

Mutual Information Criteria for Feature Selection

Zhihong Zhang and Edwin R.Hancock

Department of Computer Science, University of York, UK

Abstract. In many data analysis tasks, one is often confronted with 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 either that features independently influence the class variable or do so only involving pairwise feature interaction. In prior work [18], we have explained the use of a new measure called multidimensional interaction information (MII) for feature selection. The advantage of MII is that it can consider third or higher order feature interaction. Using dominant set clustering, we can extract most of the informative features in the leading dominant sets in advance, limiting the search space for higher order interactions. In this paper, we provide a comparison of different similarity measures based on mutual information. Experimental results demonstrate the effectiveness of our feature selection method on a number of standard data-sets.

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 the feature subset which is most relevant to classification. Mutual information provides a principled way of measuring the mutual dependence of two variables, and has been used by a number of researchers to develop information theoretic feature selection criteria. For example, Batti [1] has developed the Mutual Information-Based Feature Selection (MIFS) criterion, where the features are selected in a greedy manner. Given a set of existing selected features S , at each step it locates the feature x_i that maximize the relevance to the class $I(x_i; C)$. The selection is regulated by a proportional term $\beta I(x_i; S)$ that measures the overlap information between the candidate feature and existing features. The parameter β may significantly affect the features selected, and its control remains an open problem. Peng et al [11] on the other hand, use the so-called Maximum-Relevance Minimum-Redundancy criterion (MRMR), which is equivalent to MIFS with

$\beta = \frac{1}{n-1}$. Yang and Moody's [15] Joint Mutual Information (JMI) criterion is based on conditional MI and selects features by checking whether they bring additional information to an existing feature set. This method effectively rejects redundant features. Kwak and Choi [8] improve MIFS by developing MIFS-U under the assumption of a uniform distribution of information for input features. It calculates the MI based on a Parzen window, which is less computationally demanding and also provides better estimates.

However, there are two limitations for the above MI feature selection methods. Firstly, they assume that each individual relevant feature should be dependent with the target class. This means that if a single feature is considered to be relevant it should be correlated with the target class, otherwise the feature is irrelevant [3]. So only a small set of relevant features is selected, and larger feature combinations are not considered. The second weakness is that most of the methods simply consider pairwise feature dependencies, and do not check for third or higher order dependencies between the candidate features and the existing features. To overcome the above problem, Zhang and Hancock [18] introduce the so called multidimensional interaction information (MII) $I(F; C) = I(f_1, \dots, f_m; C)$ to select the optimal subset of features. The main reason for using $I(F; C)$ as feature selection criterion is that: because $I(F; C)$ is a measure of the reduction of uncertainty in class C due to the knowledge of feature vector $F = \{f_1, \dots, f_m\}$, selecting features that maximize $I(F; C)$, from an information theoretic perspective, translates into selecting those features that contain the maximum information about class C .

In prior work [18], we have proposed a graph-based method to feature selection. In this feature selection scheme, the original features are clustered into different clusters based on dominant-set clustering and each cluster just includes a small set of features. As dominant set clustering can group most of the informative features into the leading dominant set based on suitable similarity measure, this allows us to limit the search space for further feature selection. The similarity measure used for clustering is based on mutual information. We compare the similarity measure with other two well known alternative measures of similarity, namely Pearson's correlation coefficient (ρ) which based on distance and the Least square regression error (e) is made. Using the Parzen window for probability distribution estimation, we then apply a greedy strategy to incrementally select the features that maximizes the multidimensional mutual information between the already selected features and the output class set.

2 Dominant-Set Clustering Algorithm

There are several different methods for clustering features, well-known examples are: k-means algorithm [9] is built for all sample, but requires a user to supply the number of clusters in advance. In addition, it can not detect clusters of arbitrary shapes. The Self Organizing Map(SOM) [14] is a type of artificial neural network which can produce a low-dimensional space for the input data objects using a neighborhood function to cluster nodes. As same with k-means

algorithm, it does not explicitly optimize any measure of the total dissimilarity to locate clusters. Again, it requires the number of clusters as user input. In this paper, we use dominant set clustering which is suitable for both subspace and high dimensional data clustering. In addition, it does not require the user to provide the number of clusters and can also handle outliers efficiently. Most importantly, it can group most of the informative features into cluster based on a suitable similarity measure.

