Ca' Foscari University

Department of Computer Science

Master's Degree programme - Second Cycle

Final Thesis

Feature Selection using Dominant-Set Clustering

Supervised by

Prof. Marcello Pelillo

Graduand

Nguyen Minh Phu

Matriculation Number: 849289

Academic Year

2014/2015

Acknowledgments

I would like to express my gratitude to my advisor Prof. Marcello Pelillo for supporting my Master's thesis. His guidance helped me much in the completion of the research and writing of this thesis. He has been both an advisor and a mentor for my M.D study. Throughout the research, he assisted me with not only insightful comments and encouragement, but also the hard questions which incented me to widen my research from various perspectives.

I will never forget all my friends: Phan Thanh Tu, Nguyen Duy Tan, Phung Nhu Kien, Mohamed Abd El Hameed, Ismail Elezi, and all my classmates, thank for being with me in my life and your sincere friendship.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout writing this thesis and my life in general.

Abstract

Feature selection techniques are essentially used in the data analysis tasks, one is frequently dealt with many features. It is computationally expensive to optimize these features that are both redundant and irrelevant. A ton of methods approach to this technique, however, they still have their own limitations. In this thesis, dominant-set clustering and multidimensional interaction information (MII) method are considered to select the most informative features from original input features set. MII not only takes into account a pairwise relation, but also examines higher order relations. Utilization of dominant-set clustering, the most informative features are selected in each dominant set. This narrows the space for higher order relation searching. As a result, the increasing of redundancy is eliminated in particular feature combinations. In this thesis, using the above method for feature selection with high dimensional data as well as comparing with other methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Feature selection is the process of picking out a group of features (or variables) that minimizes redundancy and maximize relevance. These features occur in high-dimensional datasets which are a significant challenge for pattern recognition and machine learning. The data set practically exists hundreds, even thousands, features, but many of them are either redundant or irrelevant. They will enhance the overfitting problem and computational burden. There is a ton of methods having been used to narrow these issues. Researchers commonly use mutual information as a good way of measuring the relevance between two features to figure out the feature selection problem. There are a lot of feature selection criteria based on mutual information.

For instance, Battiti [3] suggested Mutual Information-Based Feature Selection (MIFS) Criterion where feature f is selected as the one that maximizes $I(C; f_i) - \beta \sum_{s \in S} I(f, s)$. Peng at al [4] developed the Maximum-Relevance Minimum-Redundancy(MRMR) criterion which is equivalent with MIFS with $\beta = \frac{1}{n-1}$. Yang and Moody's [5] proposed Joint Mutual Information (JMI) finding the information between a joint random variable and the targets, given by $\sum_{k=1}^{n-1} I(X_n X_k; Y = I(X_n, Y) - \frac{1}{n-1} \sum_{k=1}^{n-1} [I(X_n, X_k) - I(X_n; X_k|Y)]$. Kwak and Choi [6] proposed an improvement to MIFS, called MIFS-U. It is more suitable to solve the problem that the information is distributed uniformly in the input features. The criterion is $I(X_n; Y) - \beta \sum_{k=1}^{n-1} \frac{I(X_k;Y}{H(X_k} I(X_n; X_k)$. In practice, using the optimal value of $\beta = 1$. Fleuret [7] has developed one of the best criteria, based on Conditional Mutual Information Maximization $\min_k [I(X_n; Y|X_k]] = I(X_n; Y) - \max_k [I(X_n; X_k) - I(X_n; X_K|Y)]$. On the other hand, in the Computer Vision literature, Lin and Tang [8] developed Conditional Infomax Feature Extraction; and Vidal-Naquet and Ullman [9] proposed Informative Fragments criterion.

However, there are two considerable limitations among these criteria.

First, the features are only considered as variance. They are completely independent to the target class. In other words, the features will be correlated with target class if it is relevant. Second, the significant limitation is that most criteria only consider the pairwise feature interaction. They do not find out the influence of other features. An advanced solution for the problem above will be introduced in this thesis, so called multidimensional interaction information (MII) [1]. In this case, the dominant-set clustering is used to select subsets of relevant features, thus each cluster has a small set of features. Much feature selection space will be limited, since most informative features can be grouped by dominant-set clustering based on similarity measure. For each dominant set, calculating the multidimensional interaction information between feature vector $F = f_1, ..., f_m$ and target class C by $I(F; c) = I(f_1, f_2, .., f_n; C)$. The Pazen window method will be used to estimate the input distribution. Finally, the method introduced in this thesis will be applied to select the feature that maximizes the multidimensional mutual information.

Chapter 2

Background Concept

2.1 Entropy and Mutual Information

Entropy (introduced by Shannon) is used to measure the unpredictability of information. In other words, how much information of a random variable uncertainty is produced. Suppose a finite set $\mathcal{Y}$ defines a random variable $\mathbf{Y}$ with possible values $\{y_1, ..., y_n\}$ and probability mass function $p(y) = \Pr(\mathbf{Y} = y), y \in \mathcal{Y}$. The normalization condition has to be satisfied:

$$\sum_{y \in \mathbf{Y}} p(y) = 1 \tag{2.1}$$

The entropy H(Y) of the random variable $\mathbf{Y}$ is defined by Shannon as:

$$\mathrm{H(Y)} = -\sum_{y \in \mathcal{Y}} p(y) \log_2 p(y) \tag{2.2}$$

where $y \log_2 y = 0$ when $y \to 0$, and the entropy is denoted in bits.

Intuitively, the entropy is the information which is missed during the process: the larger the entropy, the less a priority information is received from the random variable value. "Note that entropy is a function of the distribution of $\mathbf{Y}$. It does not depend on the actual values taken by a random variable $\mathbf{Y}$, but only on the probabilities" [11].

Let consider the random variables $\mathbf{Y}$ and $\mathbf{X}$ are dependent. The conditional entropy (an amount of the information's unpredictability) is obtained on variable $\mathbf{Y}$ if the variable $\mathbf{X}$ is given. In other words, it is a

measure of the degree of dependence between two such random variables. The conditional entropy H(Y|X) is defined as

$$\mathrm{H(Y|X)} = -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log_2 p(y|x) \tag{2.3}$$

where $p(y|x)$ is the conditional probability distribution for the random variable $\mathbf{Y}$ given random variable $\mathbf{X}$. It can be calculated based on Bayes' theorem:

$$p(y|x) = \frac{p(y,x)}{px} = \frac{p(x|y)p(y)}{px} \tag{2.4}$$

Before presenting mutual information, Let inspect the distance between two distributions, called relative entropy or Kullback-Leibler distance. It is defined as

$$D(p||q) = \sum_{y\in\mathcal{Y}} p(y) \log_2 \frac{p(y)}{q(y)} \tag{2.5}$$

where $p(y)$ and $q(y)$ are two probability mass functions, and there is a convention when $p(y)$ or $q(y)$ is equal zero that $0 \log_2 \frac{0}{0} = 0$, $0 \log_2 \frac{0}{q(y)} = 0$, and $0 \log_2 \frac{p(y)}{0} = \infty$.

Mutual information (MI) of two random variables is a measure of the amount of information that obtained about one random variable based on about another. It is considered as the reduction in certainty about a random variable on account of the knowledge of the other. Higher mutual information, larger reduction in uncertainty or more relevant between two random variables. Lower mutual information, smaller reduction in uncertainty or less related between two random variables Let $p(y,x)$ is a joint probability distribution of two random variables $\mathbf{Y}$ and $\mathbf{X}$, $p(y)$ and $p(x)$ are two marginal probability distributions.

