Communications of the Korean Statistical Society
2009, Vol. 16, No. 1, 143–155

Variable Selection Based on Mutual Information

Moon Y. Huh[1,a], Byong Su Choi[b]

[a]Dept. of Statistics, Sungkyunkwan Univ., [b]Dept. of Mutimedia Engineering, Hansung Univ.

Abstract

Best subset selection procedure based on mutual information (MI) between a set of explanatory variables and a dependent class variable is suggested. Derivation of multivariate MI is based on normal mixtures. Several types of normal mixtures are proposed. Also a best subset selection algorithm is proposed. Four real data sets are employed to demonstrate the efficiency of the proposals.

Keywords: Best subset selection, feature selection, mutual information, normal mixture.

1. Introduction

Feature selection or variable selection has been one of the most important topics in data analysis. For instance, the stepwise variable selection method is widely applied in developing a best model in various regression problems. However, such a variable selection method is available rather limited applications. The screening variables before analysis may be more conventional method and has been used more often than the model based variable selection procedure. Note, however, that the conventional method may also fail when the explanatory variables are of complex types. Liu and Motoda (1998) gives a good survey on this topic.

Recently, various methods for screening or ranking of complex types of variables have been investigated specially in machine learning area. Examples of the study for variable ranking for the complex data types are ReliefF by Kononenko *et al.* (1997), and MDI (measure of departure from independence) by Lee and Huh (2003). Another approach for variable selection is using mutual information (MI) based on Shannon's entropy theory (Shannon, 1948) to measure the association between explanatory variables and a class variable. In fact, many researcher have payed attention to the mutual information as an ideal measure of association. See, for example, Cover and Thomas (1991), Darbellay (1999), and Joe (1989). The most desirable property of MI is that it can measure all kinds of dependency between variables and between groups of variables unlike the correlation coefficient or the rank correlation coefficient. It is well-known that they can only account for linear relationships or monotone dependencies between two variables. Beside the excellency, it seems that further studies should be done to make MI as a practical measure of association, because the efficient estimation of MI is an unsolved problem yet.

Some works have been done to investigate the distributional properties of the estimator of MI (Brillinger, 2004; Christensen, 1997 and Hutter, 2002). Tourassi *et al.* (2001) investigated the property of application of MI for complex type variable selection. These works, however, are based on measuring the association between discrete variables because they first discretize continuous variables

This Research was financially supported by Hansung University in the year of 2007

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan Univ., Myungrun-Dong, Chongro-Ku, Seoul 110-745, Korea. E-mail: myhuh@skku.edu

for MI. Obtaining MI through discretizing continuous variables raises several problems. Firstly, discretization itself is an NP-hard problem (Nguyen and Skowron, 1995), and the MI depends upon the result of discretization. Secondly, for multivariate case, discretization will produce many empty cells even with large sample sizes, and computing MI will be unstable. Consider, for example, the case of 3 variables with 10 categories for each variable. This will introduce $10^3 = 1,000$ cells, and most of the cells will be empty even with large number of observations. This is especially a severe problem when there are huge number of variables with not so large sample sizes which is usually the case with gene selection problems.

We do not have the above problems of discretization if we compute MI between continuous explanatory variables and a discrete class variable directly from the data. This requires density estimation, and Parzen filter approach has been used (Torkkola and Campbell, 2000; Kojadinovic, 2005). Density estimation usually requires lots of computing efforts, and it is well known that this nonparametric approach does not provide efficient result.

In next section, we proposes an easy–to–compute models for MI between continuous explanatory variables and a discrete class variable using normal mixtures. The moment estimation approach is employed to obtain entropy. We consider simple 1-component normal and 2-components normal mixture to estimate the unknown multivariate density of the continuous explanatory variables when the category of a class variable is given. In Section 3, we suggest an algorithm for best subset selection. The algorithm is not a perfect enumeration, but tries to find a sub-optimal subset based on the information between explanatory and class variables. In Section 4, four real data sets from UCI database (Merz and Murphy, 1996) are employed to demonstrate the efficiency of the proposed models. The evaluations of the models are performed using 10–fold cross validation with logistic regression and J48 decision tree implemented in Weka (Witten and Frank, 1999). We also provide the process of data visualization using DAVIS (Huh and Song, 2002) to visually confirm the efficiency of the results of subset variable selection, whenever possible.