2.1 Concept of Dominant Set

The dominant set[10], is a combinational concept in graph theory that generalizes the notion of a maximal complete subgraph from simple graphs to edge-weighted graphs. In fact, dominant sets turn out to be equivalent to maximal cliques. The definition of the dominant set simultaneously emphasizes internal homogeneity and together with external inhomogeneity. Thus it is can be used as a general definition of a "cluster". To provide an example, assume there are N training samples, each having 5 feature vectors. In order to capture the dominant features from these 5 features (represented as F_1, \dots, F_5), we construct a graph $G = (V, E)$ with node-set V , edge-set $E \subseteq V \times V$ and edge weight matrix W whose elements are in the interval $[0, 1]$. Each vertex represents a feature and the edge between two features represents their pairwise relationship. The weight on the edge reflects the degree of relevance between two features. Therefore, we represent the graph G with the corresponding edge-weight or weighted relevance matrix. In our example, in Fig. 1, features $\{F_1, F_2, F_3\}$ form the dominant set, since the edge weights "internal" to that set (0.6, 0.7 and 0.9) are larger than the sum of those between the internal and external features (which is between 0.05 and 0.25).

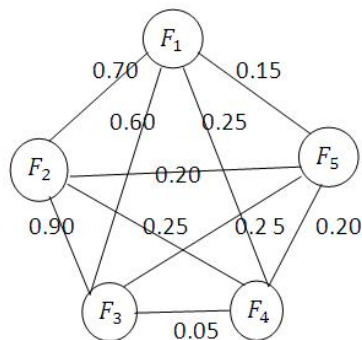


Fig. 1. The subset of features $\{F_1, F_2, F_3\}$ is dominant

For the graph $G = (V, E)$ above, we can locate the dominant set by finding the solutions of a quadratic program that maximizes the functional

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} . \quad (1)$$

subject to $\mathbf{x} \in \Delta$, where $\Delta = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0 \text{ and } \sum_{i=1}^n x_i = 1\}$ and \mathbf{W} is the relevance weight matrix between features. The dominant set corresponds in the strict sense with solutions of the quadratic program. Let u denote a strict local solution of the above program. It has been proved by [10] that $\sigma(u) = \{i | u_i > 0\}$ is equivalent to a dominant set of the graph represented by the edge-weight matrix \mathbf{W} . In addition, the local maximum of $f(u)$ indicates the ‘‘cohesiveness’’ of the corresponding cluster. The replicator equation can be used to solve the program using the iterative update equation:

$$x_i(t+1) = x_i(t) \frac{(\mathbf{W} \mathbf{x}(t))_i}{\mathbf{x}(t)^T \mathbf{W} \mathbf{x}(t)} . \quad (2)$$

where $x_i(t)$ corresponded to the i -th feature vector at iteration t of the update process.

2.2 Dominant-Set Clustering Algorithm

Pavan et al have demonstrated that the concept of a dominant set provides an effective framework for iterative pairwise clustering. Consider a set of features represented by an undirected edge-weighted graph with no self-loops. Let the graph be denoted by $G = (V, E, \omega)$ where $V = 1, \dots, n$ is the vertex set, $E \subseteq V \times V$ is the edge set, and ω is the weight function. Each vertex represents a feature and the weight residing on the edge between two nodes represents the pairwise affinity of the corresponding features. To cluster the features into coherent groups, a dominant set of the weighted graph is iteratively located, and then removed from the graph. This process is repeated until the node-set of the graph is empty. The main property of a dominant set is that the overall similarity among the internal features is greater than that between the external features and the internal features.