The mutual information $\mathrm{I(Y,X)}$ is defined as the Kullback-Leibler distance between the product of two marginal probability distributions and joint probability distribution:

$$\mathrm{I(Y,X)} = \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(y,x) \log_2 \frac{p(y,x)}{p(y)p(x)} \tag{2.6}$$

The relation between the entropy and mutual information is illustrated

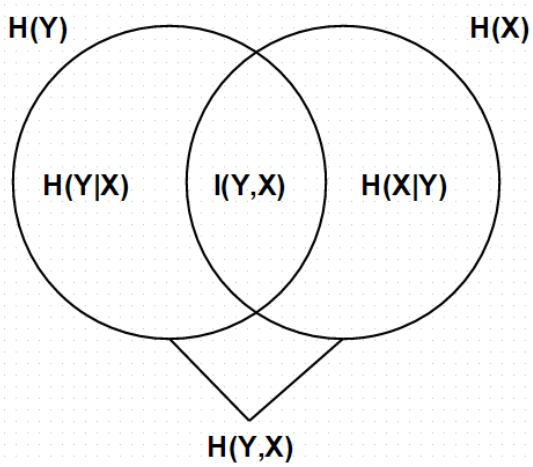


Figure 2.1: Relation between entropy and mutual information

in figure 2.1. Mutual information is a relative entropy, thus it should be greater than or equal zero ($\mathrm{I(Y,X)} \geq 0$). When the equality is present, it indicates that two random variables are completely irrelevant. Since these such properties of mutual information are really useful in selection feature context, and how to apply the mutual information in this prospect will be mentioned in the next chapters.

2.2 Basic Graph-Theoretic Definitions and Concepts

Conceptually, a graph is the relationship between nodes (called vertices) and lines (called edges). These vertices are connected by edges as figure 2.2 shows

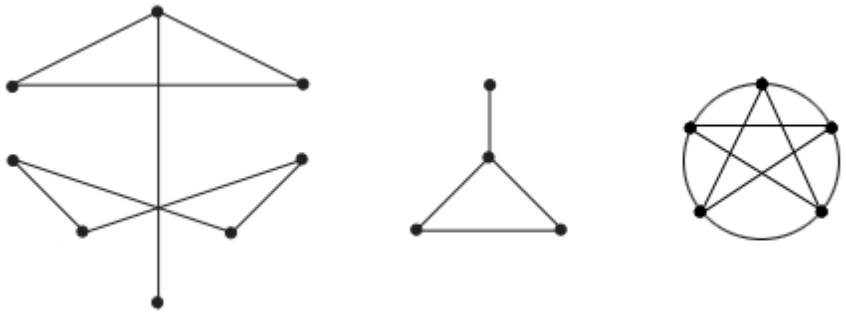


Figure 2.2: A graph example [12]

Let V is a finite set of vertices, and E is a set of edges. A simple graph G is a pair of sets (V, E). In the graph, the vertices and the edges represent

the objects and the relationship among these objects respectively. A measure of edge value is called edge weight.

There are various types of graphs, they depend on the measure of edge weights and the direction of edges. An edge is a bi-directed pairwise vertex (or no direction), called an undirected graph; whereas a directed graph is a graph that there exists an ordered pair of vertices, in other words, a directed edge. When the measure of edge weight is a value of either 0 or 1, called an unweighted graph, put differently, there exists edge or not between a pair of vertices. On the other hand, a graph is given a numerical edge weight, called a weighted graph. There exists an edges between a vertex and itself, this graph is a self-loop graph.

In the context of this thesis, the undirected edge-weighted graph without self-loop will be assumed as a graph $G = (V, E, \omega)$ where V is the finite set, called a vertex set with possible value $\{1, ..., n\}$, $E \subseteq V \times V$ is the edge set, and ω is the weighted function with domain $E \to R^*$. There are several useful concepts which, related to graph theory, should be known to understand more about graph.

Firstly, **a clique** is a subgraph of a simple graph $G = (V, E)$ with a vertex subset $S \subseteq V$ such that every two vertices in S is adjacent each other. It means there exists a maximal complete subgraph where all vertices are connected.

A Maximal Clique of a graph G is a clique that cannot extend or add more vertices. In other words, it is a clique with the largest size (the number of vertices of a graph) given in a graph. The number of vertices in the largest clique of graph G, called the **clique number** of the graph G.

2.3 Cluster analysis

Everitt (1980) stated that "A cluster is a set of entities which are alike, and entities from different clusters are not alike". In other words, cluster analysis (clustering) is a process of dividing objects into groups (clusters) where objects will be relevant if they are contained in the same cluster, whereas they will be irrelevant if objects are outside such cluster or in the other clusters. Cluster analysis is a significant role in numerous areas: machine learning, biology, pattern recognition, statistics, data mining...

The goal of cluster analysis is to partition the given input into meaningful and (or) useful groups from the given input data which can be presented as a $n \times m$ matrix where, m is the number of variables, or feature while n is the number of observation or samples of given data. In general,

there are two variations of the cluster analysis problem, including: central clustering, and pairwise clustering. It depends on what type of data set which need to cluster. For example, figure 2.3 will show clustering using graph-theoretic method.

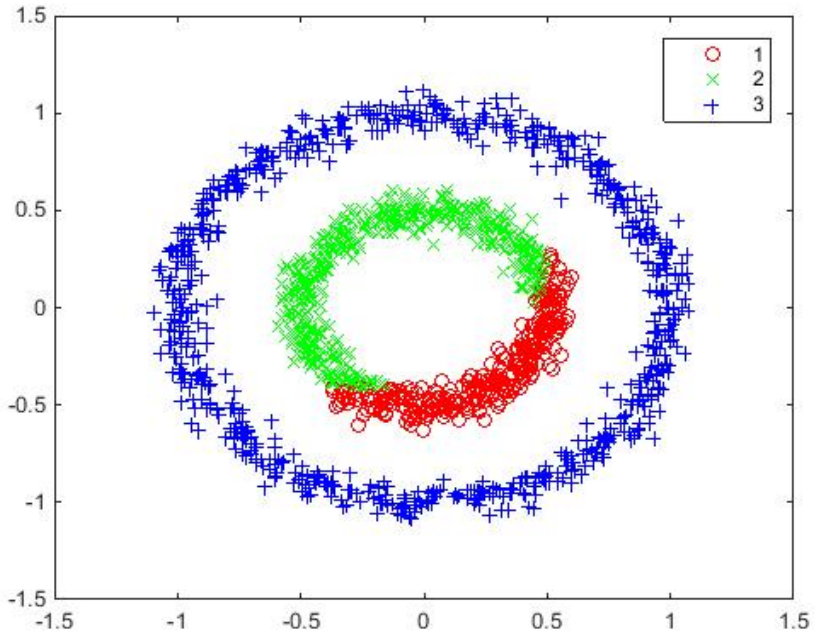


Figure 2.3: An example of graph-theoretic clustering

2.3.1 Central clustering

Central clustering is also known as feature based clustering where the objects, need to be clustered, represented as the feature vector. In other words, a multidimensional space of data points will be clustered in this context. One of the most popular algorithm is used in central clustering is *K-mean algorithm.*

Assume that the data set $X = x_i$ for all $i = 1, .., n$ is the D-dimensional variable with n samples, K is the number of clusters which are grouped from a given data set. Intuitively, the distance between points inside a cluster will be smaller than the distance between inside points and outside points. The goal of *K-mean algorithm* is find the minimizing Euclidean distance, called sum of squared errors (SSE), between variable x_i and mean of k^{th} cluster μ_i

in order to optimize the partitioned cluster. The SSE is defined as

$$SSE = \sum_{i=1}^{n} \sum_{k=1}^{K} ||x_i - \mu_k||^2 \tag{2.7}$$

The basic clustering algorithm for iterative procedure of *K-mean* as belows:

1. Choose the initial k-partition to take action as a cluster centroid.
2. Assign each data point to the nearest cluster.
3. Compute the cluster center based on the current cluster.
4. Repeat step 2 and 3 until the cluster centroids are no change.

As the figure 2.4 shown, the center of black circles is the cluster centroid. It will be recomputed until all objects are clustered, on other words, the centers is stable.

