ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 161-166

Searching optimal feature subset using mutual information

D. Huang and Tommy W.S. Chow

Dept of Electronic Engineering, City university of Hong Kong

Abstract

A novel feature selection methodology is proposed with the concept of mutual information. The proposed methodology effectively circumvents two major problems in feature selection process: to identify the irrelevancy and redundancy in the feature set, and to estimate the optimal feature subset for classification task.

1. Introduction

Feature selection is a necessary preprocessing stage in classification especially when one is required to manage large or even overwhelming quantities of data. A process of feature selection should be guided by certain criteria. It is expected that feature selection criteria can address two problems: to identify the salient features and to estimate the optimal number of features from a large feature set. Many well-known feature selection criteria are based on the statistics of input variables, i.e., the probability [1] and the covariance of input variable [2]. To perform feature selection by counting the data points is sensitive to noises [3]. The major shortcoming of the covariance-based criteria is their sensitivity to data transformation. Also, the high order statistics are not taken into account in [1,2]. Recently, the mutual information (MI) is used by feature selection methods [4-7]. The advantages of MI have been detailed in [4,8,9]. Generally, many MI-based feature selection methods [4,5] are a forward selection process in which features are selected one by one. For this type of forward selection process, MI between selected input variables and output class labels monotonically increases because addition of input variables cannot decrease MI [10]. Also, with increasing of the number of selected features, the rate of increased MI gradually decreases to zero where all relevant features have been selected. With these characteristics, MI-based criteria can not only effectively select features but also roughly estimate optimal feature subset. However, estimating MI is hard because of requirement for the conditional density functions and the highly computational complexity. Many MI-based feature selection algorithms [4-7] used histogram as density estimator. In high dimensional space, histogram is neither effective nor accuracy [13]. Accordingly, MIFS [4] and MIFS-U [5] estimated 2-dimensional MI and analyzed high-dimensional MI with 2-dimensional MI. The experimental results have shown their effectiveness. However, no direct estimate to high-dimensional MI causes problems. First, as no monotonic criteria are available, these methods depend on generalization performance to estimate optimal feature subset, which is highly computationally demanding and classifier-dependent. Second, these methods are suboptimal in the sense that selected features are individually, instead of globally, considered. Third, no analytic guides are provided to avoid redundancy among the selected features.

In this paper, a novel toward-optimal feature selection methodology (OFS-MI) is proposed. The proposed OFS-MI consists of two MI-based criteria and a forward searching algorithm. The advantages of OFS-MI include: 1) these MI-based criteria enable us systematically to estimate the optimal feature sets, which is very important when the given data set contain huge quantities of features; 2) the direct estimate of MI is able to give optimal solution; 3) minimizing the redundancy in selected feature set can be operated in a principled way.

2. Quadratic MI and its estimation

In the proposed methodology, the quadratic MI [8] is used in order to greatly reduce computational demanding,

$$I_{CS}(X;C) = \log \frac{(\sum_c \int\ p(x,c)^2 dx)(\sum_c p(c))^2 (\int p(x)^2 dx)}{(\sum_c p(c) \int p(c,x) p(x) dx)^2}\ , \qquad (1)$$

where X and C are continuous input variables and output class labels respectively. The underlying probability density functions (pdf) are required to calculate (1). In order to further reduce computational complexity, a supervised clustering algorithm is proposed in this paper. The parameter $\sigma_{\min}$ is selected as 0.05 because all the data sets used in this paper are all normalized to have zero means and unit variances.

```
Initializing: For class $c_k$ , create a cluster $u_k$ , its center $g_k = \sum_{x_i \in class c} x_i / m_k$ and its label $l_k = k$ , End For.
Clustering: While (cardinality of training pattern set ($X$) > 0)
    Randomly selecting $x$ from $X$. Find out $g_i$ , minimizing $\| x - g_j \|, \forall j$ .
    If the class of $x = l_i$ , $x \to u_i$ , recalculating $g_i = mean(all\ x \in u_i)$
    Else $m = m + 1$ ; New cluster $u_m$ , $g_m = x$ , $l_m$ = class of $x$ ; $X \leftarrow X \setminus \{x\}$ ; End While
Data selecting: Selected pattern set $SX$={};
   For every cluster $u_i$ , $X_i = \{$all the $x \in u_i\}$ , $\sigma_i = \max(\mathrm{var}(X_i))$ , $J_i$ = cardinality ( $X_i$ ).
   If $\sigma_i > \sigma_{\min}$ Sampling 10 data patterns from $X_i$ into $SX$ when $J_i > 10$ ; Or $X_i \to SX$ when $J_i \le 10$ ,
   Else $x = median(X_j) \to SX$ ; End If; End For.
```

