is obvious that exhaustive or complete search is prohibitive with a large number of features. On the other hand, an optimal subset is defined based on the full population where the true data distribution is known. It is generally assumed that a training data set is only a small portion of the full population, especially in a high-dimensional space. Therefore, it is not proper to search for an optimal subset from the training data as over-searching the training data can cause over-fitting . We next present our approximation method.

3. ALGORITHM

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, and (2) selecting predominant features from relevant ones. As shown in Figure 4, 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 S 0 list based on a predefined threshold δ, 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 S 0 list to select predominant features. A

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1320 - 1324
RECEIVED : 29/10/2016. PUBLISHED: 07/11/2016
November 7, 2016

```
input:  S(F1, F2, ..., FN, C)   // a training data set
        δ                       // 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 > δ)
5         append Fi to S'list;
6     end;
7     order S'list in descending SUi,c value;
8     Fj = getFirstElement(S'list);
9     do begin
10      Fi = getNextElement(S'list, Fj);
11      if (Fi <> NULL)
12        do begin
13          if (SUi,j ≥ SUi,c)
14            remove Fi from S'list;
15          Fi = getNextElement(S'list, Fi);
16        end until (Fi == NULL);
17      Fj = getNextElement(S'list, Fj);
18    end until (Fj == NULL);
19  Sbest = S'list;
20  end;
```

Fig.2.Algorithm

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 S 0 list (line 8) and continues as follows. For all the remaining features (from the one right next to Fj to the last one in S 0 list), if Fj happens to form an approximate Markov blanket.

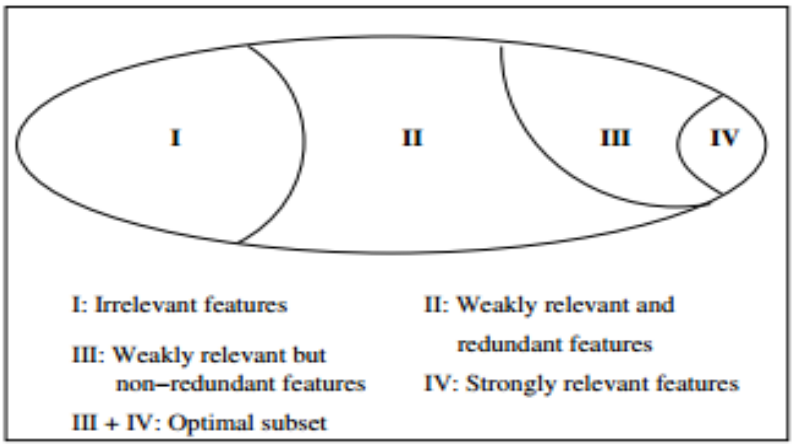

Fig.3. A view of feature relevance and redundancy

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1320 - 1324
RECEIVED : 29/10/2016. PUBLISHED: 07/11/2016
November 7, 2016

4. ANALYSIS

The efficiency of a feature selection algorithm can be directly measured by its running time over various data sets. As to effectiveness, a simple and direct evaluation criterion is how similar the selected subset and the optimal subset are, but it can only be measured over synthetic data for which we know beforehand which features are irrelevant or redundant. For real-world data, we often do not have such prior knowledge about the optimal subset, so we use the predictive accuracy on the selected subset of features as an indirect measure. In terms of the above criteria, we limit our comparisons to the filter model as FCBF is a filter algorithm designed for high-dimensional data. We choose representative algorithms from both approaches (i.e., individual evaluation and subset evaluation).

Title	$FCBF_{(log)}$	$FCBF_{(0)}$	ReliefF	CFS-SF	FOCUS-SF
Lung-cancer	0.001	0.02	0.09	0.05	0.08
Promoters	0.001	0.02	0.06	0.03	0.16
Splice	0.20	0.55	0.89	0.55	16.59
USCensus90	0.30	0.50	2.94	0.52	77.67
CoIL2000	0.25	0.50	4.25	1.98	143.94
Chemical	0.05	0.05	1.36	0.28	6.56
Musk2	0.53	0.88	9.55	4.84	85.78
Arrhythmia	0.06	0.08	1.19	0.78	13.70
Isolet	0.42	3.05	10.05	93.94	107.33
Multi-Features	1.19	19.42	11.42	71.00	67.56

Table.1.Result analysis

One algorithm, from individual evaluation, is ReliefF (Robnik-Sikonja and Kononenko, 2003) which searches for nearest neighbors of instances of different classes and weights features according to how well they differentiate instances of different classes. Another algorithm, from subset evaluation, is a variation of CFS In various machine learning domains, there are two forms of high-dimensional data. Traditionally, the dimensionality is usually thought high if data contains tens or hundreds of features. In this form of data, the number of instances is normally much larger than the dimensionality. In new domains such as text categorization and genomic microarray analysis, the dimensionality is in the order of thousands or even higher, and often greatly exceeds the number of instances. Therefore, we evaluate our method in comparison with others on high-dimensional data of both forms.

CONCLUSION

In this paper, 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 proposed a new framework of efficient feature selection via relevance and redundancy analysis, and a correlation-based method which uses C-correlation for relevance analysis and both C- and F-correlations for redundancy analysis. A new feature selection algorithm FCBF is implemented and evaluated 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 for feature selection in supervised learning in domains where data contains many irrelevant and/or redundant features.

REFERENCES

[1] Absil, Pierre-Antoine, Mahony, Robert E., and Sepulchre, Rodolphe. Optimization Algorithms on Matrix Manifolds. Princeton University Press, 2008.

[2] Bellet, Aurelien, Habrard, Amaury, and Sebban, Marc. A ´ survey on metric learning for feature vectors and structured data. CoRR, abs/1306.6709, 2013.

[3] Chechik, Gal, Sharma, Varun, Shalit, Uri, and Bengio, Samy. Large scale online learning of image similarity through ranking. In IbPRIA, pp. 11–14, 2009.

[4] Davis, Jason V., Kulis, Brian, Jain, Prateek, Sra, Suvrit, and Dhillon, Inderjit S. Information-theoretic metric learning. In International Conference on Machine Learning (ICML), 2007.

[5] Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-jia, Li, Kai, and Li, Fei-fei. Imagenet: A large-scale hierarchical image database. In Proc. IEEE CVPR, 2009.

[6] Do, Huyen, Kalousis, Alexandros, Wang, Jun, and Woznica, Adam. A metric learning perspective of svm: on the relation of lmnn and svm. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2012.