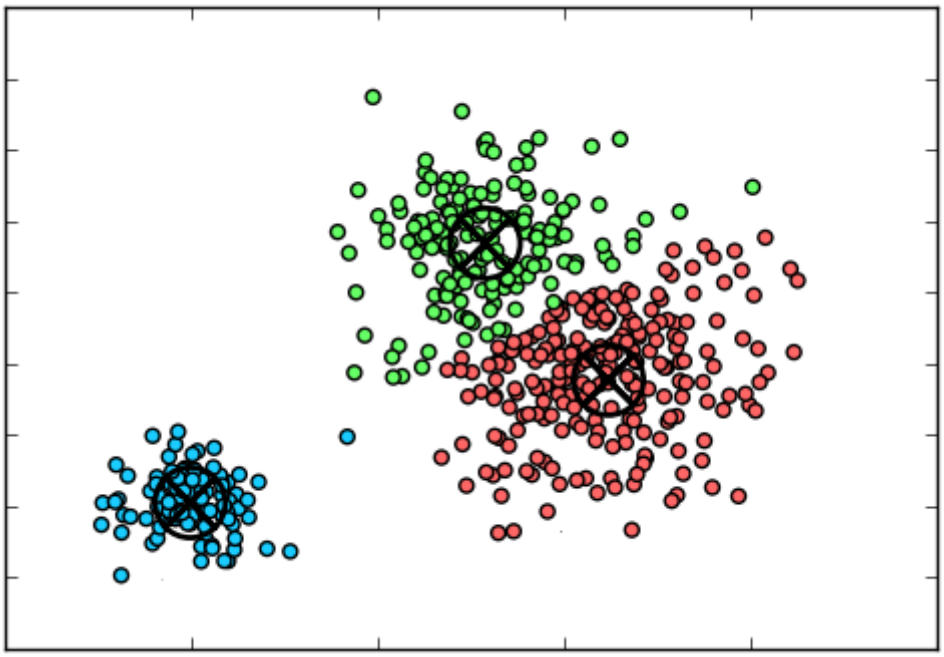


Figure 2.4: A K-mean algorithm example

2.3.2 Graph-theoretic clustering

In many applications, the first step is clustering to extract or select features from the huge data sets. It is useful to cut down problem from the groups of variables in a multidimensional feature space. Such algorithm, *K-mean algorithm*, is the most common method to solve this issue.

However, in other contexts, it is only possible to obtain a pairwise similarity information between objects. For instance, in several fields, the input data just present the relevance between objects. Thus, it is so difficult to apply the *K-mean algorithm* in this case. But there is a possible way to measure the similarity or relevance between data points, in other words, the pairwise clustering method will be utilized in this situation.

A graph-theoretic clustering approach considers the similarity (affinity) matrix as an input data. According to the affinity matrix, the method will base on several criteria to group the input data into different clusters. There exist various algorithms in this situation, including Normalized Cut, and Dominant-set clustering introduced more detail in chapter 3.

Before interpreting dominant-set clustering, Let check out the Normalized Cut [13] method first. In this scheme, suppose the graph G as the graph is presented in section 2.2 with the affinity matrix W. This method will cut the graph G into two subgraph X and Y where, $X \cup Y = V$ and $X \cap Y = \emptyset$, as shown in figure 2.5. The edges, connect X and Y, will be removed and the cost of this process is defined as:

$$cut(X,Y) = \sum_{u \in X, v \in Y} W(u,v) \tag{2.8}$$

The goal is to minimize the subgraph *cut* while the main graph G is grouped. However, minimizing the value of *cut* will cause the isolated node problem. This means that there exist nodes which do not belong G any more. To deal with this problem, the value of *cut* will be divided by the total weight of connections between the nodes X and Y to all the nodes in the graph G, this calculation is called *Normalized Cut* given as below:

$$Ncut(X,Y) = \frac{cut(X,Y)}{asso(X,V)} + \frac{cut(X,Y)}{asso(Y,V)} \tag{2.9}$$

where $asso(X,V) = \sum_{u \in X, v \in V} W(u,v)$, $asso(Y,V) = \sum_{u \in Y, v \in V} W(u,v)$.

Thus the main goal of *Normalized Cut* is to minimize the $Ncut(X,Y)$. When the minimal $Ncut(X,Y)$ is detected, two subgraphs, regions, or cluster X

and Y will be separated where have less edge weight between them and high internal edge weights.

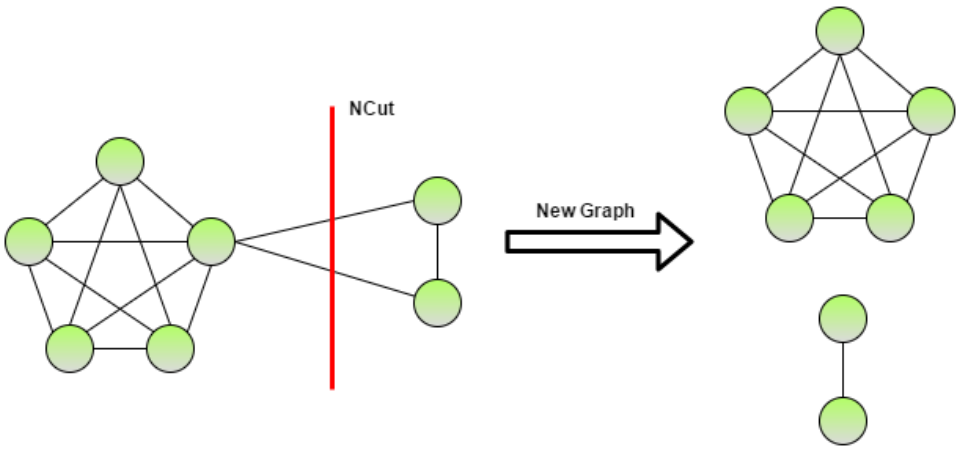


Figure 2.5: A Normalized cut algorithm example

Chapter 3

Dominant-set Clustering

3.1 Overview

In the pairwise data clustering problem, dominant sets [2] are expressed as a graph-theoretic concept generalized as a maximal clique to the graph that is undirected edge-weighted without self-loop as the graph $G = (V, E, \omega)$ in section 2.2. Vertices in the graph represent the neighborhood relationship between two features. An edge-weigh reflects similarity between pairs of connected vertices. And then the graph G represents a similarity matrix $n \times n$ where the value of the matrix is the edge-weights denoted a_{ij}. If there exist edges between two vertices (or i and $j \in E$), such value $a_{ij} = w(i, j)$; Otherwise it is zero. According to the definitions given by Massimiliano Pavan and Marcello Pelillo [2]:

Definition 1:

"*Let $S \subseteq V$ be a non empty subset of vertices and $i \in S$ is a vertex in subset S. The (average) weight degree of a vertex i with respect to S is defined as a sum of edge weights connecting i to all the points in S divided by the cardinality of S. It is denoted by* $\mathrm{awdeg_S(i)}$ *and mathematically defined as:*"

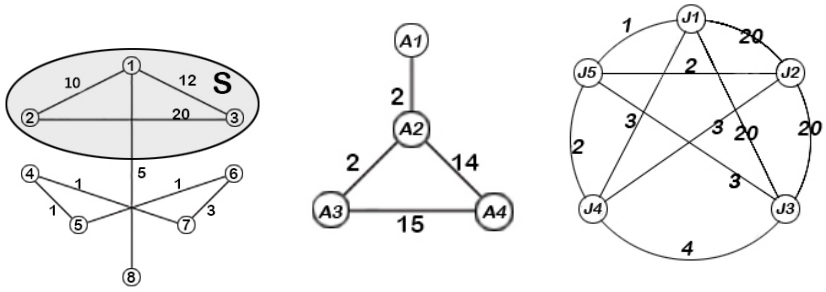
$$\mathrm{awdeg}_S(i) = \frac{1}{|S|} \sum_{j \in S} a_{ij} \tag{3.1}$$

Note that the average weight degree of i is equal to zero if the corresponding subset only contains node i. Furthermore, let consider vertex j is not belong to subset S, the relative similarity between two vertices i and j ($\phi_S(i, j)$), with respect to the average similarity between vertex i and its

neighbors in subset S, is defined as the difference between the edge-weight between two vertices and the average weight degree of vertex i:

$$\phi_S(i,j) = a_{ij} - \text{awdeg}_S(i) \tag{3.2}$$

From the equation 3.2, it is clear that the measure of relative similarity can be positive or negative depend on the value of two terms in the equation.