3 Feature Similarity Measure

There are different similarity measure methods that can be used for clustering and different methods may lead to different cluster results. As a result, we need to carefully select the most suitable measure to use. In general, the Euclidean distance is widely used as the distance or similarity measure for clustering [7]. However, Euclidean distance only accounts for a data which follows a particular distribution [16], it is not effective to reflect functional similarity such as positive and negative correlation and interdependency. Rao [12] introduced two approaches to measure the linear dependency between variables, namely, a) Pearson’s correlation coefficient (ρ), b) Least square regression error (e).

Pearson’s correlation coefficient (ρ): The Correlation coefficient (ρ) between two random variables x and y is defined as:

$$\rho(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}. \quad (3)$$

where $var()$ denotes the variance of a variable and $cov(x, y)$ is the covariance between two random variables. From the above definition, we can see that Pearson’s correlation coefficient quantifies the linear dependency between two variables x and y . When the $\rho(x, y)$ is large (i.e. 1 or -1), this implies that variable x and variable y are closely related, otherwise, when $\rho(x, y)$ is equal to 0, this means that two variables are totally unrelated. As a result, the method can be used to detect positive and negative correlation. However, there are two limitations which unsuit the utility of Pearson coefficient to used for dominant set clustering. First, it is not robust to outliers and as a result it may assign a high similarity score to a pair of dissimilar features. Second, as it is sensitive to rotation and invariant to scaling, the two pairs of variables having different variances may give the same value of the similarity measure.

Least square regression error (e): The dependency of two variables x and y can be modeled by the linear model, $y = a + bx$. As a result, the degree of dependency between them can be measure by the error in predicting y from the linear model. The parameters of the model a and b can be learned by minimizing the mean square error as follows:

$$e(x, y)^2 = \frac{1}{n} \sum (e(x, y)_i)^2. \quad (4)$$

where $e(x, y)_i = y_i - a - bx_i$, $a = \bar{y}$, $b = \frac{cov(x, y)}{var(x)}$ and $e(x, y) = var(y)(1 - \rho(x, y)^2)$. From this definition, we can see that the least square regression error (e) quantifies the amount of variance of y unexplained by the linear model. As with Pearson’s correlation coefficient (ρ), it is sensitive to rotation and scaling.

4 Feature Selection Using Dominant-Set Clustering

In this paper we aim to utilize the dominant-set clustering algorithm for feature selection. Using a graph representation of the features, there are three steps to the algorithm, namely a) computing the relevance matrix $\mathbf{W} = (\mathbf{w}_{ij})_{n \times n}$ based on the mutual information between feature vectors, b) dominant-set clustering to cluster the feature vectors and c) selecting the optimal feature set from leading dominant set 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.

4.1 Computing the Similarity Matrix

Instead of using the Euclidean distance, Pearson’s correlation coefficient (ρ) or the least square regression error (e), our similarity measure employs an mutual

information measure to evaluate the interdependence of features. The use of this mutual information measure allows dominant set clustering to discover the informative features and group them into cluster. In accordance with Shannon's information theory [13], 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:

$$H(Y) = - \sum_{y \in Y} P(y) \log P(y) . \quad (5)$$

where $P(y)$ is the prior probability density function over Y . The remaining uncertainty in the variable Y if the variable X is known is defined by the conditional entropy $H(Y|X)$

$$H(Y|X) = - \int_x p(x) \left\{ \sum_{y \in Y} p(y|x) \log p(y|x) \right\} dx . \quad (6)$$

where $p(y|x)$ denotes the posterior probability for variable Y given another random variable X . After observing the variable vector x , the amount of additional information gain is given by the mutual information (MI)

$$I(X;Y) = H(Y) - H(Y|X) = \sum_{y \in Y} \int_x p(y,x) \log \frac{p(y,x)}{p(y)p(x)} dx . \quad (7)$$

