

A Graph-based Approach to 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. 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. To this end we derive a new hyper graph-theoretic approach to feature selection based on dominant-set clustering. We base our feature selection criterion on the multidimensional interaction information (MII). 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 select the features that are most relevant to classification while reducing redundancy. 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 [7] 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 [9] 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 [5] 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 [2]. 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, we 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 .

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

The primary reason for using the approximation $\hat{I}(F; C)$ 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 for features using large training samples. To overcome this problem, in this paper, we propose a graph-based approach to feature selection. In this 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 $\hat{I}(F; C)$. Instead we can directly use the multidimensional interaction information $I(F; C)$ criterion for feature selection. Using the Parzen window for probability distribution estimation, we then apply a greedy strategy to incrementally select the feature that maximizes the multidimensional mutual information between the already selected features and the output class set.

2 Dominant-Set Clustering Algorithm

Concept of Dominant Set: The dominant set[6], is a combinational concept in graph theory that generalizes the notion of a maximal complete sub-

graph 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 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).

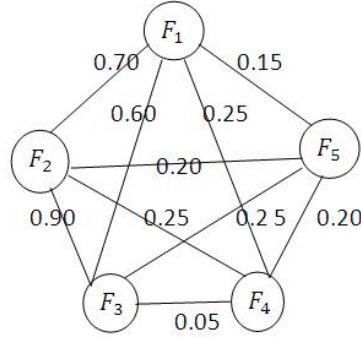


Fig. 1. The subset of features $\{F_1, F_2, F_3\}$ is dominant

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 [6] 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 $x_i(t)$ is correspondent to the i -th feature vector at iteration t of the update process.

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 Selection Using Dominant-Set Clustering

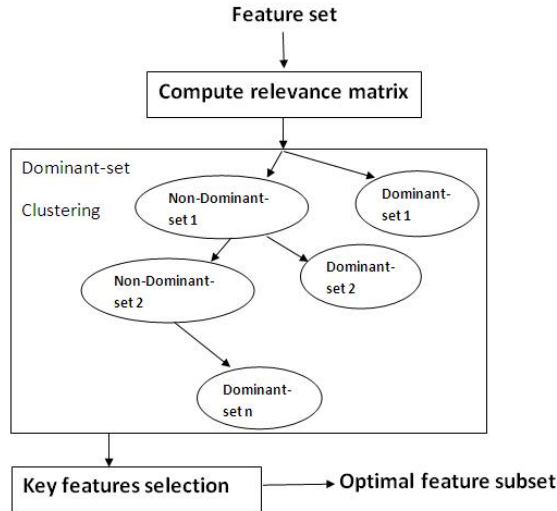


Fig. 2. The flowchart of our approach for feature selection

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 each dominant set using the multidimensional interaction information (MII) criterion. Fig. 2 shows a schematic view of the proposed method for feature selection. 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 [8], 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:

$$H(Y) = - \sum_{y \in Y} P(y) \log P(y) . \quad (3)$$

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 $H(Y|X)$

$$H(Y|X) = - \int_x p(x) \left\{ \sum_{y \in Y} p(y|x) \log p(y|x) \right\} dx . \quad (4)$$

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)

$$I(X;Y) = H(Y) - H(Y|X) = \sum_{y \in Y} \int_x p(y,x) \log \frac{p(y,x)}{p(y)p(x)} dx . \quad (5)$$

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 [10]:

$$\mathbf{W}(F_{k1}, F_{k2}) = \frac{2I(F_{k1}, F_{k2})}{H(F_{k1}) + H(F_{k2})}. \quad (6)$$

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 Parzen-Rosenblatt window method to estimate the probability density function of random variables F_{k1} and F_{k2} [7]. 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} .

Dominant-set Clustering: As illustrated in Fig. 2, 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.

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

$$I(F; C) = I(f_1, \dots, f_m; C) = \sum_{f_1, \dots, f_m} \sum_{c \in C} P(f_1, \dots, f_m; c) \times \log \frac{P(f_1, \dots, f_m; c)}{P(f_1, \dots, f_m)P(c)}. \quad (7)$$

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 [5]. 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 [4] to approximate $I(F; C)$, and this is arguably the most simple yet effective evaluation criterion for selecting features. The approximation is given as

$$\begin{aligned}
I(F; C) \approx \widehat{I}(F; C) &= \sum_i I(f_i; C) - \sum_i \sum_{j>i} I(f_i; f_j) \\
&+ \sum_i \sum_{j>i} I(f_i; f_j | C) . \tag{8}
\end{aligned}$$

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

$$\begin{aligned}
P(f_i, \dots, f_m) &= P(f_1)P(f_2|f_1)P(f_3|f_2, f_1) \times P(f_4|f_3, f_2, f_1) \cdots P(f_m|f_1, f_2 \dots f_{m-1}) \\
P(F; C) &= P(f_1, \dots, f_m; C) = P(C)p(f_1|C)P(f_2|f_1, C)P(f_3|f_1, f_2, C) \\
&\times P(f_4|f_1, f_2, f_3, C) \cdots P(f_i|f_1, \dots, f_m, C) . \tag{10}
\end{aligned}$$

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 $|S| = k$.