(a) (b) (c)

Figure 3.1: An example edge-weighted graph

For example, referring figure 3.1, we have

$V = \{(1,2,3,8,), (4,5,6,7), (A1, A2, A3, A4), (J1, J2, J3, J4, J5)\}$

Let S be a subset of V including $\{1,2,3\}$, the weight degree of vertex 1 in subset S is:

$\text{awdeg}_S(1) = \frac{1}{3}(10+12) = 7.33$

Vertex $\{8\}$ is not belong to S, we have the similarity between $\{1\}$ and $\{8\}$:

$\phi_S(1,8) = a_{1,8} - \text{awdeg}_S(1) = 5 - 7.33 = -2.33$

As stated by definition 1, it is possible to allocate a weight to a node. The following recursive definition will interpret this issue in more detail.

Definition 2:

"*Let $S \subseteq V$ be a non-empty subset of vertices and $i \in S$. The weight of i with respect to S is:*"

$$w_S(i) = \begin{cases} 1, & \text{if } |S| = 1. \\ \sum_{j \in S\setminus\{i\}} \phi_{S\setminus\{i\}}(i,j) w_{S\setminus\{i\}}(j), & \text{if } Otherwise. \end{cases} \tag{3.3}$$

Intuitively, the weight of i w.r.t. S is the measure of the similarity between node i and the nodes in subset $S \setminus \{i\}$.

"*Moreover, the total weight of S is defined to be:*"

$$W(S) = \sum_{i \in S} W_s(i) \tag{3.4}$$

Observe that $w_{\{i,j\}}(j) = w_{\{j,i\}}(j) = a_{ij}$, $\forall i, j \in V (i \neq j)$.

For example, as of figure 3.1a, it shows that $w_S(3) > w_S(2)$, and $w_S(2) > w_S(1)$. Figure 3.1c shows that $w_{J1,J2,J3,J4,J5}(J5) < 0$ and $w_{J1,J2,J3,J4,J5}(J2) > 0$. This can be intuitively achieved by observing the amount of edge-weight connected to node $J5$ is considerably smaller than subset $\{J1, J2, J3\}$. Whereas the measure of edge-weight connected to node $J2$ is considerably larger than subset $\{J1, J4, J5\}$.

3.2 Dominant-set clustering

The concept of a dominant set (cluster) will be represented by the following definition.

3.2.1 Definition 3 (Dominant Set):

"*A non-empty subset of vertices $S \subseteq V$ such that $W(T) > 0$ for any non-empty $T \subseteq S$, is said to be* dominant *if:*"

1. $w_S(i) > 0$, $\forall i \in S$.
2. $W_{S \cup \{i\}}(i) < 0$, $\forall i \notin S$.

For example, in figure 3.1c, set $\{J1, J2, J3\}$ is dominant. Definition 3 highlights that dominant set express main cluster properties that the vertices in the same subset are more similar to each other than to the other subsets. This is intuitively obtained by the matter that the edge weights connecting them are considerably higher than the others. This fact is the reason why a dominant set is studied as a cluster.

In the feature selection context [1], suppose that a data set with N observations and 5 feature vectors. For an example on dominant set, the graph $G = (V, E)$ is created with vertex set V including 5 nodes $(\{F_1, ..., F_5\})$, E is the edge set, the similarity matrix W contains the

edge-weight with possible value from 0 to 1. In this case, the edge presents the relative information of pairwise features, the edge-weight between two features represents a measure of relevance between a pairwise feature. As of figure 3.2, it turns out that the subset features $\{F_3, F_4, F_5\}$ is the dominant set because the edge-weighted values inside subset is considerably greater than the total of the amount of edge weights between the inside features and outside features, the value is $0.5, 0.7, 0.9)$ and (from $0.05to0.26$) respectively.

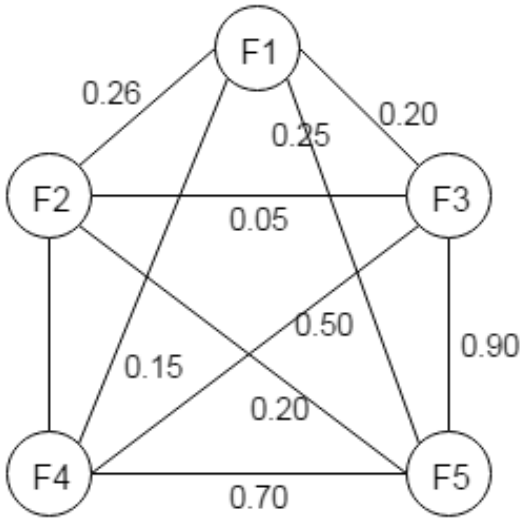


Figure 3.2: The dominant set of subset $\{F_3, F_4, F_5\}$[1]

3.2.2 The Algorithm

Mathematically, the dominant set is located by a quadratic equation of which the solutions are to maximize:

$$f(x) = \frac{1}{2} x^T W x \tag{3.5}$$

subject to $x \in \Delta$

where Δ is a standard simplex, $\Delta = \{x \in R^n : x \geq 0 \text{ and } \sum_{i=1}^{n} x_i = 1\}$, and W is the relevance weight matrix between features.

Solving the quadratic equation requires lots of effort. According to [2], to reduce the complexity, an equivalent formula can be used:

$$x_i(t+1) = x_i(t) \frac{(Wx(t))_i}{x(t)^T W x(t)}. \tag{3.6}$$

where $x_i(t)$ represents the feature vector i at a time t in updating state.

Pavan and Pelillo [2] have proved that the dominant set is very useful in term of clustering based on pairwise affinities between the features. The "*Dominant-set clustering algorithm*" is with the assumption that the input data set of feature vector will be represented by the graph which is without self-loop and bi-directed edge weights as the graph $G = (V, E, \omega)$ in the section 3.1. The algorithm is presented as follows:

- Initial the empty dominant set
- Repeat

  Using replicator dynamics to iteratively find out a dominant set of features (the vertices of the graph)

  Remove these vertices from the graph.
- Until there is no vertex in the graph.

As the results, the clusters will be obtained as dominant sets. As far as discussing above, the features in the same dominant set are highly relevant, whereas they are highly irrelevant between external features and the features outside the dominant set. In other words, the measure of similarity among the feature inside the dominant set is significantly larger than comparable with the total similarity between the outside features and features inside.

Chapter 4

Applying Dominant-Set Clustering On Feature Selection

4.1 An Overview

Before thinking more about the algorithm, let take a look at what is the feature selection (attributes or variable selection). This is a process of picking up features in the given input data, such that these features are the most appropriate for our purposes. In general, feature selection is like dimensionality reduction. They both finding a way to cut down the quantity of features (or columns of a matrix) in the input data set.

However, these methods are mathematically different. In feature selection context, selecting the features is useful and meaningful, in other words, include and exclude features from the input, but they are not changed while in the dimensional reduction field, combining features to build the new one.

Feature selection help to increasing the accuracy of model since the redundant and irrelevant features will be detected and taken out from the input. Thus the model is less complex and more straightforward to comprehend and interpret. Isabelle Guyon [14] stated that:

"The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data."

These features ussually take place in high-dimensional datasets which are a significant challenge for pattern recognition and machine learning.

There are a variety of methods having been used to tackle the high-dimension feature selection. One of the most widely known methods is that PCA which used for dimensional reduction based on subset of input features.