2. Estimation of Mutual Information

Mutual information (MI) between two random variables X and Y is defined as:

$$I(X;Y) = \sum_{x\epsilon\mathcal{X}} \sum_{y\epsilon\mathcal{Y}} \Pr(x,y) \log \frac{\Pr(x,y)}{\Pr(x)\Pr(y)} \tag{2.1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the finite sets of values of all possible values for X and Y, respectively. It is straightforward to see that MI can be rewritten as follows:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{2.2}$$

where H denotes the entropy of a random variable, and is defined to be

$$H(X) = -\sum_{x\epsilon\mathcal{X}} \Pr(x) \log \Pr(x) \tag{2.3}$$

$$H(X|Y) = -\sum_{y\epsilon\mathcal{Y}} \Pr(y) \sum_{x\epsilon\mathcal{X}} \Pr(x|y) \log \Pr(x|y) \tag{2.4}$$

A natural extension of MI to the case when X is continuous and Y is discrete is using the integral instead of summation. Doing this, $H(X)$ and $H(X|Y)$ are defined as follows:

$$H(X) = -\int f(x) \log f_X(x) dx \tag{2.5}$$

$$H(X|Y) = -\sum_{k}^{K} p_k \int f_{X|Y}(x|k) \log f_{X|Y}(x|k) dx \tag{2.6}$$

where $f_X(x)$ is the marginal density of X, $f_{X|Y}(x|k)$ is the conditional density of X given $Y = k$ with $p_k = f_Y(k), k = 1, \ldots, K$, where K being the number of categories of Y.

Now we assume $f_{X|Y}(x|k)$ to be a normal density with mean μ_k and variance σ_k^2. Then it is easy to check that

$$H(X|Y) = \frac{1}{2}(1 + \log 2\pi) + \sum_{k=1}^{K} p_k \log\left(\sigma_k^2\right).$$

Substituting σ_k^2 and p_k by their estimates, we can obtain the estimate of $H(X|Y)$. Unlike the simplicity in obtaining the estimate of $H(X|Y)$, the estimation procedure for $H(X)$ is a little bit complicate. The marginal density of X is given by

$$f_X(x) = \sum_{k=1}^{K} f_{X,Y}(x,k) = \sum_{k=1}^{K} p_k f_{X|Y}(x|k) = \sum_{k=1}^{K} p_k \phi(x|\mu_k, \sigma_k)$$

where $\phi(x|\mu, \sigma)$ denotes the density of a normal random variable with mean μ and standard deviation σ. Since $H(X)$ is the expected value of $-\log f_X(x)$, a natural estimate can be obtained by the method of moments which would give the sample mean of the realization of $-\log f_X(x)$. Substituting μ_k, σ_k and p_k by their estimates, we have the following result.

$$\widehat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n} \log \widehat{f_X}(x_i) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \widehat{p}_k \phi(x_i|\widehat{\mu}_k, \widehat{\sigma}_k)\right)$$

where n is the sample size and x_i's are the observed values of X. This approach can be easily extended to the case of p-dimensional X as follows.

1. Estimation of MI: 1-component normal model.

MI for $p-$dimensional continuous variable X and discrete class variable Y can be estimated as follows:

$$\widehat{I}(X|Y) = \widehat{H}(X) - \widehat{H}(X|Y) \tag{2.7}$$

where

$$\widehat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n} \log \sum_{k=1}^{K} \hat{p}_k \phi(x_i; \hat{\mu}_k, \hat{\Sigma}_k) \tag{2.8}$$

and

$$\widehat{H}(X|Y) = \frac{p}{2}(1 + \log 2\pi) + \sum_{k=1}^{K} \hat{p}_k \log|\hat{\Sigma}_k| \tag{2.9}$$

$X|Y = k$ is assumed to be normally distributed. However, the normality assumption might not be appropriate in some cases. Previous studies have suggested to approximate $f_{X|Y}(x|k)$ with normal mixtures. For example, Wang (2001) showed that the mixture of normal distributions provides a useful extension of the normal distribution for modeling data with fatter-than-normal tails or with skewness. Thus we now assume that $f_{X|Y}(x|k)$ is the mixture of normal densities. However, a question might

be arisen. How many normal distributions might be appropriate? MCLUST (Fraley and Raftery, 2002) could give a partial answer to the question. It selects the number of distributions by the Bayesian Information Criterion and uses the EM algorithm to estimate parameters. However, the procedure requires lots of computing time to determine the number of components for normal mixtures when the dimension of X is large. This makes it difficult to employ the algorithm of MCLUST directly to compute the MI when the number of variables is large.

We have run several experiments and have observed that the 2-components mixture works quite well for our purpose. Hence, we propose to assume that the density $f_{X|Y}(x|k)$ is a 2-components normal mixture as given in the following:

$$f_{X|Y}(x|k) = (1-\pi_k)\phi(x|\mu_{k1},\sigma_{k1}) + \pi_k\phi(x|\mu_{k2},\sigma_{k2}) \tag{2.10}$$

where $0 \le \pi_k \le 1$. Then the marginal density of X can be written as

$$f_X(x) = \sum_{k=1}^{K} p_k((1-\pi_k)\phi(x|\mu_{k1},\sigma_{k1}) + \pi_k\phi(x|\mu_{k2},\sigma_{k2})). \tag{2.11}$$

One way to estimate the 5 parameters $(\pi_k, \mu_{k1}, \mu_{k2}, \sigma_{k1}, \sigma_{k2})$ for each category $k = 1,\ldots,K$ is to employ the EM algorithm. Plugging in these estimates to (2.11), we can estimate the marginal density of X. Then, as before, $H(X)$ is estimated by the sample mean of the realization of $-\log f(x)$,

$$\widehat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n}\log\widehat{f_X}(x_i) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\hat{p}_k\widehat{f_{X|Y}}(x|k)\right). \tag{2.12}$$

Similarly, $H(X|Y)$ can be estimated as follows:

$$\widehat{H}(X|Y) = -\sum_{k=1}^{K}\hat{p}_k\frac{1}{n_k}\sum_{i=1}^{n_k}\log\widehat{f_{X|Y}}(x_{ki}|k) = -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\log\widehat{f_{X|Y}}(x_{ki}|k) \tag{2.13}$$

where $\hat{p}_k = n_k/n$, with n_k being the number of observations belonging to the k-th category of the target variable Y.

