Thesis for the Degree of Master of Engineering

Feature Selection in Supervised Learning Problems using Data Mining Approach

By

Farnaz Pirasteh

Department of Chemical Engineering

The Graduate School

Pukyong National University

August 2015

Feature Selection in Supervised Learning Problems using Data Mining Approach

데이터 마이닝 접근법을 이용한 지도학습 문제에서의 통계적 자질 선택

Advisor: Prof. Jay Liu

By

Farnaz Pirasteh

A thesis submitted in partial fulfillment of the requirement
for the degree of Master of Engineering
in Department of Chemical Engineering, Graduate School,
Pukyong National University

August 2015

Feature Selection in Supervised Learning Problems using Data Mining Approach

A dissertation

By

Farnaz Pirasteh

Approved by:

(Chairman) Do Jin Im

(Member) Prof. Pyung-Hoi Koo

(Member) Prof. Jay Liu

August 2015

Contents

LIST OF TABLES

LIST OF FIGURES

Farnaz Pirasteh

Department of Chemical Engineering, The Graduate School,
Pukyong National University

Abstract

When analyzing high dimensional massive data, it is desirable to identify a few important features that affect a certain outcome of interest using feature selection methods. Feature selection methods play more important role when the number of features (p) far exceeds the number of observations (n), which makes the traditional statistical methods infeasible for data analysis. This study presents quantitative and qualitative analysis results of applying feature selection methods in two case studies, multivariate calibration for determining soil carbonate content and cancer prediction using gene expression data, as a regression and a classification problems, with comparison of their performance on each case study. Feature selection methods compared include Least Angle Regression algorithm (LARS), Least Absolute Shrinkage and Selection operator (Lasso), Genetic Algorithm (GA), and classical methods such as forward and stepwise selection. Selected subsets by each method are used for the input of Support Vector Machines (SVM) for supervised modeling. Root Mean Square Error Prediction (RMSEP), Mean Squared Error (MSE) and Prediction

Error (PE) and Area Under the Curve are quantitative criteria used to compare the methods. Due to a small number of samples, bootstrapping also applied when building models in order to obtain more reliable results. The results of case study 1 show high ability of LARS and Lasso in extracting effective features on carbonate content determination, as well as choosing the most true features (wavenumbers), while the least prediction errors were (RMSEP, MSE, and PE) obtained by applying subset selected by LARS. On the other hand, in case study 2 also LARS and Lasso show high accuracy in predicting prostate cancer.

CHAPTER 1. INTRODUCTION

Technological innovations have had deep impact on scientific research as well as on society during the past several decades and allowed us to collect massive amount of data with very low cost and a short period of time. Now a days, observations with texts, curves, images or movies, along with many other typical variables, are frequently encountered in scientific research due to the technological development. For example, in gene expression data, huge numbers of pixels are in hand with limited number of images. These kinds of examples abound in computational biology, climatology, geology, neurology, health science, economics, to list just a few. One common theme in various fields such as science, engineering and the humanities beside difference in their concerns, is massive and high dimensional data have been collected and it will be costly and time consuming to analyze or extract new knowledge or information from these amount of data. These high dimensional datasets along with many new scientific problems create golden opportunities and significant challenges for the development of statistical sciences. The availability of high dimensional data along with new scientific problems has caused statistical methods to be useful and applicable in these problems regarding dimension reduction and variable selection.

Thus, high-dimensionality has significantly challenged traditional statistical theories. Many new insights need to be unveiled and many new phenomena need to be discovered. There is little doubt that the high dimensional data analysis will be the most important research topic in statistics in the 21st century (Donoho 2000). Variable selection and feature extraction are fundamental to knowledge discovery from massive data. Many variable

selection criteria have been proposed in the literature. Variable selection using Akaike information criterion (AIC) or Bayesian information criterion (BIC), and traditional variable selection methods such as forward and stepwise selection involves a combinatorial optimization problem, which is NP-hard (**N**on-deterministic **P**olynomial-time hard), with computational time increasing exponentially with the dimensionality. The expensive computational cost, especially in case of few observations with hundreds or thousands of variables, makes traditional procedures infeasible for high-dimensional data analysis. Clearly, innovative variable selection procedures are needed to cope with high-dimensionality.