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 and variable Y are closely related, otherwise, when $I(X;Y)$ is equal to 0, this means that two variables are totally unrelated. Therefore, in our feature selection scheme, the relevance of pairs of feature vectors is computed using mutual 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 , so we can represent the k^{th} feature vector for the N training samples as the long vector $F_k = \{f_k^1, f_k^2, \dots, f_k^N\}$. The entropy of the feature vector F_k where $(k = 1, 2, \dots, K)$ can be computed using Equation (3). For two feature vectors F_{k1} and F_{k2} , their mutual information $I(F_{k1}, F_{k2})$ can be computed by Equation (5). The relevance degree between two feature vectors F_{k1} and F_{k2} can be defined as [17]:

$$\mathbf{W}(F_{k1}, F_{k2}) = \frac{2I(F_{k1}, F_{k2})}{H(F_{k1}) + H(F_{k2})} . \quad (8)$$

where $k1, k2 \in K$ and the higher the value of $\mathbf{W}(F_{k1}, F_{k2})$ the more relevant are the features F_{k1} and F_{k2} . Otherwise, if $\mathbf{W}(F_{k1}, F_{k2}) = 0$, the two features are

totally unrelated. In addition, for the above computation, we use the Parzen-Rosenblatt window method to estimate the probability density function of random variables F_{k1} and F_{k2} [11]. 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 Σ 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 $d = 2$, $p(x)$ estimates the joint density of variables such as F_{k1} and F_{k2} .

4.2 Dominant-set Clustering

The dominant-set clustering algorithm commences from the relevance matrix and iteratively bi-partitions the features into a dominant set and a non-dominant set. It therefore produces the dominant-set progressively and hierarchically. The clustering process stops when all the features are grouped into one of the dominant-sets. We can formulate the dominant-set clustering algorithm in the following: a) Initialize \mathbf{W}^t by the similarity matrix \mathbf{W} , where $t = 1$. b) Calculate the local solution of Equation(1) by Equation(2): u^t and $f(u^t)$. c) Get the dominant set: $DS^t = \sigma(u^t)$. d) Split out DS^t from \mathbf{W}^t and get a new similarity matrix \mathbf{W}^{t+1} . e) If \mathbf{W}^{t+1} is not empty, $\mathbf{W}^t = \mathbf{W}^{t+1}$ and $t = t + 1$, then go to step b; else exit

4.3 Selecting Key Features

In accordance with Shannon's information theory [13], 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:

$$H(Y) = - \sum_{y \in Y} P(y) \log P(y) . \quad (9)$$

where $P(y)$ is the prior probability density function over Y . The remaining uncertainty in the variable Y if the variable X is known is defined by the conditional entropy $H(Y|X)$

$$H(Y|X) = - \int_x p(x) \{ \sum_{y \in Y} p(y|x) \log p(y|x) \} dx . \quad (10)$$

where $p(y|x)$ denotes the posterior probability for variable Y given another random variable X . After observing the variable vector x , the amount of additional information gain is given by the mutual information (MI)

$$I(X; Y) = H(Y) - H(Y|X) = \sum_{y \in Y} \int_x p(y, x) \log \frac{p(y, x)}{p(y)p(x)} dx. \quad (11)$$

In addition, for the above computation, we use Parzen-Rosenblatt window method to estimate the probability density function of random variables F_{k1} and F_{k2} [11]. 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 Σ 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 $d = 2$, $p(x)$ estimates the joint density of variables such as F_{k1} and F_{k2} .

