

# A Hypergraph-based Approach to Feature Selection

Zhihong Zhang and Edwin R.Hancock\*

Department of Computer Science, University of York, UK,  
{zhihong, erh}@cs.york.ac.uk

**Abstract.** In many data analysis tasks, one is often confronted with the problem of selecting features from very high dimensional data. The feature selection problem is essentially a combinatorial optimization problem which is computationally expensive. To overcome this problem it is frequently assumed that either features independently influence the class variable or do so only involving pairwise feature interaction. To overcome this problem, we draw on recent work on hyper-graph clustering to extract maximally coherent feature groups from a set of objects using high-order (rather than pairwise) similarities. We propose a three step algorithm that, namely, i) first constructs a graph in which each node corresponds to each feature, and each edge has a weight corresponding to the interaction information among features connected by that edge, ii) perform hypergraph clustering to select a highly coherent set of features, iii) further selects features based on a new measure called the multidimensional interaction information (MII). The advantage of MII is that it incorporates third or higher order feature interactions. This is realized using hypergraph clustering, which separates features into clusters prior to selection, thereby allowing us to limit the search space for higher order interactions. Experimental results demonstrate the effectiveness of our feature selection method on a number of standard data-sets.

**Keywords:** Hypergraph clustering; Multidimensional interaction information(MII)

## 1 Introduction

High-dimensional data pose a significant challenge for pattern recognition. The most popular methods for reducing dimensionality are variance based subspace methods such as PCA. However, the extracted PCA feature vectors only capture sets of features with a significant combined variance, and this renders them relatively ineffective for classification tasks. Hence, it is crucial to identify a smaller subset of features that are informative for classification and clustering. The idea underpinning feature selection is to a) reduce the dimensionality of the feature space, b) speed up and reduce the cost of a learning algorithm, c) obtain

---

\* Edwin Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

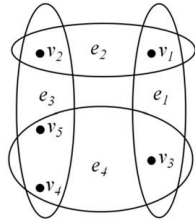
the feature subset which is most relevant to classification. In practice, however, optimal feature selection requires  $2^n$  feature subset evaluations, where  $n$  is the original number of features and many problems related to feature selection are shown to be NP-hard [2]. Traditional feature selection methods address this issue by partitioning the original feature set into distinct clusters formed by similar features [3]. However, all of the above methods are weakened by only considering pairwise relations. In some applications higher-order relations are more appropriate to the classification task on hand, and approximating them in terms of pairwise interactions can lead to a substantial loss of information.

To overcome the above problem, in this paper, we propose a hypergraph-based approach to feature selection. Hypergraph clustering is capable of detecting high-order feature similarities. In this feature selection scheme, the original features are clustered into different groups based on hypergraph clustering and each group includes just a small set of features. In addition, for each group, a new feature selection criterion referred to as multidimensional interaction information (MII)  $I(F; C)$  is applied to feature selection. In contrast to existing feature selection criterion, MII is sensitive to the relations between feature combinations and can be used to seek third or even higher order dependencies between the relevant features. However, the limitations of the MII criterion are that it requires an exhaustive “combinatorial” search over the feature space and demands estimation of the joint probability distribution for features using large training samples. So most existing works use MII based on the second-order feature dependence assumption [1]. Since hypergraph clustering separates features into clusters in advance, this allows us to limit the search space for higher order interactions directly using the MII criterion  $I(F; C)$  for feature selection. Using the Parzen window for probability distribution estimation, we apply a greedy strategy to incrementally select the features that maximize the multidimensional mutual information between the current selected features and the output class set.

## 2 Hypergraph Clustering Algorithm

**Concept of hypergraph:** A hypergraph is defined as a triplet  $H = (V, E, s)$ , where  $V = \{1, \dots, n\}$  is the node-set,  $E$  is a set of non-empty subsets of  $V$  or hyperedges and  $s$  is a weight function which associates a real value with each edge. A hypergraph is a generalization of a graph. Unlike graph edges which consisting pairs of vertices, hyperedges are arbitrarily sized sets of vertices. Examples of a hypergraph are shown in Fig. 1. For the hypergraph, the vertex set is  $V = \{v_1, v_2, v_3, v_4, v_5\}$ , where each vertex represents a feature, and the hyper-edge set is  $E = \{e_1 = \{v_1, v_3\}, e_2 = \{v_1, v_2\}, e_3 = \{v_2, v_4, v_5\}, e_4 = \{v_3, v_4, v_5\}\}$ . The number of vertices constituting each hyperedge represent the order of the relationship between features.