Computational challenges from high-dimensional statistical endeavors forge cross fertilizations among applied and computational mathematics, machine learning, and statistics. For example, (Donoho and Elad 2003) showed that the NP-hard best subset regression can be solved by a least-square problem, which can be handled by a linear programming when the solution is sufficiently sparse. (Balabina and Smirnov 2011) used feature selection methods in biodiesel data as a critical step in data analysis for vibrational spectroscopy (infrared, Raman, or near infrared spectroscopy (NIRS)). Many other studies can be found regarding application of feature selection methods in high dimensional problems.

Subset selection (Kohavi and Wrappers 1997) and ridge regression (Hoerl and Kennard 1970), two well-known techniques for improving the Ordinary Least Squares (OLS) estimates in high dimensional problems, both have drawbacks. Subset selection provides interpretable models but can be extremely variable because it is a discrete process, i.e., regressors are either retained or dropped from the model. Small changes in the data can result in very different models being selected and this can reduce its prediction accuracy.

Tibshirani proposed a new technique, called least absolute shrinkage and selection operator or Lasso (Tibshirani 1996). It shrinks some coefficients and sets others to 0, and hence tries to retain the good features of both subset selection and ridge regression. Lasso was applied to many variable selection problems in high dimensional cases as well as other various applications; Shah considered prediction and estimation using Lasso algorithm (Shah 2012). Zhao and Yu proved that a single condition, which is called the irrepresentable condition, is almost necessary and sufficient for Lasso to select the true model both in the classical fixed p setting and in the large p setting as the sample size n gets large (Zaho and Yu 2006). In 2000 Osborne et. al presented a research which makes two contributions to computational problems associated with implementing the Lasso (Osborne et al. 2000). A new method for variable selection and shrinkage in Cox's proportional hazards model using Lasso was presented by Tibshirani (Tibshirani 1997). This method which is a variation of the Lasso proposal of Tibshirani, reduces the estimation variance while providing an interpretable final model. Zou (Zou 2012) presented the adaptive Lasso and showed that the adaptive lasso enjoys the oracle properties; namely, it performs as well as if the true underlying model were given in advance and can be solved by the same efficient algorithm for solving the Lasso.

On the other hand as an innovation algorithm for Lasso, LARS (Least Angle Regression) was presented by Efron et. al (Efron et al. 2004). They introduced LARS as a useful and less greedy version of traditional forward selection method that uses a simple mathematical formula to accelerate the computations. LARS relates to the classic model-selection method known as forward selection or "forward stepwise regression", described in (Weisberg 1980).

LARS used in a comparison by Wang conducted an extensive simulation study over a vast range of data settings and selection methods (Wang 2009). A study on comparison of different variable selection methods in high dimensional cases as well as a chemometric application of LARS to a near-infrared spectroscopy data was presented by Stodden (Stodden 2012).

As a comparison of this or similar studies to our research, the number of variables dealing with in the problem is much less, as well as more sample size, that makes classical or ordinary feature selection methods useful for extracting effective variables; also bigger sample size let the statistician not to be worry about inaccurate results or small training or testing subset, which we are dealing with in our study.

In this thesis, the issues of variable selection in high dimensional supervised learning problems are addressed using two representative case studies: multivariate calibration for determining soil carbonate content and prostate cancer prediction using gene expression data. The rest of this thesis is organized as follows. In Chapter 2, methods that we use in this study are outlined and briefly described. In chapter 3, problem description and data for each case is explained. Results and conclusion are outlined in chapter 4 and finally in chapter 5 we come to a conclusion.

CHAPTER 2. METHODS

2.1 Feature Selection Methods

Feature selection has become the focus of research in various areas of applications where datasets with tens or hundreds of thousands of features are involved. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry, to list just a few.

Feature selection can be defined as a process that chooses a minimum subset of *m* features from the original set of *p* features, so that the feature space is optimally reduced according to a certain evaluation criterion (Novakovic 2010). As the dimensionality of a domain expands, the number of feature *p* increases. Finding the best feature subset is usually intractable (Domingos and Pazzani 1997) and many problems related to feature selection have been shown to be NP-hard (Blum and Rivest 1992). Researchers have studied various aspects of feature selection. Feature selection algorithms can be categorized into filters (Almuallim and Dietterich 1991), wrappers (Domingos and Pazzani 1997) and embedded approaches (Blum and Langley 1997). Filters methods evaluate quality of selected features, independently from the classification algorithm, while wrapper methods require application of a model to evaluate this quality. Embedded methods perform feature selection during learning of optimal parameters.