2. Estimation of MI: 2-component normal mixture approximation

MI for the continuous variable X and discrete class variable Y can be estimated as follows:

$$\widehat{I}(X|Y) = \widehat{H}(X) - \widehat{H}(X|Y) \tag{2.14}$$

where

$$\widehat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n}\log\sum_{k=1}^{K}\hat{p}_k((1-\hat{\pi}_k)\phi(x_i|\hat{\mu}_{k1},\hat{\sigma}_{k1}) + \hat{\pi}_k\phi(x_i|\hat{\mu}_{k2},\hat{\sigma}_{k2})) \tag{2.15}$$

and

$$\widehat{H}(X|Y) = -\frac{1}{n}\sum_{k=1}^{K}\sum_{i}^{n_k}\log((1-\hat{\pi}_k)\phi(x_{ki}|\hat{\mu}_{k1},\hat{\sigma}_{k1}) + \hat{\pi}_k\phi(x_{ki}|\hat{\mu}_{k2},\hat{\sigma}_{k2})). \tag{2.16}$$

The estimates of $\pi_k, \mu_{k1}, \mu_{k2}, \sigma_{k1}$ and σ_{k2} are obtained by using the EM algorithm for the 2-components normal mixture.

This estimation process applies equally to the case of multidimensional X. In this case, the parameters $(\pi_k, \mu_{k1}, \mu_{k2}, \sigma_{k1}, \sigma_{k2})$ in the above equations will be replaced by $(\pi_k, \mu_{k1}, \mu_{k2}, \Sigma_{k1}, \Sigma_{k2})$, where μ_{k1}, μ_{k2} are $p-$dimensional mean-vector, and $(\Sigma_{k1}, \Sigma_{k2})$ are $p \times p$ dimensional variance-covariance matrix. We can also apply the EM algorithm to estimate theses parameters.

We now consider a further simplified approximation formula to estimate MI assuming all the $p-$variables are independent of each other.

3. Estimation of MI: X is multidimensional and independent.

Assuming the independence of the continuous p-dimensional variable X, MI between X and discrete class variable Y can be estimated as follows:

$$\widehat{I}(X|Y) = \widehat{H}(X) - \widehat{H}(X|Y) \tag{2.17}$$

where

$$\widehat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n} \log \sum_{k=1}^{K} \hat{p}_k \phi(x_i; \hat{\mu}_k, \hat{\mathrm{D}}_k) \tag{2.18}$$

and

$$\widehat{H}(X|Y) = \frac{p}{2}(1 + \log 2\pi) + \sum_{k=1}^{K} \hat{p}_k \log \prod_{i=1}^{p} \hat{\sigma}_{ki}^2 \tag{2.19}$$

with $\hat{\mathrm{D}}_k$ being the diagonal matrix of $\hat{\sigma}_{ki}^2$'s.

When X is discrete, density $f_X(x)$ is always less than 1, and the entropy is non-negative. However, $f_X(x)$ can be greater than 1 with continuous X, and this may cause the negative estimate of $H(X)$. We can safely set the negative entropy to 0 since this happens when the corresponding entropy is very small.

3. Best Subset Selection Strategy

Selecting optimal set of variables with p competing variables requires computing 2^p combinations of multivariate MI's. When there are 10 variables, this requires investigating more than 1,000 different sets of variables. A few heuristic approaches for subset variable selection strategies with discretized variables have been suggested (see, for example, Battiti, 1994). These approaches are basically based on the successive evaluation of the univariate MI's. However, the subset selection procedures based on univariate MI may yield catastrophic result when there is a high interrelationships among the variables. This will be demonstrated in the numerical examples of the next section.

Classical variable selection procedures used in linear regression models work well when the input variables are nearly independent (Miller, 1990). Other selection methods were suggested. See for example the work of Collett (2003). The methods considered only a portion of the complete enumerations of the subsets of the variables. We suggest another heuristic variable selection algorithm that practically considers most of the enumerations of the subset combinations of the variables. The algorithm consists of 4 steps.