**Hypergraph Clustering Algorithm:** Let  $H = (V, E, s)$  be a hypergraph clustering problem. We can locate the hypergraph cluster by finding the solutions



**Fig. 1.** Hypergraph example

of the following non-linear optimization problem that maximizes the functional

$$f(\mathbf{x}) = \sum_{e \in E} s(e) \prod_{i \in e} x_i . \quad (1)$$

subject to  $\mathbf{x} \in \Delta$ , where  $\Delta = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0, \sum_{i=1}^n x_i = 1\}$  and  $s$  is a weight function which associates a real value with each edge. The local maximum of  $f(x)$  can be solved using the Baum-Eagon inequality and leads to the iteratively updated lambda:

$$z_i = \frac{x_i \partial_i f(x)}{\sum_{j=1}^n x_j \partial_j f(x)}, i = 1, \dots, n . \quad (2)$$

where  $f(x)$  is a homogeneous polynomial in the variables  $x_i$  and  $z = \mathcal{M}(x)$  is a growth transformation of  $x$ . The Baum-Eagon inequality  $f(\mathcal{M}(x)) > f(x)$  provides an effective iterative means for maximizing polynomial functions in probability domains.

### 3 Feature Selection Using Hypergraph Clustering

In this paper we aim to utilize the hypergraph clustering algorithm for feature selection. Using a hypergraph representation of the features, there are three steps to the algorithm, namely a) computing the relevance matrix  $\mathbf{S}$  based on the interaction information among feature vectors, b) hypergraph clustering to cluster the feature vectors and c) selecting the optimal feature set from each cluster using the multidimensional interaction information (MII) criterion. In the remainder of this paper we describe these elements of our feature selection algorithm in more detail.

**Computing the Relevance Matrix:** In accordance with Shannon’s information theory, the uncertainty of a random variable  $Y$  can be measured by the entropy  $H(Y)$ . For two variables  $X$  and  $Y$ , the conditional entropy  $H(Y|X)$  measures the remaining uncertainty about  $Y$  when  $X$  is known. The mutual information (MI) represented by  $I(X; Y)$  quantifies the information gain about  $Y$  provided by variable  $X$ . The relationship between  $H(Y)$ ,  $H(Y|X)$  and  $I(X; Y)$  is  $I(X; Y) = H(Y) - H(Y|X)$ . As defined by Shannon, the initial uncertainty for the random variable  $Y$  is expressed as:  $H(Y) = -\sum_{y \in Y} P(y) \log P(y)$ , where

$P(y)$  is the prior probability density function over  $y \in Y$ . The remaining uncertainty in the variable  $Y$  if the variable  $X$  is known is defined by the conditional entropy  $H(Y|X) = -\int_x p(x)\{\sum_{y \in Y} p(y|x) \log p(y|x)\}dx$ , where  $p(y|x)$  denotes the posterior probability for variable  $y \in Y$  given another random variable  $x \in X$ . After observing the variable vector  $x$ , the amount of additional information gain is given by the mutual information (MI)  $I(X;Y) = \sum_{y \in Y} \int_x p(y,x) \log \frac{p(y,x)}{p(y)p(x)} dx$ .

From the above definition, we can see that mutual information quantifies the information which is shared by two variables  $X$  and  $Y$ . When the  $I(X;Y)$  is large, this implies that variable  $x \in X$  and variable  $y \in Y$  are closely related, otherwise, when  $I(X;Y)$  is equal to 0, this means that two variables are totally unrelated. Analogically, the conditional mutual information of  $X$  and  $Y$ , denoted as  $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$ , represents the quantity of information shared by  $X$  and  $Y$  when  $Z$  is known. The conditioning on a third random variable may either increase or decrease the original mutual information. That is, the difference between the conditional mutual information and the simple mutual information, referred to as the Interaction Information is:

$$I(X;Y;Z) = I(X;Y|Z) - I(X;Y) . \quad (3)$$

The interaction information measures the influence of the variable  $Z$  on the amount of information shared between variables  $\{Y, X\}$ , the value can be positive, negative, or zero. A zero value means that the relation between  $X$  and  $Y$  is entirely because of  $Z$ . A positive value means that  $X$  and  $Y$  are independent of each other. However, when combined with  $Z$ ,  $X$  and  $Y$  are correlated with each other. A negative value indicates that  $Z$  can account for or explain the correlation between  $X$  and  $Y$ . The extension of interaction information to  $n$  variables is defined recursively,

$$I(\{X_1, \dots, X_n\}) = I(\{X_1, \dots, X_{n-1}\}|X_n) - I(\{X_1, \dots, X_{n-1}\}) . \quad (4)$$