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

$$I(F; C) = I(f_1, \dots, f_m; C) = \sum_{f_1, \dots, f_m} \sum_{c \in C} P(f_1, \dots, f_m; c) \times \log \frac{P(f_1, \dots, f_m; c)}{P(f_1, \dots, f_m)P(c)}. \quad (12)$$

The main reason for using $I(F; C)$ as a feature selection criterion is that: because $I(F; C)$ 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 [8]. 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

$$I(F; C) \approx \hat{I}(F; C) = \sum_i I(f_i; C) - \sum_i \sum_{j>i} I(f_i; f_j) + \sum_i \sum_{j>i} I(f_i; f_j|C). \quad (13)$$

By using $\widehat{I}(F; C)$ instead of $I(F; C)$, it is possible to locate a subset of informative features by implementing a greedy “pick-one-feature-at-a-time” selection procedure. Given K features, out of which m are to be selected ($m < K$), this involves two steps: 1) select the first feature f'_{max} that maximizes $I(f'; C)$, and 2) select $m - 1$ subsequent features that maximize the criterion in Equation (8), i.e., select the second feature f''_{max} that maximizes $I(f''; C) - I(f''; f'_{max}) + I(f''; f'_{max}|C)$, select the third feature f'''_{max} that maximizes $I(f'''; C) - I(f'''; f'_{max}) - I(f'''; f''_{max}) + I(f'''; f'_{max}|C) + I(f'''; f''_{max}|C)$ and so on.

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

$$P(f_i, \dots, f_m) = P(f_1)P(f_2|f_1) \times P(f_3|f_2, f_1) \cdots P(f_m|f_1, f_2 \dots f_{m-1}), \quad (14)$$

$$\begin{aligned} P(F; C) = P(f_1, \dots, f_m; C) &= P(C)p(f_1|C)P(f_2|f_1, C)P(f_3|f_1, f_2, C) \\ &\times P(f_4|f_1, f_2, f_3, C) \cdots P(f_i|f_1, \dots, f_m, C). \end{aligned} \quad (15)$$

In our feature selection scheme, the original features are clustered into different dominant-sets based on dominant-set clustering and each dominant-set just includes a small set of features. Therefore, for each dominant set, 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 dominant set, we repeat this procedure until $|S| = k$.

5 Classification

After finding the discriminating features, we apply the variational EM (VBEM) algorithm [2] to fit a mixture of Gaussians model to the selected feature subset. After learning the mixture model, we use the a posteriori probability, see Equation(16), to classify sample. Given a sample, we first compute its selected feature vector b through feature selection. Then we compute its a posteriori probabilities r_c , the mean vectors \hat{b}_c , and the precision matrices A_c , where $c \in c_1, \dots, c_l$ and

l is the number of class for the data. For example, in binary class, if $r_{c_1} > r_{c_2}$ then the sample is classified as class c_1 . Otherwise, the sample is classified as c_2 . The posterior probabilities are given by

$$r_{nk} \propto \pi_k |A_k|^{-\frac{1}{2}} \exp\left\{\frac{1}{2}(x_n - \mu_k)^T A_k (x_n - \mu_k)\right\}. \quad (16)$$

where $k = 1, \dots, K$ is the mixture component, $n = 1, \dots, N$ denotes the data index. Model parameters π_k , μ_k and A_k are respectively a priori probability, the mean of selected feature vectors and precision matrices of the k^{th} component. In the variational Bayesian EM (VBEM) algorithm, all of these model parameters are characterized by hyper-parameters, which take into account the uncertainty in the parameter estimation. The parameters r_{nk} are called posteriori probability because they represent the responsibility the k^{th} component takes in explaining the n^{th} observation. The posteriori probability can be arranged into a matrix $R = (r_{nk})$ and will have to satisfy the following conditions:

$$0 \leq r_{nk} \leq 1. \quad (17)$$

6 Experiments and Comparisons

The data sets used to test the performance of our proposed algorithm are the benchmark data sets from the NIPS 2003 feature selection challenge and the UCI Machine Learning Repository. Table. 1 summarizes the properties of these data-sets. Our proposed feature selection method (referred to as the DS*plus*MII method) (which utilizes the multidimensional interaction information (MII) criterion and dominant-set clustering for feature selection) involves grouping a set of informative features into cluster from the original feature set by dominant-set clustering and then applying MII criterion into the cluster for feature selection. In order to examine the performance of our proposed method DS*plus*MII, we need to know how meaningful the cluster obtained based on mutual information is and what more useful information they contain. In view of this, we should first examine how discriminative the features in the leading dominant set. Next, we could use the extracted features for classification to check the performance. Our proposed scheme for evaluation and comparison can be outlined as follows: a) the study of the cluster performance obtained by different similarity measure methods(i.e., the Pearson's correlation coefficient (ρ) and Least square regression error (e)). b) the study of classification results based on the selected feature subset captured by MII in the dominant sets and compared with other MI-based criterion methods(i.e., the MRMR algorithm [11] and the MIFS algorithm [1]).