The objective of feature selection is three-fold: (1) improving the prediction performance of the predictors, (2) providing faster and more cost-effective predictors, and (3) providing a better understanding of the underlying process that generated the data.

Various feature selection methods were presented for different problems, from simple ones with few number of features and samples to high dimensional data with thousands of features and thousands of sample size, from regression problems to classification ones, and in any research field. Choosing the best feature selection method among all available methods is one of the first challenges in this area.

In this study we cover two classical feature selection methods and also some innovative algorithms, Lasso, LARS and GA. A description of each method is coming in the subsubsections 2.1.1 to 2.1.5.

2.1.1 Forward Selection

One of the classical feature selection approaches is called forward selection (Wilkinson and Dallal 1981). In this approach, one adds features to the model one at a time. At each step, each feature that is not already in the model is tested for inclusion in the model. The most significant of these features is added to the model, so long as its p-value is below some pre-set level. It is customary to set this value above the conventional 0.05 level at say 0.10 or 0.15, because of the exploratory nature of this method.

Thus we begin with a model including the feature that is most significant in the initial analysis, and continue adding features until none of remaining features are "significant" when added to the model. Note that this multiple use of hypothesis testing means that the

real type I error rate for a feature (i.e. the chance of including it in the model given it isn't really necessary), does not equal the critical level we choose. In fact, because of the complexity that arises from the complex nature of the procedure, it is essentially impossible to control error rates and this procedure must be viewed as exploratory.

The process can be summarized as follows:

- Start with no features in the model.
- For all predictors in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than α_{crit} (level of significance).
- Continue until no new predictors can be added.

The α_{crit} is sometimes called the "p-to-remove" and does not have to be 5%. If prediction performance is the goal, then a 15-20% cut-off may work best, although methods designed more directly for optimal prediction should be preferred.

2.1.2 Stepwise Selection

Stepwise selection (Efroymson 1960; Hocking 1976) is a method that allows moves in direction, dropping or adding features at the various steps. Backward stepwise selection involves starting off in a backward approach and then potentially adding back features if they later appear to be significant. The process is one of alternation between choosing the least significant feature to drop and then re-considering all dropped features (except the most recently dropped) for re-introduction into the model. This means that two separate significance levels must be chosen for deletion from the model and for adding to the model. The second significance must be more stringent than the first.

On the other hand, Stepwise selection is a semi-automated process of building a model by successively adding or removing features based solely on the t-statistics of their estimated coefficients. Properly used, the stepwise regression option in Statgraphics (or other stat packages) puts more power and information at your fingertips than does the ordinary multiple regression option, and it is especially useful for sifting through large numbers of potential independent features and/or fine-tuning a model by poking features in or out. Improperly used, it may converge on a poor model while giving you a false sense of security. It's a bit like doing carpentry with a chain saw: you can get a lot of work done quickly, but you may end up doing more harm than good if you don't read the instructions, remain sober, and keep a firm grip on the controls.

2.1.3 Least Absolute Shrinkage and Selection Operator (Lasso)

Suppose that we have data $(\mathbf{x}^i, y_j), i = 1,2, \dots, n$, where $\mathbf{x}^i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ are the predictor features and y_is are the responses. As in the usual regression set-up, we assume either that the observations are independent or that the y_is are conditionally independent given the x_{ij}s. We assume that the x_{ij} are standardized so that $\sum_i \frac{x_{ij}}{N} = 0$, $\sum_i \frac{x^2{}_{ij}}{N} = 1$.

Letting $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$, the Lasso estimate $(\hat{\alpha}, \widehat{\boldsymbol{\beta}})$ is defined by

$$(\hat{\alpha}, \widehat{\boldsymbol{\beta}}) = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_j |\beta_j| \le t. \qquad (2-1)$$

Here $t \ge 0$ is a tuning parameter and α and β are unknown parameters. Now, for all t, the solution for α is $\hat{\alpha} = \bar{y}$. We can assume without loss of generality that $\bar{y} = 0$ and hence omit

α. Computation of the solution to equation (2-1) is a quadratic programming problem with linear inequality constraints. The parameter $t \geq 0$ controls the amount of shrinkage that is applied to the estimates. Let $\hat{\beta}_j^{\circ}$ be the full least squares estimates and let $t_0 = \sum |\hat{\beta}_j^{\circ}|$. Values of $t < t_0$ will cause shrinkage of the solutions towards 0, and some coefficients may be exactly equal to 0. For example, if $t = t_0/2$, the effect will be roughly similar to finding the best subset of size $p/2$.