In our feature selection scheme, the high-order relevance of features is computed using interaction information. Suppose there are  $N$  training samples, each having  $K$  feature vectors. The  $k^{th}$  feature vector for the  $l^{th}$  training sample is  $f_k^l$ , and so we can represent the  $k^{th}$  feature vector for the  $N$  training samples as the long vector  $F_k = \{f_k^1, f_k^2, \dots, f_k^N\}$ . For three feature vectors  $F_{k1}, F_{k2}$  and  $F_{k3}$ , their interaction information  $I(F_{k1}, F_{k2}, F_{k3})$  can be computed by Equation (3). The relevance degree among three feature vectors  $F_{k1}, F_{k2}$  and  $F_{k3}$  can be defined as

$$\mathbf{S}(F_{k1}, F_{k2}, F_{k3}) = \frac{3I(F_{k1}, F_{k2}, F_{k3})}{H(F_{k1}) + H(F_{k2}) + H(F_{k3})} . \quad (5)$$

where  $k1, k2, k3 \in K$  and the higher the value of  $\mathbf{S}(F_{k1}, F_{k2}, F_{k3})$  the more relevant are the features  $F_{k1}, F_{k2}$  and  $F_{k3}$ . Otherwise, if  $\mathbf{S}(F_{k1}, F_{k2}, F_{k3}) = 0$ , the three features are totally unrelated. In addition, for the above computation, we use Parzen-Rosenblatt window method to estimate the probability density function of random variables  $F_{k1}, F_{k2}$  and  $F_{k3}$ . The Parzen probability density

estimation formula is given by:  $p(x) = \frac{1}{N} \phi(\frac{x-x_i}{h})$ , where  $\phi(\frac{x-x_i}{h})$  is the window function and  $h$  is the window width. Here, we use a Gaussian as the window function, so  $\phi(\frac{x-x_i}{h}) = \frac{1}{(2\pi)^{\frac{d}{2}} h^d |\Sigma|^{\frac{1}{2}}} \exp(\frac{(x-x_i)^T \Sigma^{-1} (x-x_i)}{-2h^2})$ , where  $\Sigma$  is the covariance of  $(x - x_i)$ ,  $d$  is the length of vector  $x$ . When  $d = 1$ ,  $p(x)$  estimates the marginal density and when  $d = 3$ ,  $p(x)$  estimates the joint density of variables such as  $F_{k1}$ ,  $F_{k2}$  and  $F_{k3}$ .

**Hypergraph Clustering:** the hypergraph clustering algorithm commences from the relevance matrix and iteratively bi-partitions the features into a foreground cluster and a background cluster. It locates the foreground cluster progressively and hierarchically. The clustering process stops when all the features are grouped into either the foreground or background cluster.

**Selecting Key Features:** The multidimensional interaction information between feature vector  $F = \{f_1, \dots, f_m\}$  and class variable  $C$  is:

$$I(F; C) = \sum_{f_1, \dots, f_m} \sum_{c \in C} P(f_1, \dots, f_m; c) \times \log \frac{P(f_1, \dots, f_m; c)}{P(f_1, \dots, f_m)P(c)}. \quad (6)$$

The main reason for using  $I(F; C)$  as a feature selection criterion is that since  $I(F; C)$  is a measure of the reduction of uncertainty in class  $C$  due to knowledge of the feature vector  $F = \{f_1, \dots, f_m\}$ , from an information theoretic perspective selecting features that maximize  $I(F; C)$  translates into selecting those features that contain the maximum information about class  $C$ . In practice, and as noted in the introduction, locating a feature subset that maximizes  $I(F; C)$  presents two problems: 1) it requires an exhaustive ‘‘combinatorial’’ search over the feature space, and 2) it demands large training sample sizes to estimate the higher order joint probability distribution in  $I(F; C)$  with a high dimensional kernel [6]. Bearing these obstacles in mind, most of the existing related papers approximate  $I(F; C)$  based on the assumption of lower-order dependencies between features. For example, the first-order class dependence assumption includes only first-order interactions. That is, it assumes that each feature independently influences the class variable, so as to select the  $m$ th feature,  $f_m$ ,  $P(f_m|f_1, \dots, f_{m-1}, C) = P(f_m|C)$ . A second-order feature dependence assumption is proposed by Guo and Nixon [5] to approximate  $I(F; C)$ , and this is arguably the most simple yet effective evaluation criterion for selecting features. The approximation is given as