Nevertheless, the sets of features, are extracted by PCA, only account for features variances. Thus this leads to poor performance in classification problem. Therefore, it is significant to consider a method for selecting a small group of informative features to deal with clustering and classification problems.

Recently, Researchers commonly use mutual information as a good way of measuring the relevance between two features to figure out the high-dimension feature selection context. Several popular criteria, are mentioned in the introduction, will be expressed as following:

MIFS

The criterion has been proposed as:

$$J = I(f_j, C) - \beta \sum_{f_v \in V} I(f_v, f_j) \tag{4.1}$$

Greedy method is used to select k features which are the most relevant from input K features. The drocedure of the algorithm is that initialization of a feature sets V. And then detecting the maximizing relevant degree between feature f_j and the class $I(f_j, C)$ until empty feature set V.

However, the method does not account for the mutual information between the output target and the set of selected features. The overlap information between possibility feature and a set of selected features will be computed by the second term ($\beta \sum I(f_j, f_v)$) in 4.1.

The value of β plays a significant role in this method, if it is too large the redundancy will overestimate. However, the value of β is a problem which needs to be studied more by researchers.

As a result, the method only examines maximizing between the features and target class without considering the dependency. Thus the MIFS algorithm may not turn out an optimal set of features, as it is able to select the unrelated features.

MRMR

In this method, it replaces parameter β in the MIFS method by $\frac{=}{1}|V|$ and V is the number of elements in a set of selected features V. It represents a mean of the second term (redundancy term) in equation 4.1. The criterion can be rewritten as:

$$J = I(f_j, C) - \frac{1}{|V|} \sum_{f_v \in V} I(f_j, f_s) \tag{4.2}$$

As a result, the MRMR method avoids the struggling in selecting parameter β problem. Since inappropriate β will lead to unbalance result between first term (relevance term) and the redundancy term in definition 4.1.

It is clear to see that the amount of redundancy term will increase when the number of selected features become larger. If the relevance term is insignificant compared with redundancy term, feature selection method bases on minimizing redundancy term to pick up features. Thus, the results of feature selection may be the irrelevant features. The MRMR method will narrow this issue by replacing the parameter β by the average of redundancy term.

In this method, each feature is supposed independence from target class. In other words, the feature selection method is a first-order incremental algorithm. Thus MRMR method has the same restriction that MIFS method holds. The numerous irrelevant and redundant features will be selected in this case.

JMI

The method considers the conditional mutual information to proceed feature selection algorithm, it is defined as:

$$J = I(f_j, C) - \frac{1}{|V|} \sum_{f_v \in V} (I(f_j, f_v) - I(f_j, f_v|C)) \tag{4.3}$$

The features are selected if they fetch the informative contribution to a set of selected features. The JMI method is an effective algorithm to discard redundant features. It expresses the similarity criterion as MRMR in the dependent pairwise features which are selected while the set of selected features increasing.

4.2 Feature Selection Algorithm

This chapter is the main part of the thesis, the feature selection problem will be solved by a graph description of features, called dominant-set cluster. In general, the algorithm has three steps:

- Firstly, the $n \times n$ relevant matrix W will be calculated. The relevance value w_{ij} of matrix $W = w_{ij}$ is derived from computing mutual information.
- Secondly, in this step the variable vectors (features) will be grouped by graph-theoretic clustering method (dominant-set clustering).
- Finally, utilizing multidimension mutual information, or multidimension interaction information, criterion to select the features from each dominant set. Figure 4.1 shows the flowchart of this method.

4.2.1 Deriving the Relevance Matrix

There are various methods to measure the relevant degree of features. However, the different results can become out by different methods in term of clustering analysis. "Euclidean distance" is one of the most common method which is widely utilized to measure the distance between objects in clustering problem.

Nevertheless, it only considers on objects with a specific distribution. In term of functional similarity which consists positive and negative correlation and interdependency, the method is not responsible for. On the other hand, two methods, is proposed by Rao, which is able to take a measure of linear dependency between features. They are called "Pearson's correlation coefficient (ρ)" and "Least square regression error (e)". The first one (ρ) is given as:

$$\rho(f_1, f_2) = \frac{cov(f_1, f_2)}{(var(f_1)var(f_2))^{1/2}} \tag{4.4}$$

where f_1 and f_2 are two features, $var(\cdot)$ is the variance of a feature, and $cov(f_1, f_2)$ denotes the covariance between two features f_1 and f_2. It is clearly that Pearson's correlation coefficient measure the linear dependency between two features based on the above formula. If the value of $\rho(f_1, f_2)$ is small, this indicates that feature f_1 and f_2 are highly irrelevant, whereas the amount of $\rho(f_1, f_2)$ is large, two features are highly relevant. Two features are completely irrelevant if the measure of $\rho(f_1, f_2) = 0$.

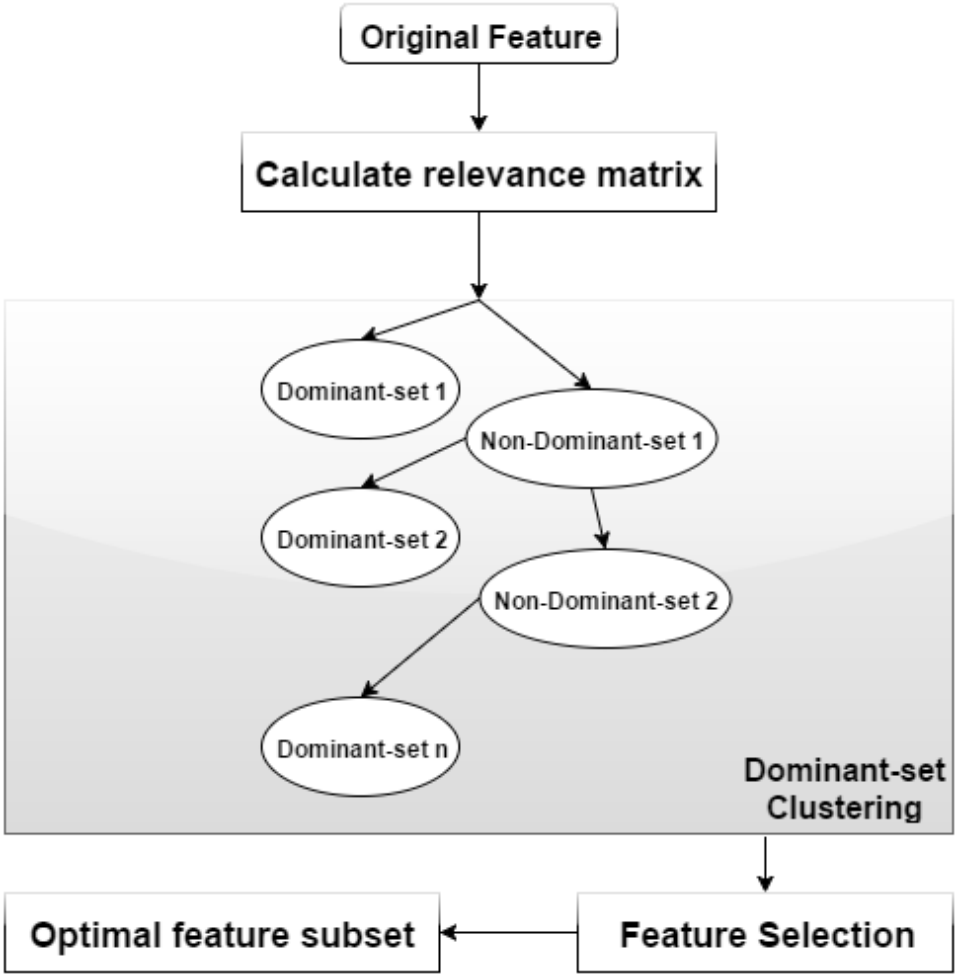


Figure 4.1: The flowchart of feature selection using dominant-set clustering [1]

The method is possible to locate the positive and negative value of correlation, Nonetheless, Pearson's correlation coefficient is not suitable to apply for the clustering method utilized in this thesis, dominant-set clustering. Since it may allocate a high relevant degree to a low relevant degree of pairwise features by reason that the method is not robust to outliers. In addition, the different variances of two pairwise features may come out the same measure of relevant degree since it is easily affected by rotation and stable on scaling.