Let $\mathcal{I}$ be the set of the indices corresponding to the initial set of variables from which the best subset is to be selected, and let $\mathcal{K}$ be the set of indices corresponding to the subset of variables selected as the best subset at current stage. The procedure is to update $\mathcal{K}$ either by adding a new variable index or by removing a subset of redundant variables. The adding process find the index i^* satisfying

$$i^* = \underset{i \in \mathcal{I}-\mathcal{K}}{\operatorname{argmax}}\ \mathrm{MI}(X_{\mathcal{K}+i}; Y).$$

Here $\mathcal{K} + i$ denotes $\mathcal{K} \cup \{i\}$, and $\mathcal{I} - \mathcal{K}$ is $\mathcal{I}$ excluding $\mathcal{K}$. After finding the index i^*, the removing process checks if there are any redundant variables in $X_{\mathcal{K}}$ which do not contribute to $\mathrm{MI}(X_{\mathcal{K}+i^*}; Y)$. This process is finding the set of indices $\mathcal{J}$ satisfying

$$\mathcal{J} = \{j | \mathrm{MI}(X_{\mathcal{K}+i^*-j}; Y) \geq \mathrm{MI}(X_{\mathcal{K}+i^*}; Y), j \in \mathcal{K}\}.$$

If this new subset of variables $X_{\mathcal{K}+i^*-\mathcal{J}}$ gives lower MI, we conclude $X_{\mathcal{K}}$ is the best subset. If not, we replace $\mathcal{K}$ with $\mathcal{K} + i^* - \mathcal{J}$, and continue the above process until $\mathcal{I} - \mathcal{K}$ becomes null.

When employing the algorithm in practice, we may have to consider the numerical instability. In other words, if we are going to continue selection procedure based on the computed value of MI, this may cause to continue selecting the next variables with a very small amount of increase in the values. This is demonstrated in the next section. However, the computation of MI is based on several steps of estimation such as the estimation of the mixture parameters by EM algorithm. Also there may be numerical errors from the iterative computation of matrices. Hence, we may need to set some allowance ϵ when comparing the magnitude of the two computed MI's for the selection procedure so that the MI increase should exceed this value if the next variable is to be selected. It can also work to control the number of variables to be selected. The amount of this value is purely dependent on domain experts. For our experiments, we apply the concept of relative increase of 1% in the values of two adjacent MI's. The algorithm is as follows:

1. (Initialization).

$\mathcal{I} = \{i \mid 1, 2, \ldots, p\}$ # *indices of all variables.*
$\mathcal{K} \leftarrow 0$ # *indices of currently selected variables.*
$\mathcal{J} \leftarrow 0$ # *indices of redundant variables.*
$\epsilon \leftarrow$ some small value

2. Apply the following stopping rule.

Let $i^* = \mathrm{argmax}_{i \in \mathcal{I}} \mathrm{MI}(X_{\mathcal{K}+i}; Y)$.
if $(\mathrm{MI}(X_{\mathcal{K}+i^*}; Y) - \mathrm{MI}(X_{\mathcal{K}}; Y))/\mathrm{MI}(X_{\mathcal{K}} + i^*; Y) \leq +\epsilon$ {
  return with $\mathcal{K}$ # *final selection: new selection does not increase MI.*
{
else {
  $\mathcal{K} \leftarrow \mathcal{K} + i^*$ # *add the new selection into* $\mathcal{K}$.
  $\mathcal{I} \leftarrow \mathcal{I} - i^*$ # *remove the selection from* $\mathcal{I}$.
}

3. Check if any variable from the redundant variable subset can increase MI when this variable is entered.

Let $j^* = \mathrm{argmax}_{j \in \mathcal{J}} \mathrm{MI}(X_{\mathcal{K}+j}; Y)$.
if $\mathrm{MI}(X_{\mathcal{K}+j^*}; Y) > \mathrm{MI}(X_{\mathcal{K}}; Y)$ {
  $\mathcal{J} \leftarrow \mathcal{J} - j^*$ # *remove the variable from the redundant subset.*
  $\mathcal{K} \leftarrow \mathcal{K} + j^*$ # *add the variable into the current subset of selection.*
}

4. Check if there is any redundant variables in the current subset selection.

Let $k^* = \text{argmax}_{k \in \mathcal{K}} \text{MI}(X_{\mathcal{K}-k}; Y)$
if $\text{MI}(X_{\mathcal{K}-k^*}; Y) > \text{MI}(X_{\mathcal{K}}; Y)$ {
  $\mathcal{K} \leftarrow \mathcal{K} - k^*$ *# remove the variable from the current subset selection.*
  $\mathcal{J} \leftarrow \mathcal{J} + k^*$ *# add the variable into the redundant subset.*
}

The algorithm considers all possible combinations starting with the largest univariate MI. There is a stopping criterion in the algorithm, and this criterion will be met in the early stage of subset selection process, unless the problem is catastrophic.