The motivation for the lasso came from an interesting proposal of (Breiman 1993). Breiman's non-negative garotte minimizes

$$\sum_{i=1}^{n}\left(y_i - \alpha - \sum_j c_j \hat{\beta}^{\circ} x_{ij}\right)^2 \quad \text{subject to } c_j \geq 0\,, \qquad \sum_j c_j \leq t. \qquad (2-2)$$

The garotte starts with the OLS estimates and shrinks them by non-negative factors whose sum is constrained. In extensive simulation studies, Breiman showed that the garotte has consistently lower prediction error than subset selection and is competitive with ridge regression except when the true model has many small nonzero coefficients.

A drawback of the garotte is that its solution depends on both the sign and the magnitude of the OLS estimates. In overfit or highly correlated settings where the OLS estimates behave poorly, the garotte may suffer as a result. In contrast, the lasso avoids the explicit use of the OLS estimates. Frank and Friedman in 1993 proposed using a bound on the L_q -norm of the parameters, where q is some number greater than or equal to 0; the lasso corresponds to q = 1 (Frank and Friedman 1993).

Orthonormal Design Case

Insight about the nature of the shrinkage can be gleaned from the orthonormal design case. Let $\mathbf{X}$ be the n × p design matrix with ij-th entry x_{ij}, and suppose that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, the identity matrix.

The solutions to equation (2-2) are easily shown to be

$$\hat{\beta}_j = \text{sign}\left(\hat{\beta}_j^\circ\right)\left(\left|\hat{\beta}_j^\circ\right| - \gamma\right)^+ \qquad (2-3)$$

where γ is determined by the condition $\sum\left|\hat{\beta}_j\right| = t$. Interestingly, this has exactly the same form as the soft shrinkage proposals (Donoho and Johnstone 1994; Donoho et al. 1995); which applied to wavelet coefficients in the context of function estimation. The connection between soft shrinkage and a minimum L_1-norm penalty was also pointed out for non-negative parameters in the context of signal or image recovery.

In the orthonormal design case, best subset selection of size *k* reduces to choosing the k largest coefficients in absolute value and setting the rest to 0. For some choice of δ this is equivalent to setting $\hat{\beta}_j = \hat{\beta}_j^\circ$ if $|\hat{\beta}_j^\circ| > \delta$ and to 0 otherwise. Ridge regression minimizes

$$\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \delta\sum_j \beta_j^2 \qquad (2-4)$$

or, equivalently, minimizes

$$\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 \text{ subject to } \sum \beta_j^2 \le t. \qquad (2-5)$$

The ridge solutions are

$$\frac{1}{1+\delta}\hat{\beta}_j^\circ$$

Where γ depends on δ or t. the garotte estimates are

$$\left(1 - \frac{\gamma}{\hat{\beta}_j^{\circ 2}}\right)^{+} \hat{\beta}_j^{\circ}.$$

Fig. 2.1. shows the form of these functions. Ridge regression scales the coefficients by a constant factor, whereas the lasso translates by a constant factor, truncating at 0. The garotte function is very similar to the lasso, with less shrinkage for larger coefficients. As our simulations will show, the differences between the lasso and garotte can be large when the design is not orthogonal.