In terms of least square regression error (e), modeling the dependency of two feature vectors f_1 and f_2 as a linear equation, $f_2 = af_1 + b$. Thus the predicting of error function, from the linear equation, is able to take a measure of the dependency degree between two feature vectors. Minimizing the mean square error is able to obtain the values of a and b of the linear

model. The mean square error is defined as:

$$e(f_1, f_2)^2 = \frac{1}{N}\sum_{j=1}^{N}((f_{2j}) - b - af_{1j})^2 \tag{4.5}$$

where $a = \frac{cov(f_1,f_2)}{var(f_1)}$, $b = \bar{(}f_2)$, and $e(f_1, f_2) = var(f_2)(1 - \rho(f_1, f_2)^2)$. It is clear from the above definition that, the method uses the linear model to measure the value of variance of f_2. As a result that it is easily affected by rotation and stable on scaling like ρ method.

As a result, utilization of mutual information to measure the relevant degree is more suitable in dominant-set clustering. As the definitions in section 2.1 about the entropy $H(Y)$ of feature vector Y and the mutual information $I(Y, X)$ of two feature vector Y and X.

The mutual information is represented in the relative information between two feature vectors. It indicates that the feature vector X and Y are highly relevant if the amount of relative information is large. Whereas the mutual information is small, it means that the feature vector X and feature vector Y are highly irrelevant. Especially, Two feature vectors are completely irrelevant when the amount of mutual information equal to zero.

The fact is the reason why the relevant degree of two features will be measured by mutual information in the feature selection context of this thesis. Let consider the input data set with N feature vectors and M observations (or training samples). The m^{th} observation of n^{th} feature is denoted by f_n^m. Following this way, it is easy to obtain the vector $F_n = \{f_n^1, \cdots, f_n^m\}$ of n^{th} feature for M observations.

According to [15], the measure of relevance between two variable vectors (feature vectors) F_{ni} and F_{nj} respectively, is defined as:

$$RD(F_{ni}, F_{nj}) = \frac{2I(F_{ni}, F_{nj})}{H(F_{ni}) + H(F_{nj})} \tag{4.6}$$

where $I(F_{ni}, F_{nj})$ is the mutual information between two variables (F_{ni}, F_{nj}), it can be computed by formula 2.6; $H(F_{ni})$ and $H(F_{nj})$ are entropy of two variables, it can be computed by formula 2.2; ni and nj belong to N; $RD(F_{ni}, F_{nj})$ is a element of relevance matrix $W(F_{ni}, F_{nj})$ mentioned above.

Note that the relevance matrix is symmetric, thus $RD(F_{ni}, F_{nj}) = RD(F_{nj}, F_{ni})$; since the graph G considered in this scheme is no self-loop, $RD(F_{ni}, F_{ni}) = 0$. More significantly, two features F_{ni}, F_{nj} are more relevant, if value of $W(F_{ni}, F_{nj})$ is highly. Whereas two features will be completely irrelevant when the $W(F_{ni}, F_{nj})$ is equal to zero.

Last but not least, to calculate the value of entropy and mutual information, the probability density function of these features has to be estimated first. In this thesis, the Parzen- Rosenblatt window method will be utilized since a measure of mutual information is more accurate, it leads to better performance.

The input data set, there is M observations of feature y, the estimate probability density function is given by the formula below:

$$\hat{p}(y) = \frac{1}{M}\sum_{i=1}^{M}\delta(y - y_i, h) \tag{4.7}$$

where y_i is the i_{th} observation; h is the window width; and $\delta(y - y_i, h)$ is the Parzen- Rosenblatt window function. According to Parzen, the estimate probability density $\hat{p}(y)$ will converge to probability density $p(y)$ if the window width h and window function are picked up appropriately. Commonly, the window function is used as Gaussian kernel. In this case, the window function will become:

$$\delta(y - y_i, h) = \frac{1}{2\pi^{d/2}h^d|\sum|^{1/2}}\exp\left(-\frac{(y - y_i)^T\sum^{-1}(y - y_i)}{2h^2}\right) \tag{4.8}$$

where d is the dimension of feature y; $\sum$ is the covariance of$(y - y_i)$. The marginal density $p(y)$ will be turned out if the value of dimension $d = 1$, while $d = 2$, approximate joint probability density, between two features, will be represented by $p(y, x)$ or $p(x, y)$.

4.2.2 Dominant Set Clustering

As discussion in chapter 3 and refer to the flowchart in figure 4.1, the input of Dominant-set clustering algorithm is the relevance matrix which is computed in section 4.2.1. According to relevance matrix, a dominant set is iteratively derived from the mutual information computation. The dominant sets will be gradually created the following way.

It is clear to see that they are hierarchically like figure 4.1 illustrates. The algorithm is terminated when there are no features to group, in the other words, all the features of the input data are clustered into dominant sets.

4.2.3 Deriving Key Features from each Dominant-set

Problem Statement:

The amount of information on output class, in the selected features, decides the accurateness of feature selection algorithms. According to Fano's inequality [11], the error estimate probability of class C with input feature vector F is defined as:

$$P_E \geq \frac{H(C|F) - 1}{\log |C|} = \frac{H(C) - I(F, C) - 1}{\log |C|} \tag{4.9}$$

where $|C|$ is the number of classes in the data; $H(C)$ and $I(F, C)$ are entropy and mutual information between classes and features. The aim is that minimizing the error estimate probability 4.9. The fact that entropy of class and class number is constant. If the mutual information between feature and class $I(F, C)$ is maximal, the measure of P_E will be minimized. Thus, to obtain the accurate feature selection, the mutual information between feature and class should be maximized.

However, minimizing mutual information $I(F, C)$ to detect a feature subset requiring a huge computation to locate on feature space. The more significant problem is that the size of training observations is numerous so that the higher order joint probability density can estimate with the high dimensional kernel. The best solution to deal with these problems is that the features are supposed to lower-order dependence.

For instance, each variable is supposed that independently affect the class feature, thus the n^{th} feature (f_n) is selected such that $P(f_n|f_1, \cdots, f_{n-1}, C) = P(f_n|C)$. In the hypothesis of second-order feature dependence [16], Indicate that replacing the information $I(F, C)$ with the approximate $\hat{I}(F, C)$, it is able to detect relative information features. The greedy algorithm will be used at the end to progressively pick up more relevant features. The approximate $\hat{I}(F, C)$ is defined as:

$$\hat{I}(F, C) \approx I(F, C) = \sum_n I(f_n, C) - \sum_n \sum_{m>n} I(f_n, f_m) + \sum_n \sum_{m>n} I(f_n, f_m|C)$$

A brief algorithm of this approach, Let consider an input data that has N features. The selected feature m $(n < N)$ is derived from two steps. Firstly, f'_{max} of maximizing mutual information $I(f', C)$ is selected. Finally, following this procedure to progressively select $(n - 1)$ features where the mutual information $I(F, C)$ is maximized, for instance, the maximizing $I(f'', C) - I(f'', f'_{max}) + I(f'', f'_{max}|C)$ to derive the feature f''_{max}.

In this thesis approach, the input features are already clustered into different subset by dominant-set algorithm. There is a small number of features in each set, therefore, the multidimensional mutual information (interaction information) $I(F, C)$ is directly utilized to selects features, the such approximation $\hat{I}(F, C)$ does not need to utilize in this case.

In practice, to perform the feature selection algorithm, firstly the estimate probability function should be considered. In this scheme, the Parzen-Rosenblatt window algorithm with a Gaussian kernel is considered.