$$I(F; C) \approx \hat{I}(F; C) = \sum_i I(f_i; C) - \sum_i \sum_{j>i} I(f_i; f_j) + \sum_i \sum_{j>i} I(f_i; f_j|C). \quad (7)$$

Although an MII based on the second-order feature dependence assumption can select features that maximize class-separability and simultaneously minimize dependencies between feature pairs, there is no reason to assume that the final optimal feature subset is formed by pairwise interactions between features. In fact, it neglects the fact that third or higher order dependencies can be lead to an optimal feature subset.

The primary reason for using the approximation  $\widehat{I}(F; C)$  for feature selection instead of directly using multidimensional interaction information  $I(F; C)$  is that  $I(F; C)$  requires estimation of the joint probability distribution of features using a large training sample. Consider the joint distribution  $P(F) = P(f_1, \dots, f_m)$ , by the chain rule of probability

$$\begin{aligned}
 P(f_i, \dots, f_m) &= P(f_1)P(f_2|f_1) \times P(f_3|f_2, f_1) \cdots P(f_m|f_1, f_2 \cdots f_{m-1}), \quad (8) \\
 P(F; C) &= P(f_1, \dots, f_m; C) = P(C)p(f_1|C)P(f_2|f_1, C)P(f_3|f_1, f_2, C) \\
 &\quad \times P(f_4|f_1, f_2, f_3, C) \cdots P(f_i|f_1, \dots, f_m, C). \quad (9)
 \end{aligned}$$

In our feature selection scheme, the original features are clustered into different groups based on hypergraph clustering and each cluster just includes a small set of features. Therefore, for each cluster, we do not need to use the approximation  $\widehat{I}(F; C)$ . Instead, we can directly use the multidimensional interaction information  $I(F; C)$  criterion for feature selection. Using Parzen windows for probability distribution estimation, we then apply the greedy strategy to select the feature that maximizes the multidimensional mutual information between the features and the output class set. As a result the first feature  $f'_{max}$  maximizes  $I(f', C)$ , the second selected feature  $f''_{max}$  maximizes  $I(f'', f', C)$ , the third feature  $f'''_{max}$  maximizes  $I(f''', f'', f', C)$ , and so on. For each cluster, we repeat this procedure until  $|S| = k$ .

## 4 Experiments and Comparisons

The data sets used to test the performance of our proposed algorithm are the benchmark data sets from the UCI Machine Learning Repository. Table. 1 summarizes the properties of these data-sets. Using the feature selection algorithm outlined above, we make a comparison between our proposed feature selection method (referred to as the *HGplusMII* method) (which utilizes the multidimensional interaction information (MII) criterion and hypergraph clustering for feature selection) and the use of multidimensional interaction information (MII) using the second-order approximation (see Equation (7)).

**Table 1.** Summary of UCI benchmark data sets

Data-set	Examples	Features	Classes
Australian	690	14	2
Breast cancer	699	10	2
Pima	768	8	2

The experimental results shown in Table. 2 demonstrate that our proposed method (i.e. *HGplusMII*) can achieve higher degree of dimensionality reduction, as it selects a smaller feature subset compared with those obtained using MII with second-order approximation. There are three reasons for this. The first

reason is that hypergraph clustering simultaneously considers the information-contribution of each feature and the correlation between features, so the structural information concealed in the data can be effectively identified. The second reason is that the multidimensional interaction information (MII) criterion is applied to each cluster for feature selection, and can consider the effects of third and higher order dependencies between the features and the class. As a result the optimal feature combination can be located so as to guarantee the optimal feature subset. The third and final reason is that second-order approximation to multidimensional interaction information (MII) simply checks for pair-wise dependencies between features and the class, and so only limited feature subsets can be obtained.