4 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. Using the feature selection algorithm outlined above, we make a comparison between our proposed feature selection method (referred to as the *DSplusMII* method) (which utilises the multidimensional interaction information (MII) criterion and dominant-sets for feature selection) and the use of multidimensional interaction information (MII) using the second-order approximation, see Equation (8).

Data-set	From	Examples	Features	Classes
Madelon	NIPS	2000	500	2
Breast cancer	UCI	699	10	2
Pima	UCI	768	8	2

Table 1. Summary of UCI and NIPS benchmark data sets

The experimental results shown in Table. 2 demonstrate that at small dimensionality, i.e. with the Breast cancer data-set(10 features and 699 examples) and the Pima data-set (8 features and 768 examples), the feature subset selected using our proposed method (i.e. *DSplusMII*) is consistent at least to some degree with those obtained using MII with second-order approximation. However, at higher dimensionality (e.g. the Madelon data set with 500 features and 2000 examples), there is a significant difference between the selected feature subsets. There are three reasons for this. The first reason is that dominant-set clustering focuses on the information-contribution of each feature, so the most informative features can be extracted. The second reason is that the multidimensional interaction information (MII) criterion is applied to each dominant set 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 multidimensional interaction information (MII) by second-order approximation simply checks for pair-wise dependencies between features and the class, and so only limited feature subsets can be obtained. When the database is large, our proposed method *DSplusMII* shows its advantage.

To illustrate the dominant-set clustering process for feature extraction in more detail, we list the dominant sets for Pima and Breast cancer data-set in Table. 3. By inspection, we can see that the first dominant set includes most of the important features. For example, in the Breast cancer data-set, the final selected features $\{f_3, f_7\}$ are all from the first dominant-set, the second dominant-set provides no further information relevant to the classification process. For the Pima data-set, most of the final selected informative features are also from the first dominant set. This reveals the advantage of our dominant-set based feature

Method	Madelon	Breast cancer	Pima
MII	{ $f_{476}, f_{49}, f_{178}, f_{131}, f_{491}, f_{299}, f_{283}, f_{121}, f_{425}, f_7, f_{385}, f_{216}, f_{458}, f_{237}, f_{310}, f_{366}, f_9, f_{499}, f_{54}, f_{346}, f_{198}, f_{368}$ }	{ f_3, f_7 }	{ f_2, f_8, f_6, f_7 }
DSplusMII	{ $f_{476}, f_{379}, f_{49}, f_{330}, f_{412}, f_{137}, f_{11}, f_{256}, f_{135}, f_{56}, f_{138}, f_{283}, f_{324}, f_{425}, f_{467}, f_{62}, f_{455}, f_{472}, f_{208}, f_{206}, f_{169}, f_{424}$ }	{ f_3, f_7 }	{ f_2, f_8, f_6, f_1 }

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

Dominant-sets	Breast cancer	Pima
Dominant set 1	{ $f_3, f_4, f_6, f_7, f_9, f_5, f_8$ }	{ $f_5, f_2, f_4, f_8, f_3, f_6$ }
Dominant set 2	{ f_1, f_2, f_{10} }	{ f_7, f_1 }

Table 3. The dominant sets for Breast cancer and Pima data-set

extraction method. It focuses on the information-contribution of each feature which is capable capturing the greatest number of informative features at a low computation cost. Additionally, it also indicates that not all of the dominant-sets located by dominant-set clustering are significant. It is for this reason that we utilize the multidimensional interaction information (MII) criterion for further feature selection.

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 [3], and given by

$$J(Y) = \frac{|S_w + S_b|}{|S_w|} = \prod_{k=1}^d (1 + \lambda_k) . \quad (11)$$

where Y denotes the feature set, $tr(S)$ is the sum of the diagonal elements of S , $\lambda_k, k = 1 \dots d$, are the eigenvalues of matrix $S_w^{-1}S_b$, and S_w and S_b are the between and within class scatter matrices.

Method	Madelon	Breast cancer	Pima
MII	1.0867	3.7939	1.3867
DSplusMII	1.1082	3.7939	1.3977

Table 4. J value comparisons for two methods on three data sets

In Table. 4, we compare the the performance of the two methods. At small dimensionality there is little difference between the two methods. However, at higher dimensionality, the features selected by our proposed *DSplusMII* method are superior to the features selected by MII based on the second-order approximation. This means that our proposed *DSplusMII* feature selection method can guarantee the optimal feature subset, as it not only focuses on the information-contribution of each feature but also considers its contribution to class.

5 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. 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. 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)
3. Devijver, P., Kittler, J.: *Pattern Recognition: A Statistical Approach*, vol. 761. Prentice-Hall London (1982)
4. 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)
5. Kwak, N., Choi, C.: Input Feature Selection by Mutual Information Based on Parzen Window. *IEEE TPAMI* 24(12), 1667–1671 (2002)
6. 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. IEEE (2003)
7. 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)
8. Shannon, C.: A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1), 3–55 (2001)
9. 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)
10. 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. IEEE (2009)