Measure of Mutual Information by Parzen-Rosenblatt window:

In the feature selection context, the input data are normally continuous features, but the class is discrete one. The multidimensional mutual information of input features $F = f_1, \cdots, f_n$ and the class C can be given as:

$$I(F;C) = I(f_1, \cdots, f_n; C) = \sum_{f} \sum_{c \in C} P(f_1, \cdots, f_n; c) \log \frac{P(f_1, \cdots, f_n; c)}{P(f_1, \cdots, f_n) P(c)} \tag{4.10}$$

In other way based on Markov chain, the multidimensional interaction information is represented as follows:

$$I(F;C) = H(C) - H(C|F) \tag{4.11}$$

In this formula, the value of class entropy $H(C)$ can be calculated by 2.2 easily since the class C is discrete value. Whereas the condition entropy $H(C|F)$ is difficult to obtain because of estimating conditional probability $p(c|f)$. It is represented as:

$$H(C|F) = -\int_{F} p(f) \sum_{c=1}^{|C|} p(c|f) \log p(c|f) \mathrm{d}f \tag{4.12}$$

In this thesis, the Parzen window method will be used to estimate the probability density function. According to Bayesian rule, it is able to rewrite $p(c|f)$ as:

$$p(c|f) = \frac{p(f|c)p(c)}{p(f)} \tag{4.13}$$

The conditional probability density function can be estimated by applying the Parzen-Rosenblatt window with the possible value of class

number is $\{1, 2, \cdots, |C|\}$ and N_c is the number of training samples in class c, it is given as:

$$\hat{p}(f|c) = \frac{1}{N_c} \sum_{i \in I_c} \delta(f - f_i, h) \tag{4.14}$$

where I_c is the indices set of observations in class c. In fact, the sum of conditional probability is equal to one, it is given that:

$$\sum_{n=1}^{|C|} p(n|f) = 1, \tag{4.15}$$

From the such formula and 4.13 equation, the equation of conditional probability 4.14 is rewritten as:

$$p(c|f) = \frac{p(c|f)}{\sum_{n=1}^{|C|} p(n|f)} = \frac{p(c)p(f|c)}{\sum_{n=1}^{|C|} p(n)p(f|n)} \tag{4.16}$$

Combining the such equation with the equation in 4.14, the conditional probability estimation is:

$$\hat{p}(c|f) = \frac{\sum_{i \in I_c} \delta(f - f_i, h_c)}{\sum_{n=1}^{|C|} \sum_{i \in I_n} \delta(f - f_i, h_n)} \tag{4.17}$$

where $p(n) \approx N_n/N_c$. When the equation 4.17 is utilized with the Gaussian kernel, the conditional probability estimation $\hat{p}(c|f)$ becomes:

$$\hat{p}(c|f) = \frac{\sum_{i \in I_c} \exp\left(-\frac{(f-f_i)^T \Sigma^{-1} (f-f_i)}{2h^2}\right)}{\sum_{n=1}^{|C|} \sum_{i \in I_n} \exp\left(-\frac{(f-f_i)^T \Sigma^{-1} (f-f_i)}{2h^2}\right)} \tag{4.18}$$

A measure of conditional entropy in equation 4.12 is rewritten as:

$$\hat{H}(C|F) = -\sum_{n=1}^{N} \frac{1}{N} \sum_{c=1}^{|C|} \hat{p}(c|f_n) \log p(c|f_n) \tag{4.19}$$

In this case, with N training observations, instead of using iteration, a summation of training observations and assume that the probability is the same at each observation.

Greedy algorithm

The final step of feature selection criterion is that using the greedy algorithm to select the key feature from each dominant set. There is the input features of each dominant set F_d with and class C. Locate the subset $S_d \subset F_d$ in each dominant set to minimize the multidimensional interaction information. The algorithm is briefly shown as below:

- Start from the first dominant set F_d of feature vectors.
- Initialize the empty set S_d.
- For all the features (f_m) belong to dominant set F_d, calculate the mutual information $I(f_m, C)$ between feature and output class.
- Assign the feature (f_m) that maximizes mutual information $I(f_m, C)$ to subset S, and remove this feature out of a dominant set F_d.
- Using greedy strategy to select features until reaching desired feature number.
  - For all features f_m in the current dominant set F_d, compute the mutual information between features and output class $I(f_m, S_d; C)$.
  - Assign the feature f_m belongs to F_d that maximize multidimensional mutual information $I(f_m, S_d; C)$, and then remove the feature out of the dominant set.
- The subset S_d is turned out with the selected features.
- Repeat the algorithm until there is no dominant set.

Chapter 5

Experiments and Experimental Results

In this chapter, there are six datasets are used for testing the algorithm, including: five datasets from UCI Machine Learning Repository and one from NIPS 2003. Table 5.1 will show a brief information about these dataset.

Data sets	Data sources	Instances	Features	Number of classes
Iris	UCI	150	4	3
Pima	UCI	768	8	2
Australian	UCI	690	14	2
Satimage	UCI	4435	36	6
Breast cancer	UCI	699	10	2
Madelon	NIPS 2013	2000	500	2

Table 5.1: Data sets used for testing in the algorithm

In the context of this thesis, a feature selection method (mentioned as *Dset*MII). The method highly groups relevant features into a different dominant set (cluster) based on dominant-set clustering algorithm. And then applying the criterion of multidimensional interaction information (MII) to select useful features in each dominant set.

Table 5.2 illustrates the results of dominant-set clusters. There are several methods to compute the relevance matrix, including: "Euclidean distance", "Pearson's correlation coefficient (ρ)", and "Least square regression error (e)". In this thesis, the algorithm applied the mutual information to calculate the similarity between features.

Dominant sets	Iris	Pima	Australian	Breast cancer
Dominant set 1	$\{f_1, f_2, f_3, f_4\}$	$\{f_1, f_2, f_6, f_8\}$	$\{f_8, f_9\}$	$\{f_3, f_4, f_6, f_7, f_8, f_9\}$
Dominant set 2		$\{f_3, f_4, f_7\}$	$\{f_5, f_7, f_{10}, f_{14}\}$	$\{f_1, f_2, f_5, f_{10}\}$
Dominant set 3		$\{f_5\}$	$\{f_2, f_3, f_4, f_6, f_{11}, f_{12}, f_{13}\}$	
Dominant set 4			$\{f_1\}$	

Table 5.2: Dominant-set clustering results

When the clustering sets come out from such algorithm. It is easy to apply the MII criterion to select features from each dominant set since the number of features is significant smaller than original features set. Table 5.3 presents the results after applying feature selection criterion. It ranks from the highest to lowest relevant degree.

Dominant sets	Iris	Pima	Australian	Breast cancer
Dominant set 1	$\{f_3, f_4, f_1, f_2\}$	$\{f_2, f_8, f_6, f_1\}$	$\{f_8, f_9\}$	$\{f_3, f_7, f_6, f_8, f_4, f_9\}$
Dominant set 2		$\{f_4, f_3, f_7\}$	$\{f_{10}, f_7, f_5, f_{14}\}$	$\{f_5, f_2, f_{10}, f_1\}$
Dominant set 3		$\{f_5\}$	$\{f_3, f_6, f_{13}, f_2, f_4, f_{12}, f_{11}\}$	
Dominant set 4			$\{f_1\}$	

Table 5.3: Selected features in each dominant set

In practice, The results of various methods demonstrated that the method, is given in this thesis (*Dset*MII), has more advantages than other methods (including: MIFS, MRMR, and JMI are mentioned in the introduction) to select features in a higher dimension. Table 5.4 shows more clearly these benefits. The "Pima" data set has 8 features, the methods show the same results with selected features whereas the "Breast cancer" data set has 10 features, there exist the different selections form third selected feature.