For a cluster u_j, let n_j^x be the number of data points in the original dataset and n_j^s be the number of data points in the selected dataset *SX*. Obviously, we have $n_j^x \ge n_j^s$. The pdfs can be estimated,

$$p(sx_i) = n_i^x / (N \times n_i^s)\ ,\ p(x) = \sum_{all\, sx_i} p(sx_i) G(x - sx_i, \Sigma_i)\ ,\ \ p(c_k) = \sum_{sx_i \in class\, c_k} p(sx_i)$$

$$p(x \mid c_k) = \sum_{sx_i \in class\, c_k} p(sx_i) G(x - sx_i, \Sigma_i), \qquad k = 1,2,..., N_c\ .$$

In the above pdf estimators, Gaussian function (2) is used

$$G(z - z_0, \Sigma) = \frac{1}{(2\pi h)^{M/2} \|\Sigma\|^{1/2}} \exp(-\frac{(z - z_0)^T\ \Sigma^{-1} (z - z_0)}{2h^2})\ , \qquad (2)$$

where Σ is determined from the covariance matrix of the overall data, h is the bandwidth of kernel function. In this paper, $\Sigma = I$ because all input data are normalized to zero mean and unit variance. The “optimal” bandwidth h in (2) is given by Silverman [11], i.e., $h = [4/(M+2)]^{1/(M+4)} n^{-1/(M+4)}$.

With the property $\int G(x-x_1,\Sigma_1)G(x-x_2,\Sigma_2)dx = G(x_1-x_2,\Sigma_1+\Sigma_2)$, the quadratic MI (1) can be calculated

$$I_{CS}(X;C) = \log \frac{V_{(c,x)^2} V_{(c)^2} V_{(x)^2}}{V_{(cx)}{}^2}$$

where $V_{(c,x)^2} = \Sigma_c \int p(c,x)^2 dx = \Sigma_k^l \Sigma_{sx_i \in \text{ class } c_k} \Sigma_{sx_j \in \text{ class c}_k} p(sx_i)p(sx_j)G(sx_i - sx_j, 2I)$,

$$V_{(c)^2} = \Sigma_k^l p(c_k)^2 = \Sigma_{k=1}^{Nl} \left(\Sigma_{sx_i \in \text{ class } l_k} p(sx_i)\right)^2 ,$$

$$V_{(x)^2} = \int p(x)^2 dx = \Sigma_{sx_i} \Sigma_{sx_j} p(sx_i)p(sx_j)G(sx_i - sx_j, 2I) ,$$

$$V_{(cx)} = \Sigma_k \int p(c_k,x)p(c_k)p(x)dx = \Sigma_k \left[(\Sigma_{sx_i \in \text{ class } c_k} p(sx_i)) \Sigma_{sx_i} \Sigma_{sx_j \in \text{ class c}_k} p(sx_j)p(sx_i)G(sx_j - sx_i, 2I) \right]$$