6.1 Cluster Performance Evaluation using Different Similarity Measures

As we mentioned before, our proposed algorithm is capable of grouping informative features in the leading dominant set by dominant set clustering based on

Table 1. Summary of UCI and NIPS benchmark data sets

Data-set	Examples	Features	Classes
Madelon	2000	500	2
Breast cancer	699	10	2
Pima	768	8	2
Australian	690	14	2

Table 2. J value comparisons of dominant set using different feature similarity measure

Data-set	Similarity Measure:MI	Similarity Measure: (ρ)	Similarity Measure: (e)
Madelon	1.1082	1.0024	1.0094
Breast cancer	5.1513	5.1513	5.1513
Pima	1.3716	1.3716	1.0177
Australian	2.2546	2.2006	1.2090

a suitable similarity measure. Different similarity measures will lead to different clustering results, which means that an unsuitable similarity measure may group less informative features into a cluster. Therefore, we should carefully select which similarity measure to use. Here, we study the clustering results obtained by using three different similarity measures for dominant-set clustering(DS). In order to examine the discriminability of the features grouped in the leading dominant set, we will use the multidimensional interaction information (MII) criterion. Then, a criterion function is used to measure the discrimination of the selected key features. This is a well known measure of class separability introduced by Devijver and Kittler [4], and given by

$$J(Y) = \frac{|S_w + S_b|}{|S_w|} = \prod_{k=1}^d (1 + \lambda_k) . \quad (18)$$

where Y denotes the feature set, λ_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. 2 shows the comparative cluster results of our mutual information based similarity measure with other two similarity measures in terms of the measured J value. The subset obtained by our mutual information based similarity measure is more discriminative, giving the highest J value.

6.2 Classification Results using Selected Feature Subset

After obtaining the discriminating features, we apply a variational Bayesian EM(VBEM) algorithm to learn a Gaussian mixture model on the selected feature subset for the purpose of classification. We compare classification results from our proposed feature selection method (referred to as the DS*plus*MII method) (which utilizes the multidimensional interaction information (MII) criterion and dominant-sets for feature selection) with those obtained using k-means algorithm

[9] and alternative existing MI-based criterion methods, namely a) Maximum-Relevance Minimum-Redundancy (MRMR), b) Mutual Information Based Feature Selection (MIFS).

Based on the feature subsets selected by our proposed *DSplusMII* method, We first examine the classification performance using different sized feature subsets by selecting the top k features ranked by their incremental gain. In the classification performance evaluation process, we employ a posteriori probability, see Equation(16), to perform classification, we got the classification accuracy by the percentage of the data, which are predicted correctly. For the purpose of comparison, we repeated the feature selection process using the k-means algorithm, MRMR algorithm and MIFS algorithm.

The Madelon data set is a 2 classes problem originally proposed in the NIPS'2003 feature selection challenge [6]. The data points grouped into 32 clusters placed on the vertices of a five dimensional hypercubes. As a result, there are only 5 informative features, but 15 redundant features and 480 probes. In Fig. 2, we present the top 14 features ranked by the incremental gain calculated by MII. The classification accuracies obtained on different feature subsets are shown in the right hand side of Fig. 2. From the figure, it is clear that using the leading 6 features (476, 339, 379, 154, 443, 456), we achieve 90% classification accuracy. Because of the unsupervised nature of the VBEM algorithm and the gaussian mixture model, the classification accuracy of 90% demonstrates the adequate separability provided by the selected feature subset. For comparison, we also visualize the classification results of using the feature subset obtained by MRMR;