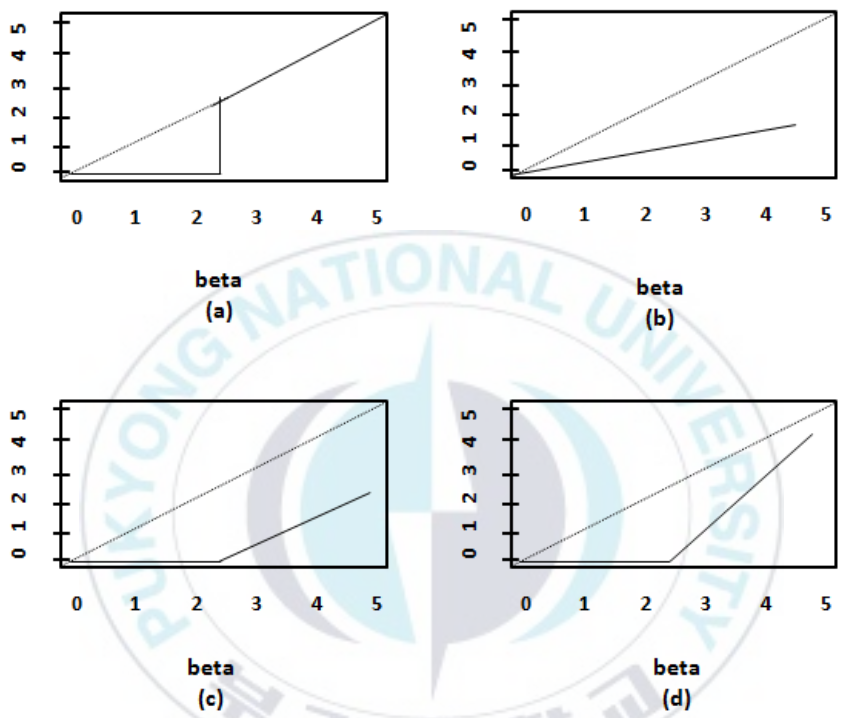

Figure 1.1. (a) Subset regression, (b) ridge regression, (c) the lasso and (d) the garotte: ———, form of coefficient shrinkage in the orthonormal design case; ....., 45°-line for reference. (Source: Tibshirani, 1996)

Geometry of Lasso

It is clear from Fig. 2.1. why the lasso will often produce coefficients that are exactly 0. Why does this happen in the general (non-orthogonal) setting. And why does it not occur with ridge regression, which uses the constraint $\sum \beta_j^2 \leq t$ rather than $\sum |\beta_j| \leq t$ Fig. 2.1. provides some insight for the case p =2.

The criterion $\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2$ equals the quadratic function

$$\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\circ}\right)^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\circ}\right)$$

(plus a constant). The elliptical contours of this function are shown by the full curves in Fig. 2.1. (a); they are centered at the OLS estimates; the constraint region is the rotated square. The lasso solution is the first place that the contours touch the square, and this will sometimes occur at a corner, corresponding to a zero coefficient. The picture for ridge regression is shown in Fig. 2.1. (b): there are no corners for the contours to hit and hence zero solutions will rarely result.

An interesting question emerges from this picture: can the signs of the lasso estimates be different from those of the least squares estimates $\hat{\beta}_j^{\circ}$ Since the features are standardized, when p =2 the principal axes of the contours are at $\pm 45°$ to the co-ordinate axes, and we can: how that the contours must contact the square in the same quadrant that contains $\hat{\beta}_j^{\circ}$. However, when $p > 2$ and there is at least moderate correlation in the data, this need not be true. Fig. 2.2. shows an example in three dimensions. The view in Fig. 2.2. (b) confirms that the ellipse touches the constraint region in an octant different from the octant in which its centre lies.

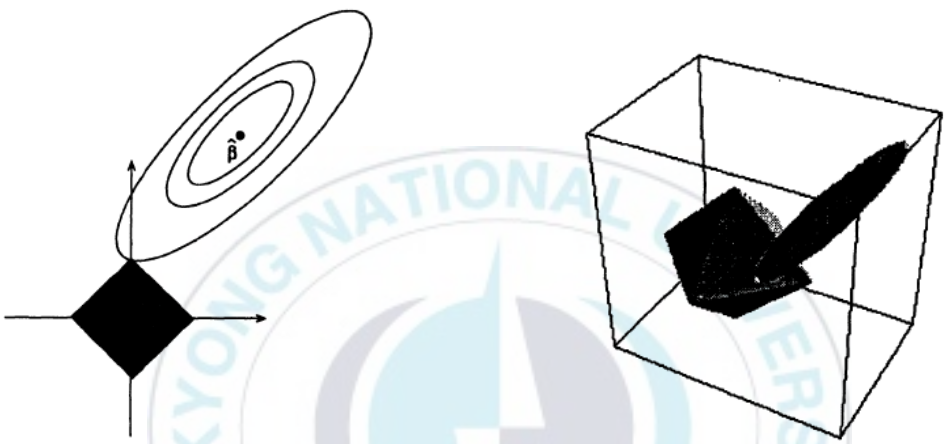

Figure. 2.2. Estimation picture for the Lasso (left), Example in which the Lasso estimate falls in an octant different from the overall least squares estimate (right) (Source: Tibshirani, 1996.)