3. OFS-MI

The proposed OFS-MI consists of two MI-based criteria and a forward searching process. Feature relevancy criterion (*FRC*) is aimed at selecting the relevant features. And the feature similarity criterion (*FSC*) is used to reduce the redundancy among the selected feature set. Below, let *F* and *S* be the candidate feature set and the selected feature set respectively. In order to determine the most relevant feature in *F*, all the candidate features are ordered by using $FRC(f) = I(f,S;C)$. Because of the trivial estimation error, adding input variables with little information about classification may lead to the decrease of *FRC*. In the proposed methodology, the problem is overcome by using the conservative stopping criteria. Feature similarity criterion ($FSC(f;S) = \arg\max\left(I(f;f_i)/H(f_i)\right)$) measures the similarity between feature f and subset $S(f \notin S)$. When $FSC(f;S) \geq \theta$, it can be concluded that f is similar to S. In this paper, $\theta = 0.5$. *FSC* can be estimated by using the quadratic MI between two continuous variables. The proposed OFS-MI may be stated as follows.

```
Step 1. Set $F \leftarrow$ “initial feature set”, $S \leftarrow$ Empty, the number of selected features $j = 0$;
Step 2. Set $F \leftarrow F \setminus \{f_k\}, S \leftarrow \{f_k\}$, $f_k$ maximizing $FRC(f) = I(f;C), \forall f \in F$; $j = 1$, $FRC_j = FRC(f_k)$
Step 3. Search $f_k \in F$ maximizing $FRC(f_k + S;C)$, and $F \leftarrow F \setminus \{f_k\}$;
Step 4. Identify the redundancy. If $FSC(f_k) \geq \theta$, Goto Step 5.
 Otherwise, $S \leftarrow \{f_k\}$, $j = j+1, FSC_j = FSC(f_k)$ Goto Step 5.
Step 5. Stop the process. If $FRC_j < \rho \arg\max(FRC_l), 1 \leq l \leq j-1$, Goto Step 6, shown as Fig 1 (a);
 Otherwise, If $(FRC_j - FRC_{j-1})/FRC_1 \leq \lambda_r$ for $n_{stop}$ times consecutively, Goto Step 6, shown as Fig 1(b);
 Otherwise, Goto Step 3.
Step 6. Estimate the appropriate number of features for classification ( $nf\_app$ ).
  $nf\_app_l \leq nf\_app \leq nf\_app_r$ where $nf\_app_l$ satisfied $FRC_{nf\_app_l} \geq \lambda_l \times \max(FRC)$, and $nf\_app_r = j-1$.
Step 7. Output the set $S$ and $nf\_app_l \leq nf\_app \leq nf\_app_r$.
```

4. Evaluations and comparison

In this section, by using 3 simulation studies consisting of synthetic data and real data, the proposed OFS-MI methodology is thoroughly examined in two perspectives: effectiveness of feature selection and correctness of estimate on optimal feature subset. A priori knowledge about the synthetic data can be used to evaluate the results of feature selection. For the real data, the feature selection results are evaluated by three feature evaluation indices. Two of them are classifier based. The K-NN rule and

MLP network are adopted for this purpose. The MLP networks are multilayer perceptrons with one hidden layer containing three neurons and trained by using fast BP provided in Matlab. k=3 in the K-NN rule. Obviously, the better generalization performance of classifier means the better feature subset. The other index, namely class separability (cs), is classifier-free. With the similar idea [12], cs is calculated by $cs = trace(S_w^{-1} S_b)$, where S_b is the between class scatter matrix, and S_w is the within class scatter matrix. A higher value of the class separability index ensures that the classes are well separated by their scatter means. The classification performances of classifier are regarded as a more convinced evaluation on *nf_app*. The above classifiers (MLP, KNN) are used for this task. In the section, the thresholds in the stopping criteria of the proposed OFS-MI are set as $\rho = 0.95$, $\lambda_l = 0.9$, $\lambda_r = 0.05$, $n_{stop} = 5$. These parameters are not crucial because, with the proposed FRC, other rational selections can work well even better.

4.1 LED dataset

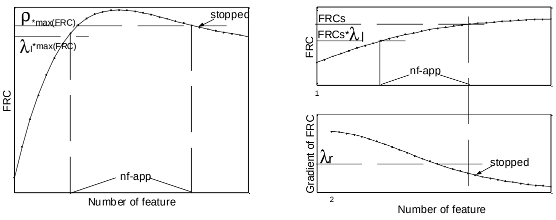
The dataset of LED display domain [13] has 24 features, in which the first 7 features determine the class label of a pattern, whilst the rest 17 features are irrelevant. In this study, 400 data patterns were generated. In order to further examine the capabilities of OFS-MI in dealing with irrelevant features and redundant features, additional 24 redundant features are added to increase the total number of feature to 48. The 24 redundant features are based on the original 24 features added with noise, generated on the basis of normal distribution $\aleph(0,0.025)$. In our simulation, it can be found that only MIFS $\beta = 0.5$ and OFS-MI can give the correct answer for this dataset. When $\beta = 1$, MIFS selected the irrelevant features because the redundancy in the selected features is taken excessive care. Based on FRC as shown in Fig 2, it can be estimated that $4 \le nf\_app \le 7$, which is consistent with prior knowledge.