Remark 1. (Standardization of MI)

When the variable X is discrete type, MI tends to have larger values when the number of levels of X gets larger. Also as the number of variables gets larger, MI tends to get larger. These make difficult to use MI for the variable selection purpose. Press *et al.* (1992) suggests the following quantity as a standardized MI (SMI).

$$\text{SMI}(X; Y) = 2\frac{I(X; Y)}{H(X) + H(Y)}$$

Extensive experiments have shown that SMI imposes too much penalty on MI with large number of variables. As a result, too few variables (2 or 3) were selected in the most of our experiments. When we use normal mixture to estimate MI for continuous X, increasing the number of variables in X does not necessarily increase the value of MI. For example, adding a variable which is highly correlated with the variables already in the model will decrease the value of MI. Hence, we will use MI rather than SMI as a measure of information for subset selection criterion.

4. Experiments with Real Data Sets

The end purpose of the variable selection strategy suggested in the paper is to find a subset of variables that have high classification ability for the class variable. The subset selection is considered to be worthwhile if the classification accuracy of selected variables is better than that of the whole variables. In our experiments, the accuracy is measured by 10-fold cross validation with logistic regression and J48 decision tree which are implemented in Weka.

The chosen 4 data sets are: Fisher's Iris data (Iris), Wisconsin Prognostic Breast Cancer data (Breast cancer), Image data (Image) and Wine data (Wine). Iris data set has 150 observations with 4 continuous variables (sepal length, sepal width, petal length, petal width) and a class variable (species) with 3 categories (Iris-setosa, Iris-virginica, Iris-versicolor). Each category has 50 observations. Breast cancer data set has 31 variables and 569 observations. First 30 variables are continuous and the last one is diagnosis information with 2 categories (malignant and benign with 212 and 357 observations each). The 3rd data set is Image data set, and has 210 observations with 18 continuous and a class variable. There are 7 categories (7 different colors) for the class variable. Among the 18 continuous variables, the 3rd, the 4th, the 6th, and the 8th variables have only 2, 3, 3 and 5 different values, respectively, and some values have only 1 or 2 observations. Hence, these 4 variables are eliminated from the analysis in this work. Wine data set has 178 observations on 13 continuous and a class variable. The class variable has 3 categories and has 59, 71, and 48 observations in each category.

Table 1 and Figure 1 give the results of the univariate MI computed using the 1-component and 2-component normal mixtures for the 4 data sets. Figure 1 shows that 1-component and 2-components models for computing univariate MI agree well for most of the situations, and especially they do for

Table 1: Variable lists in the order of the magnitude of univariate MI's. Data sets are, I: Iris, B: Breast Cancer, M: Image, and W: Wine. Models are, 1C: 1-component normal(single normal), and 2C: 2-components normal mixture.

Data	Model	Variable lists																	
I	1C	3	4	1	2														
	2C	3	4	1	2														
B	1C	23	21	24	3	1	4	14	13	22	2	5	6	7	8	9	10	11	12
	2C	15	16	17	18	19	20	25	26	27	28	29	30						
M	1C	6	8	12	5	7	10	2	11	14	9	1	3	4	13				
	2C	5	8	12	6	7	10	2	11	9	14	4	1	3	13				
W	1C	7	13	12	10	1	6	4	9	5	2	3	8	11					
	2C	7	13	10	12	1	6	9	4	5	2	3	8	11					

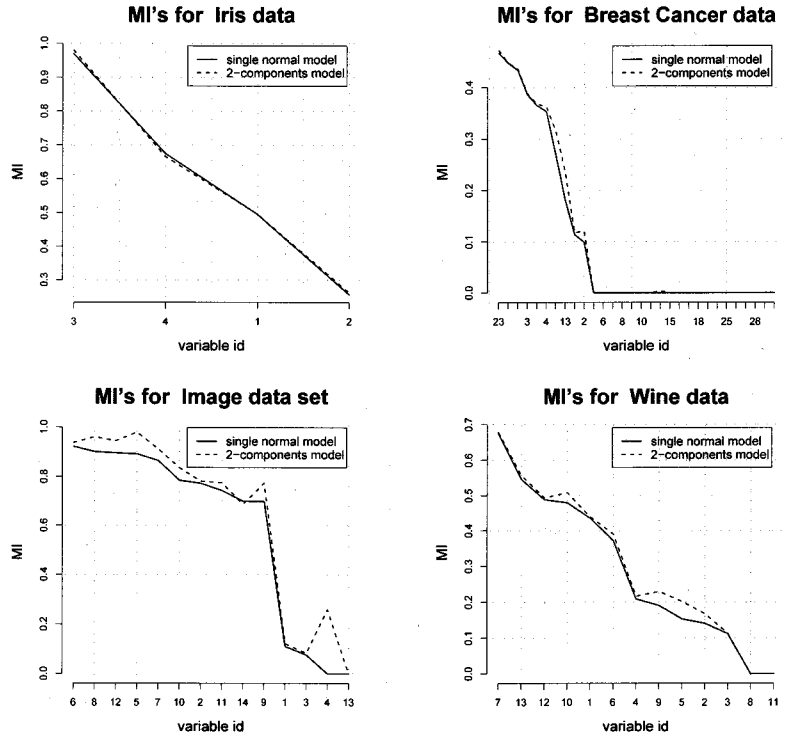

Figure 1: *Univariate MI's of Iris, Breast cancer, Image and Wine data sets under 1-component normal and 2-components normal mixture models. The variable id's are arranged in decreasing order of MI's.*

the variables with larger MI's which will play important roles in subset variable selection. Thus using the minimum description length principle, 1-component normal model is suggested instead of the time-consuming 2-component normal mixture assumption to compute univariate MI's.