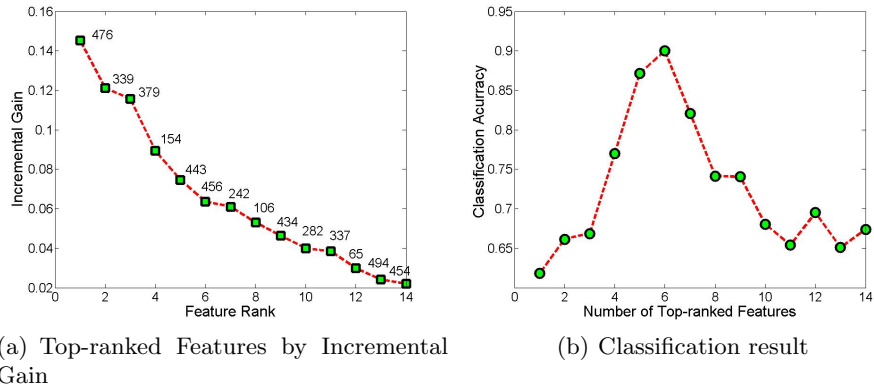


Fig. 2. The result on Madelon data set for our algorithm. The values of the Incremental gain for the top 14 features are presented in the left part along with the feature indices, while the classification accuracies are plotted in the right part

In Fig. 3, the top-ranked features ranked by MRMR are presented in the left hand part, and the classification accuracies using the top-ranked features

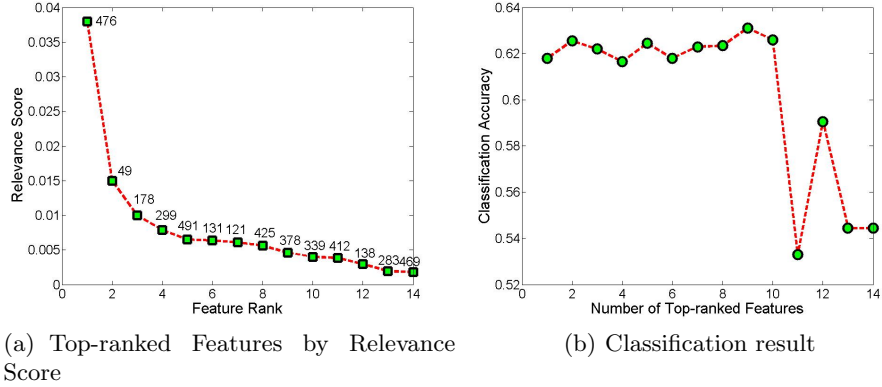


Fig. 3. The result on Madelon data set using MRMR for feature ranking. The values of the relevance score for the top 14 features are presented in the left part along with the feature indices, while the classification accuracies are plotted in the right part

incrementally are presented in the right hand part. The best result is about 63.1% using 9 features, which is much worse than the result of our algorithm as shown in Fig. 2. The poor classification performance may be explained by our observation that most of the selected top features are not in the 1st dominant set and ranked very low by *DSplusMII*. On the other hand, we find that for MRMR there is a tendency to overestimate the redundancy between features, since they neglect the conditional redundancy term $I(x_i, S|C)$. As a result some important features can be discarded, which in turn leads to information loss.