4.2 Sonar dataset

This dataset [13] consists of 208 patterns including 104 training and test patterns. It has 60 input features and two output classes. FRC and its gradient are shown in Fig 3 (a, b). The results in Table 1 show that MLP-based and K-NN classifier together with the proposed OFS-MI could achieve the best performances at all cases. And, class separability criterion (cs) is much better than MIFS and MIFS-U. Using OFS-MI, it is estimated that $9 \le nf\_app \le 15$. There are up to 40 features selected by the OFS-MI in order to validate the estimation result of $9 \le nf\_app \le 15$. The generalization accuracy values of K-NN, MLP are visualized in Fig 3(c). The curves suggest that all these classifiers are able to generalize the best with $9 \le nf\_app \le 15$, so the estimation result on nf_app is plausible.

4.3 cDNA data for ovarian cancer classification

There are 253 data samples and each sample contains 15154 features (genes) in this dataset [14]. We partition the dataset into two parts, 150 for training and 103 for test. For this dataset, there is no supervised clustering strategy because of the huge feature set. Parzen window estimator is used as pdf estimator, i.e., the whole data is used as selected dataset *SX*. In order to effectively dealing with the dataset containing a large quantity of feature, a special strategy is adopted with consideration that many features are useless. In this study, the 15154 features are ranked on the base of *I(f;C)*, and the 600 top features are selected into feature set *F* at step 1 for iterative feature selection process. The MIFS and MIFS-U methodology were also tested on this dataset. They were implemented in the same way with the proposed OFS-MI, i.e., the top 600 individually relevant features are selected into the feature set *F* before the iterative feature selection. The comparisons of results are shown in Table 1. Obviously, the proposed methodology is much more effective than other methods when handling this huge dataset. Based on the curves in Fig 4(a,b), we have $6 \le nf\_app \le 10$.And the curves in Fig 4(c) suggest that the estimation on nf_app is correct.



(a) (b)

Fig 1. Stopping criteria.

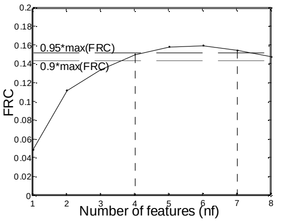
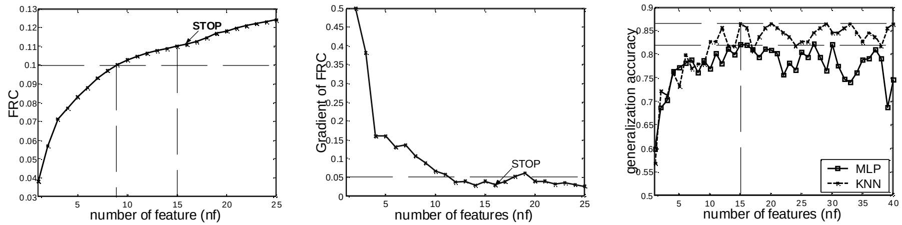
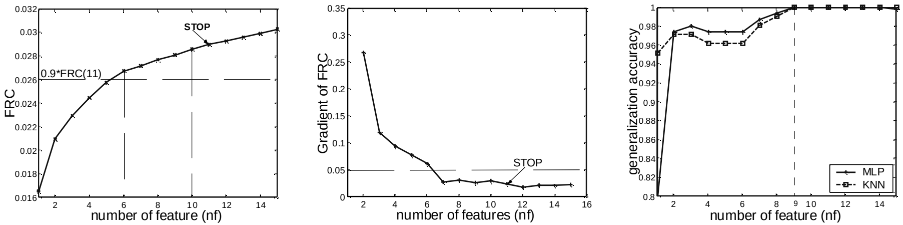


Fig 2. The FRC curve on the LED data.



(a) (b) (c)

Fig 3. The results of OFS-MI and the corresponding estimations on sonar dataset.