To confirm the validity of the univariate MI, visual process of data exploration is provided using DAVIS. For all the plots, colors represent the class information. The variables are rearranged in the order of the magnitude of the univariate MI's with 1-component normal model. Figure 2 is the FEDF

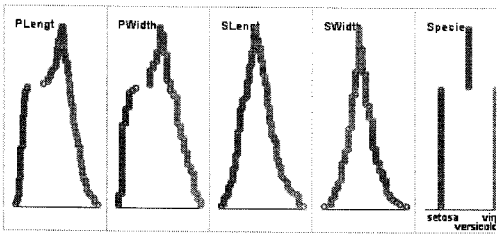

Figure 2: *FEDF plot of the Iris data set.*

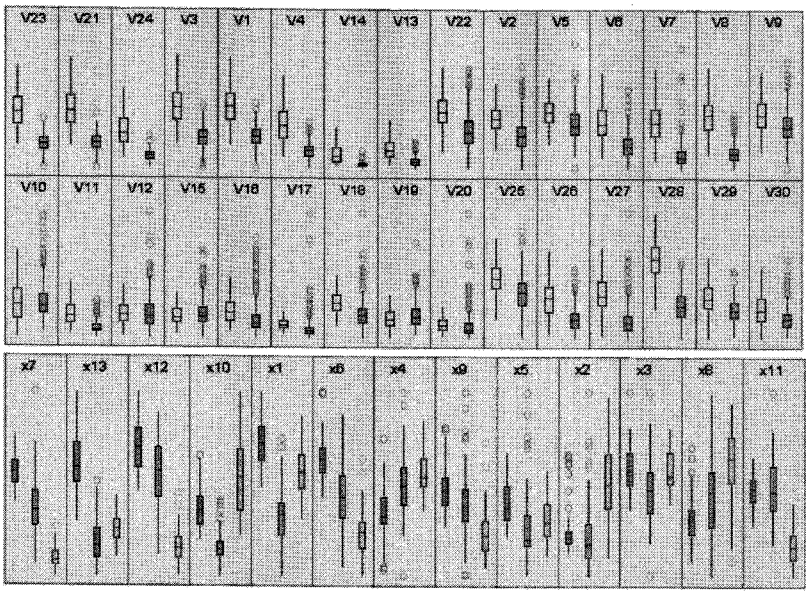

Figure 3: *Box-plots of Breast cancer (top) and Wine (bottom) data sets.*

(Huh, 1995) of the Iris data set. It shows that Petal Length and Petal Width clearly have the power of classifying the class variable while Sepal Length and Sepal Width do not give much information for the class variable. Also the plot shows that the distributions of Petal Length and Width do not have symmetry and have 2 modals. Figure 3 gives the boxplots of the Breast Cancer and the Wine data sets. It can be observed that for the first few variables, say 5 or 6 variables in both data sets, the portion of overlapping area of boxplots is relatively small. Thus they can serve as good predictors for the class variable. In particular, it seems that V23, V21 and V24 in the Breast cancer data set have high classification ability.

However, these variables could be correlated. If they are highly correlated each other and a variables is selected among them, then it would not be suggested to select another variables in this group because they would share common information about the class variable. Also it is well known that the correlation among explanatory variables would reduce the prediction power in many classification models. Thus it would not be desirable to select only variables with large value of univariate MI for the best subset of variables.

Our variable selection strategy could safely avoid such kind of problem. To demonstrate this point. We may refer to Figure 4 which is the scatterplot matrix of some variables in the Breast cancer data. It shows clearly that V23, V21 and V24 have the ability to classify the class variable, but they