In addition, when the higher dimension is taken into account that the results are considerable differences among the methods. Table 5.5 illustrates this point on 500 features data set.

The reason why *Dset*MII has more advantage in high dimension is that

Method	Breast cancer	Pima
*Dset*MII	$\{f_3, f_7, f_8, f_6, f_4\}$	$\{f_2, f_8, f_6, f_1, f_4\}$
MIFS	$\{f_3, f_7, f_1, f_{10}, f_2\}$	$\{f_2, f_8, f_6, f_7, f_5\}$
MRMR	$\{f_3, f_7, f_8, f_6, f_2\}$	$\{f_2, f_8, f_6, f_7, f_5\}$
JMI	$\{f_3, f_7, f_8, f_4, f_1\}$	$\{f_2, f_8, f_6, f_1, f_4\}$

Table 5.4: Selected features on different methods

Method	Madelon dataset
*Dset*MII	$\{f_{476}, f_{242}, f_{393}, f_{65}, f_{337}\}$
MIFS	$\{f_{476}, f_{49}, f_{178}, f_{131}, f_{491}\}$
MRMR	$\{f_{476}, f_{49}, f_{178}, f_{299}, f_{491}\}$
JMI	$\{f_{476}, f_{339}, f_{242}, f_{154}, f_{282}\}$

Table 5.5: Selected features from higher dimension

dominant-set clustering using the mutual information to compute the relative information between features, so the most relevance features will be will be selected. Secondly, after clustering, each dominant set utilizes multidimension interaction information to select the feature, therefore, the class and features are examined by higher order. This is considered as an optimized feature subset. And the final one is that the pMII method tends to consider a pairwise of class and features, therefore, all the features relationship could not be checked. Throwback table 5.4 and 5.5, it is clearer to see that other methods, including: MIFS, MRMR, and JMI criteria tend to consider pairwise features and class, therefore, the results are significantly different with the higher order method.

And then Scatter Separability Criterion is applied to measure the selected featuers quality. On this criterion, S_w is the within class scatter matrix which measures the scatter of observation from the mean of cluster and S_b is the between class sacatter matrix which measures scatter of cluster means are from the total mean. They are defined as,

$$S_w = \sum_{i}^{cL} \pi_i E\{(X - \mu_i)(X - \mu_i)^T | \omega_i\} = \sum_{i}^{cL} \pi_i \Sigma_i$$

$$S_b = \sum_{i}^{cL} \pi_i (\mu_i - E\{X\})(\mu_i - E\{X\})^T$$

Where cL is the cluster number, X is feature vectors, π_i is the probability of a sample in cluster ω_i, μ_i is the sample mean vector of cluster i^{th}, Σ_i is a sample covariance matrix of cluster i^{th}, and $E\{Z\}$ is expected value. And the

J value measures the selected feature discrimination is proposed by Devijiver and Kittler, defined as:

$$J(X) = \frac{|S_w + S_b|}{S_w} = \prod_{i=1}^{cL}(1 + \lambda_i)$$

where $\lambda_i, i = 1, \cdots, cL$ are the eigenvalues of matrix $S_w^{(-1)}S_b$. where X is the feature set and $|S|$ is the diagonal sum of elements of S. Table 5.6 shows the results of the comparison between the such methods.

Method	Breast cancer	Pima	Madelon
*Dset*MII	2.718	1.190	1.0353
MIFS	1.0065	1.0329	1.0316
MRMR	2.5059	1.0329	1.0312
JMI	1.0065	1.190	1.0189

Table 5.6: Comparison of J values of different methods

Last but not least, after capturing the key features. The classification algorithm is applied to evaluate these features accuracy. The algorithm is utilized in this experiment is a perceptron with 10-fold cross-validation. As shown in table 5.7, 5.8, and 5.9, the accuracy of classification is compared by various methods with the number of selected features are 5, 4, 3 respectively.

As the results shown, the MII feature selection using dominant-set clustering method performs better than other methods, including: MIFS, MRMR, and JMI. It is significant the performance obtains around 85% when 5 selected features are used with *Dset*MII method and keep going the identical accuracy with 3 and 4 selected features (Australian data set). However, the results, are obtained with the Pima data set, tend to have the same accuracy because of less dimension. It indicates that there are several informative features used to deal with classification problem.

Method	Australian	Pima
*Dset*MII	85.1%	62.76%
MIFS	67.25%	61.20%
MRMR	57.25%	61.20%
JMI	57.24%	61.72%

Table 5.7: The classification accuracy comparison with 5 features selected by 4 methods

Method	Australian	Pima
*Dset*MII	84.20%	64.84%
MIFS	67.54%	63.67%
MRMR	43.33%	63.67%
JMI	84.20%	64.84%

Table 5.8: The classification accuracy comparison with 4 features selected by 4 methods

Method	Australian	Pima
*Dset*MII	84.35%	63.93%
MIFS	55.22%	63.93%
MRMR	49.28%	63.93%
JMI	84.35%	63.93%

Table 5.9: The classification accuracy comparison with 3 features selected by 4 methods

The reasons why the *Dset*MII is effective in term of feature selection comparison with others is that the method measures the relevance between features based on mutual information, hence, the most relative information between features and class are selected. More significant, the selected feature subset examines informative present for each feature together with the correlation between features, therefore, it is effectively able to locate the potential information about input features.

Chapter 6

Conclusion

In this thesis, we presented the multidimension interaction information criterion using dominant-set clustering for feature selection based on greedy strategy which is a "pick-one-feature-at-a-time" method. In this method, dominant-set clustering uses mutual information to calculate the relative information of features, in other words, relevance matrix is computed, for that reason, the most informative information of features will be clustered in the order dominant sets. On the other hand, the method considers high order dependencies of class and features, hence, it is able to deal with "the problem of overestimated redundancy". The result of this method shows that the higher measure of mutual information between features, higher ranking in the selected features.

Bibliography

[1] Zhihong Zhang, Edwin R.Hancock: A Graph-based Approach to Feature Selection. In: Chapter Graph -Based Representations in Pattern Recognition Volume 6658 of the series Lecture Notes in Computer Science pp 205-214

[2] Massimiliano Pavan, Marcello Pelillo: A New Graph-Theoretic Approach to Clustering and Segmentation . In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1. IEEE (2003)

[3] Roberto Battiti: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Transactions on Neural Networks 5(4), 537-550 (2002)

[4] Hanchuan Peng, Fuhui Long, and Chris Ding: 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)

[5] Howard Hua Yang, John Moody: Feature Selection Based on Joint Mutual Information. In: Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis. pp. 22-25 (1999)

[6] Nojun Kwak and Chong-Ho Choi: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE TPAMI 24(12), 1667-1671 (2002)

[7] Francÿois Fleuret: Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research 5, 1531-1555 (2004)

[8] Dahua Lin, Xiaoou Tang: Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion. In European Conference on Computer Vision.

[9] Michel Vidal-Naquet, Shimon Ullman: Object recognition with informative features and linear classification. IEEE Conf. on Computer Vision and Pattern Recognition.

[10] C. E. Shannon: A Mathematical Theory of Communication. Reprinted with corrections from The Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656, July, October, 1948.

[11] Thomas M. Cover, Joy A. Thomas: Elements of Information Theory, Second edition, A John Wiley & Sons, INC., Publication.

[12] Edward A. Bender, S. Gill Williamson: Unit GT: Basic Concepts in Graph Theory 2010.

[13] Jianbo Shi and Jitendra Malik: Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 22, NO. 8, August 2000.

[14] Isabelle Guyon, Andre Elisseeff: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 (2003).

[15] Feng Zhang, Ya-Jun Zhao, Jun-fen Chen: Unsupervised Feature Selection Based On Feature Relevance. Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009.

[16] Baofeng Guo and Mark S. Nixon: Gait Feature Subset Selection by Mutual Information. IEEE Transactions on Systems, MAN, and Cybernetics-part a: Systems and Humans, Vol. 39, No. 1, January 2009.