(a) (b) (c)

Fig 4. The results of OFS-MI and the corresponding estimations on cDNA dataset

5. Discussions and Conclusion

The major objectives of this paper are to effectively select the features and to determine the optimal or near-optimal feature subset. A novel feature selection methodology, OFS-MI, is successfully developed by using two MI-based criteria. The proposed OFS-MI was thoroughly examined by different classification problems. And both priori knowledge and the generalization results corroborate that the estimation on optimal feature subset is plausible. Also, the comparisons with other MI-based methods show the effectiveness of the proposed OFS-MI.

Sonar dataset									
nf	Class separability *cs*			KNN generalization accuracy			MLP generalization accuracy		
	9	12	15	9	12	15	9	12	15
MIFS $\beta=0$	0.93	1.02	1.20	**0.779**	0.846	0.846	0.783	0.793	0.750
MIFS $\beta=1$	0.43	0.48	0.57	0.730	0.712	0.692	0.691	0.718	0.760
MIFS-U $\beta=1$	0.98	1.05	1.05	0.759	0.759	0.759	0.686	0.710	0.764
OFS-MI	**1.10**	**1.24**	**1.38**	**0.779**	**0.856**	**0.865**	**0.785**	**0.818**	**0.820**
CDNA dataset									
nf	Class separability *cs*			KNN generalization accuracy			MLP generalization accuracy		
	5	7	9	5	7	9	5	7	9
MIFS $\beta=0$	10.3	14.9	16.4	0.950	0.971	0.971	0.967	0.891	0.980
MIFS $\beta=1$	**21.2**	21.6	22.5	0.970	0.971	0.971	0.802	0.980	0.973
MIFS-U $\beta=1$	19.5	21.9	24.8	**0.971**	0.971	0.971	**0.980**	0.980	0.982
OFS-MI	13.4	**22.1**	**49.3**	0.961	**0.981**	**1.00**	0.973	**0.987**	**1.00**

Table 1. The comparison of MIFS $\beta=0$, MIFS $\beta=1$, MIFS-U $\beta=1$ and OFS-MI. *nf* is the number of selected features. We outline in boldface the best values of indices at all cases.

References:

1. P.M. Narendra, K. Fukunaga, "A branch and bound algorithm for feature selection, " IEEE. Trans on computers, C-26(9), pp917-922, 1977.
2. H. Lin, H. Motoda, M. Dash, "A monotonic measure for optimal feature selection," In Proceedings of European Conference on Machine Learning, pp101-106, 1998.
3. M. Dash, H. Liu, "Feature selection for classification," Intelligent data analysis, pp 131-156, 1997.
4. R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Tran. on Neural network, vol. 5, No 4, 1994, pp537-550.
5. Nojun Kwak, C-H Choi, "Input feature selection for classification problems," IEEE. Trans on Neural networks, vol 13 no 1, pp143-159, 2002.
6. B.Bonnlander A.S.Weigend, "Selecting input variables using mutual information and nonparametric density estimation," In Proceedings of 1994 International Symposium on Artificial Neural network, Taiwan, 1994, pp42-50.
7. Ani, M. Deriche, "An optimal feature selection technique using the concept of mutual information," In Proceedings of A.A ISSPA, Malaysia, 2001
8. J.C.Principe, J. Fisher III, Xu D, "Information theoretic learning," in Unsupervised adaptive filtering, eds. S.Haykin, New York, NY: Wiley, 2000.
9. W.T Li, "Mutual information functions versus correlation functions," Stat. Phys., vol. 60, no.5/6, pp 823-837, 1990
10. T.M.Cover, J.A.Thomas, "Elements of information theory", New York: John Wiley, 1994.
11. Silverman, B.W., Density estimation for statistics and data analysis, London, Chapman-Hall, 1986.
12. P.A. Devijver, J.Kittler, "Pattern recognition: A statistical approach," Englewood Cliffs: Prentice Hall, 1982.
13. Y. Moon, B. Rajagopalan, U. Lall, "Estimation of mutual information using kernel density estimators," Phys. Rev. E, 52(3), 3, pp. 2318-2321, 1995
14. Data available in UCI machine learning repository at http://www.ics.uci.edu/~mlearn/MLRepository.html.
15. Data available at http://clinicalproteomics.steem.com.