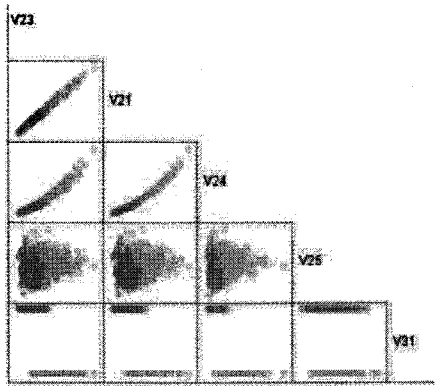

Figure 4: *Scatterplot matrix of some variables in the Breast cancer data set.*

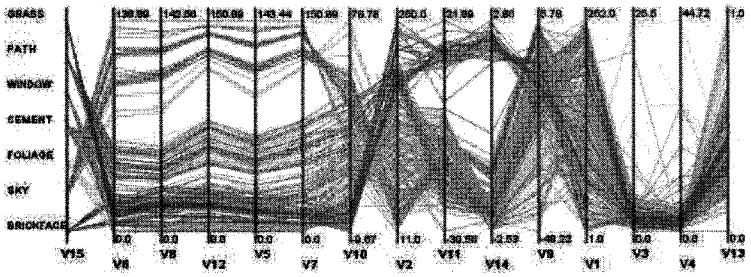

Figure 5: *Parallel plot of the Image data set.*

are correlated each others. Suppose that $X = (V23, V21)$. Then the marginal distribution of X may concentrates most of its mass on a small area constructed by the data points. Since the area with high mass is narrow, $f_X(x)$ would have large value, say greater than 1, on this area, which in tern gives negative or small value of $\widehat{H}(X)$. The same phenomenon may happen for the estimation of $H(X|Y)$. For each value of k, $f_{X|Y}(x|k)$ would be large. Thus both of $H(X)$ and $H(X|Y)$ would have small or even negative values. Hence $\widehat{I}(X;Y)$ would be small or zero. On the other hand, we can observe from Figure 3 that $V25$ have relatively less power than $V21$ or $V24$, but the scatterplot of $V23$ *vs* $V25$ shows that they are uncorrelated and the combination of these two variables might have power to classify the class variable because most data points are grouped by colors. Since the whole data is widely spread than the grouped data, it is obvious that MI$(V23, V25; Y)$ should have large value. In fact the estimates of MI$(V23, V25; Y)$ under 1-component normal with and without independence, and 2-components normal mixture with and without assumptions are $0.519, 0.530, 0.519$ and 0.532, respectively. These values are considerably larger than 0.444, 0.414, 0.419 and 0.450, the estimates of MI$(V23, V21; Y)$. Thus we can safely choose $(V23, V25)$ instead of $(V23, V21)$ which have large univariate MI's.

Figure 5 gives the parallel coordinates of the Image data set. The plot shows that the first 5 variables of large MI's also have high power of classification, and the lines go parallel with each other which suggests the variables are highly correlated. It seems that most of the variables are quite informative for the class variable except 3 variables $(V1, V3, V4)$.

We have shown graphically that the estimates of univariate MI could identify the variables with high classification ability. Now we will demonstrate the subset selection strategy would work well.

Table 2: The summary of subset variable selection and the results of cross validation for each of the 4 data sets and 4 different types of models. The first line for each model is the number of variables selected, second line is the selected variable id's, third line is the % accuracy of the 10-fold cross validation for logistic and J48 decision tree. Considered models are: U: univariate MI, 1CI: 1-component independent, 1CD: 1-component dependent, 2CI: 2-component independent, 2CD: 2-component dependent, A: all variables and Cfs: CfsSubsetEval.

Models	Iris	Breast	Image	Wine
	2	3	4	6
U	3, 4	23, 21, 24	6, 8, 12, 5	7, 13, 12, 10, 1, 6
	96/96	92/90	78/76	94/94
	2	3	4	6
1CI	3, 4	23, 28, 22	6, 14, 2, 1	7, 1, 11, 13, 10, 5
	96/96	96/94	86/90	97/94
	4	4	5	6
1CD	3, 1, 4, 2	23, 25, 22, 15	6, 8, 2, 13, 1	7, 10, 13, 11, 5, 1
	96/96	97/94	90/81	97/94
	4	4	4	5
2CI	3, 4, 1, 2	23, 28, 22, 11	5, 2, 14, 1	7, 1, 11, 13, 5
	96/96	96/94	85/90	96/92
	3	6	5	5
2CD	3, 1, 4	23, 25, 22, 11, 28, 27	5, 2, 14, 11, 6	7, 13, 11, 1, 5
	96/96	96/93	91/88	96/92
	2	9	6	8
Cfs	3, 4	2, 7, 8, 14, 21, 23, 24, 27, 28	2, 6, 9, 11, 13, 14	1,5,6,7,10,11,12,13
	96/96	96/93	89/88	96/94
All	96/96	93/93	86/90	97/94

To this end, we have selected variables from those 4 data sets as the best subsets for classification. The selections are done under 4 different models proposed in the paper with $\epsilon = 0.01$. Table 2 gives the results of the selections. The accuracies of classification, which are measured by the 10-fold cross validation with logistic regression and J48 decision trees, are given. For comparison, we also provide the accuracies of classification based on all variables, the first few number of variables having largest univariate MI's and the subset of variables selected by `CfsSubsetEval` in Weka.

All of the cases, the numbers of variables selected by the 4 proposed models is less than or equal to those by `CfsSubsetEval` while the accuracies of 4 models are comparable. Especially 1-component normal with independence model gives quite good results with simplicity. Thus this model is preferable.

Most of cases, the accuracy could be improved by adding variables. This can be done by adjusting the value of ϵ. For example, if we set $\epsilon = 0$, then the 1-component normal with independent model selects 10 variables from the Wine data set including 6 variables in Table 2. The 10 variables give 99% and 94% classification accuracies. We believe that it is a matter of choice. Perhaps, $\epsilon = 0.01$ is appropriate for most applications.