**Table 2.** The experiment results on three data-sets.

Method	Australian	Breast cancer	Pima
MII	$\{f_8, f_{14}, f_5, f_{13}\}$	$\{f_3, f_8, f_7\}$	$\{f_2, f_8, f_6, f_7\}$
HGplusMII	$\{f_8, f_9, f_5\}$	$\{f_3, f_7, f_9\}$	$\{f_2, f_6, f_1\}$

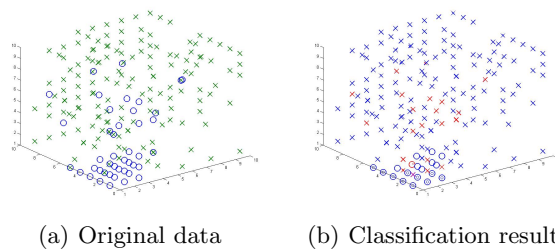
After obtaining the discriminating features, we compute a scatter separability criterion to evaluate the quality of the selected feature subset. This is a well known measure of class separability introduced by Devijver and Kittler [4], and given by  $J(Y) = \frac{|S_w + S_b|}{|S_w|} = \prod_{k=1}^d (1 + \lambda_k)$ , where  $Y$  denotes the feature set,  $\lambda_k$ ,  $k = 1 \dots d$ , are the eigenvalues of matrix  $S_w^{-1}S_b$ , and  $S_w$  and  $S_b$  are the between and within class scatter matrices.

**Table 3.** J value comparisons for two methods on three data sets

Method	Australian	Breast cancer	Pima
MII	2.2832	5.0430	1.3867
HGplusMII	2.3010	5.1513	1.3942

In Table. 3, we compare the the performance of the two methods. We find that the effective feature subsets can be obtained using our proposed HGplusMII method, e.g., for dataset Australian and Pima, it can achieve a higher discriminability power based on fewer features. This means that our feature selection method can guarantee the optimal feature subset, as it not only achieves higher degree of the dimensionality reduction but also obtains better discriminability power.

After obtaining the discriminating features, we apply a variational EM algorithm to learn Gaussian mixture model on the selected feature subset for the purpose of classification. For the Breast Cancer dataset, we visualize the classification results using the selected feature subset. The classification accuracy achieved using the selected feature subset is 96.3% which is superior to the ac-



**Fig. 2.** Classification result visualized on 3rd, 7th and 9th features.

curacy of 95.4% achieved by RD-based method [7]. The classification results are shown in Fig. 2. The left hand panel is the data with correct labeling, and the right hand panel is the classification results with the misclassified data highlighted. Because of the unsupervised nature of the variational EM algorithm and the Gaussian mixture model, the classification accuracy of 96.3% demonstrates the adequate class separability provided by the selected feature subset.

## 5 Conclusions

This paper has presented a new graph theoretic approach to feature selection. The proposed feature selection method offers two major advantages. First, hypergraph clustering simultaneously considers the significance of both the features and the correlation between features, and therefore the structural information concealed in the data can be more effectively utilized. Second, the MII criteria takes into account high-order feature interactions with the class, overcoming the problem of overestimated redundancy. As a result the features associated with the greatest amount of joint information can be preserved.

## References

- [1] Balagani, S., Phoha, V.: On the Feature Selection Criterion Based on an Approximation of Multidimensional Mutual Information. *IEEE TPAMI* 32(7), 1342–1343 (2010)
- [2] Blum, L., Rivest, L.: Training a 3-Node Neural Network is NP-complete. *Neural Networks* 5(1), 117–127 (1992)
- [3] Covões, T., Hruschka, E., de Castro, L., Santos, Á.: A Cluster-based Feature Selection Approach. *Hybrid Artificial Intelligence Systems* pp. 169–176 (2010)
- [4] Devijver, A., Kittler, J.: *Pattern Recognition: A Statistical Approach*, vol. 761. Prentice-Hall London (1982)
- [5] Guo, B., Nixon, S.: Gait Feature Subset Selection by Mutual Information. *IEEE TSMC, Part A: Systems and Humans* 39(1), 36–46 (2008)
- [6] Kwak, N., Choi, H.: Input Feature Selection by Mutual Information Based on Parzen Window. *IEEE TPAMI* 24(12), 1667–1671 (2002)
- [7] Zhang, F., Zhao, Y.J., Fen, J.: Unsupervised Feature Selection based on Feature Relevance. *ICMLC* 1, 487–492 (2009)