Table 3. The classification accuracy on the top features selected by different methods in the Breast Cancer data set

No.of Features Selected	<i>DSplusMII</i>	MRMR	MIFS
2	88.84%	88.84%	88.84%
3	96.3%	87.98%	84.4%
4	96.3%	87.55%	82.51%

Table 4. The classification accuracy on the top features selected by different methods in the Pima data set

No.of Features Selected	<i>DSplusMII</i>	MRMR	MIFS
2	74.09%	74.09%	74.09%
3	75.91%	75.91%	75.91%
4	72.79%	70.31%	70.31%

Table 5. The classification accuracy on the top features selected by different methods in the Australian data set

No.of Features Selected	DS <i>plus</i> MII	MRMR	MIFS
3	83.77%	68.84%	64.35%
4	83.77%	69.13%	64.35%
5	83.77%	69.28%	83.62%

The experimental results in Table. 3, 4 and 5 show that DS*plus*MII is, by and large, superior to the other feature clustering and feature selection methods by selecting a smaller set of discriminative features than the others as reflected by the classification results. As shown by the results, DS*plus*MII outperforms MIFS and MRMR algorithms in all cases except in the Pima dataset, in which all the four methods yield a comparable classification rate. It is interesting to note that the performance achieves a 96.3% when using the 3 features selected by DS*plus*MII and maintain at the same accuracy even when more features are selected(see Table. 3). Similarly, 83.77% is achieved when 3 features are selected by DS*plus*MII and its performance remains at this level even when more features are selected(see Table. 5). This implies that the discriminative information exists in a small set of features which can be used to fit the mixture Gaussian models to the data. In addition, in breast cancer, we find out that the leading 4 selected features are all from the first dominant set found by dominant set clustering. This again supports the fact that the first dominant set captures the greatest number of informative features. From Table. 4, it is clear that using the leading three features, then all the four methods achieve 75.91% classification accuracy, which is higher than that obtained using other sized feature subsets. Using fewer or more features both deteriorate the accuracy. This implies that classification of samples is based on a very few of the most important features.

7 Conclusions

This paper has presented a new graph theoretic approach to feature selection. The proposed feature selection method offers two major advantages. First, dominant-set clustering can capture the most informative features based on MI-based similarity measure. Second, the MII criteria takes into account high-order feature interactions, overcoming the problem of overestimated redundancy. As a result the features associated with the greatest amount of joint information can be preserved.

References

1. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks* 5(4), 537–550 (2002)
2. Bishop, C.: *Pattern Recognition and Machine Learning*, vol. 4. Springer New York (2006)

3. Cheng, H., Qin, Z., Qian, W., Liu, W.: Conditional Mutual Information Based Feature Selection. In: IEEE International Symposium on Knowledge Acquisition and Modeling. pp. 103–107 (2008)
4. Devijver, P., Kittler, J.: Pattern Recognition: A Statistical Approach, vol. 761. Prentice-Hall London (1982)
5. Guo, B., Nixon, M.: Gait Feature Subset Selection by Mutual Information. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 39(1), 36–46 (2008)
6. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.: Feature extraction, foundations and applications (2006)
7. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering 16(11), 1370–1386 (2004)
8. Kwak, N., Choi, C.: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE TPAMI 24(12), 1667–1671 (2002)
9. MacQueen, J., et al.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. California, USA (1967)
10. Pavan, M., Pelillo, M.: A New Graph-Theoretic Approach to Clustering and Segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1 (2003)
11. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1226–1238 (2005)
12. Rao, C.: {Linear statistical Inference and Its Applications} (1965)
13. Shannon, C.: A Mathematical Theory of Communication. ACM SIGMOBILE Mobile Computing and Communications Review 5(1), 3–55 (2001)
14. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T.: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America 96(6), 2907 (1999)
15. Yang, H., Moody, J.: Feature Selection Based on Joint Mutual Information. In: Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis. pp. 22–25 (1999)
16. Yu, J., Tian, Q., Amores, J., Sebe, N.: Toward robust distance metric analysis for similarity estimation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 316–322. IEEE (2006)
17. Zhang, F., Zhao, Y., Fen, J.: Unsupervised Feature Selection based on Feature Relevance. In: International Conference on Machine Learning and Cybernetics. vol. 1, pp. 487–492 (2009)
18. Zhang, Z., Hancock, E.: A graph-based approach to feature selection. Graph-Based Representations in Pattern Recognition pp. 205–214 (2011)