Based on these observations for subset variable selection, we might conclude following statements.

1. The 1-component multivariate normal with independence model would generally provide reasonable accuracy of classification with simplicity.

2. Classification based on subset selection generally gives higher accuracy than using all the variables.
3. Generally, 3 or 4 number of variables will be enough for subset selection for classification.
4. Subset selection based on univariate MI's give consistently lower accuracy for all the cases considered.
5. When a variable is selected, other variables which are highly correlated with this selected variable should not be selected.
6. A variable with low univariate MI but not being correlated with the variables having high univariate MI's should not be ignorable.

5. Conclusion

We suggested to compute multivariate MI directly by estimating the multivariate density under various forms of normal mixture assumptions, and empirically demonstrated that 1-component normal model with independent assumption is efficient for most of the cases considered. The study also showed that those variables with almost 0 univariate MI's could have high impact on the multivariate MI while those variables with high univariate MI's but highly correlated with each other would have very little impact on the value. We considered only the case of continuous type explanatory variables. However, the complex type can be easily extended by considering the fact $f(X_1, X_2) = f(X_1|X_2)f(X_2)$ where X_1 is continuous and X_2 is discrete. The experimentation was run using *R* (Ihaka and Gentleman, 1996). We are currently working to implement the algorithm using C++ and Java to expedite the computing speed. It will be worthwhile to investigate the proposed models for the data which are complex types and whose volumes are huge like the microarray data.

References

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning, *IEEE Transactions on Neural Networks,* **5**, 537–550.

Brillinger, D. R. (2004). Some data analyses using mutual information, *Brazilian Journal of Probability and Statistics,* **18**, 163–183.

Christensen, R. (1997). *Log-linear Models and Logistic Regression,* Springer, New York.

Collett, D. (2003). *Modelling Binary Data,* 2nd ed., Chapman & Hall/CRC.

Cover, T. M. and Thomas, J. A. (1991). *Element of Information Theory,* John Wiley & Sons.

Darbellay, G. A. (1999). An estimator of the mutual information based on a criterion for independence, *Computational Statistics & Data Analysis,* **32**, 1–17.

Fraley, C. and Raftery, A. E. (2002). MCLUST: Software for model-based clustering, density estimation and discriminant analysis, Technical report No. 415, Department of Statistics, University of Washington.

Huh, M. Y. (1995). Exploring multidimensional data with the flipped empirical distribution function, *Journal of Computational and Graphical Statistics,* **4**, 335–343.

Huh, M. Y. and Song, K. Y. (2002). DAVIS: A Java-based data visualization system, *Computational Statistics,* **17**, 411–423.

Hutter, M. (2002). Distribution of mutual information, In *Advances in Neural Information* Processing Systems 14, editor T. G. Dietterich and S. Becker and Z. Ghahramani, MIT Press, Cambridge, MA, 399–406.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics,* **5**, 299–314. http://www.r-project.org

Joe, H. (1989). Relative entropy measures of multivariate dependence, *Journal of the American Statistical Association,* **84**, 157–164.

Kojadinovic, I. (2005). Relevance measures for subset variable selection in regression problems based on k-additive mutual information, *Computational Statistics & Data Analysis,* **49**, 1205–1227.

Kononenko, I., Simec, E. and Robnik-Sikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with RELIEFF, *Applied Intelligence,* **7**, 39–55.

Lee, S.-C. and Huh, M. Y. (2003). A measure of association for complex data, *Computational Statistics & Data Analysis,* **44**, 211–222.

Liu, H. and Motoda, H. (1998). *Feature Extraction, Construction and Selection: A Data Mining Perspective,* 2nd Printing, Kluwer Academic Publishers.

Merz, C. J. and Murphy, P. M. (1996). UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California, Irvine, CA(http://www.ics.uci.edu/~mlearn/MLRepository.html)

Miller, A. J. (1990). *Subset Selection in Regression*, Chapman & Hall/CRC, London.

Nguyen, H. S. and Skowron, A. (1995). Quantization of real value attributes. *Proceedins of Second Joint Annual Conf. on Information Science, Wrightsville Beach, North Carolina,* 34–37.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal,* **27**, 379–423 and 623–656.

Torkkola, K. and Campbell, W. M. (2000). Mutual information in learning feature transformations, *In Proceeding ICML'2000, The Seventeenth International Conference on Machine Learning.*

Tourassi, G. D., Frederick, E. D., Markey, M. K. and Floyd, C. E. Jr. (2001). Application of the mutual information criterion for feature selection in computer-aided diagnosis, *Medicine Physicist,* **28**, 2394–2402.

Wang, J. (2001). Generating daily changes in market variables using a multivariate mixture of normal distributions, *Proceedings of the 33nd conference on Winter simulation, IEEE Computer Society.*

Witten, I. and Frank, E. (1999). Data Mining, Morgan and Kaufmann. http://www.cs. waikato.ac.nz/ml/weka

Received October 2008; Accepted November 2008