2.1.4 Least Angle Regression (LARS)

The LARS algorithm works well especially in the case of $p \gg n$. It's a new feature selection algorithm, which is guaranteed to find the best ranking among all possible inputs. The LARS provides a stepwise approximation to Lasso.

Least Angle Regression (LARS) relates to the classic model-selection method known as Forward Selection, or "forward stepwise regression," described in (Weisberg 1980; Section 8.5) given a collection of possible predictors, we select the one having largest absolute correlation with the response y, say x_{j1}, and perform simple linear regression of y on x_{j1}. This leaves a residual vector orthogonal to x_{j1}, now considered to be the response. We project the other predictors orthogonally to x_{j1} and repeat the selection process. After k steps this results in a set of predictors $x_{j1}, x_{j2}, \ldots, x_{jk}$ that are then used in the usual way to construct a k-parameter linear model. Forward Selection is an aggressive fitting technique that can be overly greedy, perhaps eliminating at the second step useful predictors that happen to be correlated with x_{j1}.

The LARS procedure works as follows. As with classic forward selection, it starts with all coefficients equal to zero, and find the predictor most correlated with the response, say x_{j1}. Then, it takes the largest step possible in the direction of this predictor until some other predictor, say x_{j2}, has as much correlation with the current residual. At this point LARS parts company with forward selection. Instead of continuing along x_{j1}., LARS proceeds in a direction equiangular between the two predictors until a third feature x_{j3} earns its way into the "most correlated" set. LARS then proceeds equiangularly between x_{j1}, x_{j2} and x_{j3}, that

is, along the "least angle direction," until a fourth feature enters, and so on. Unlike Lasso, features cannot be removed from active set.

LARS builds up estimates $\hat{\mu} = \mathbf{X}\hat{\boldsymbol{\beta}}$ in successive steps, each step adding one covariate to the model, so that after k steps just k of the $\hat{\beta}_j$'s are non zero. Fig. 4.2. illustrates the algorithm in the situation $m = 2$ covariates (m is the number of covariates), $X = (x_1, x_2)$. In this case the current correlations depend only on the projection $\bar{y}_2$ of y into the linear space $\pounds(X)$ (L1 norm) spanned by x_1 and x_2

$$c(\hat{\mu}) = X'(y - \hat{\mu}) = X'(\bar{y}_2 - \hat{\mu}).$$

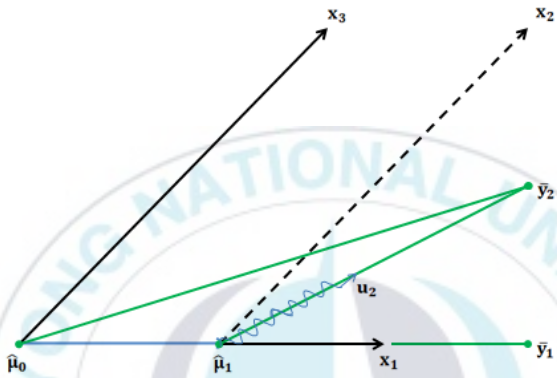

Figure 2.3. The LARS algorithm in case of m=2 covariates; $\bar{\mathbf{y}}_2$ is the projection of $\mathbf{y}$ into $\pounds(\mathbf{x_1}, \mathbf{x_2})$. Beginning at $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$, the residual vector $\hat{\mathbf{y}}_2 - \hat{\boldsymbol{\mu}}_0$ has greater correlation with $\mathbf{x_1}$ than $\mathbf{x_2}$; the next LARS estimate is $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\gamma}}_1 \mathbf{x_1}$, where $\hat{\boldsymbol{\gamma}}_1$ is chosen such that $\hat{\mathbf{y}}_2 - \hat{\boldsymbol{\mu}}_1$ bisects the angle between $\mathbf{x_1}$ and $\mathbf{x_2}$; **then** $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\gamma}}_2 \mathbf{u_2}$, where $\mathbf{u_2}$ is the unit bisector; $\hat{\boldsymbol{\mu}}_2 = \hat{\mathbf{y}}_2$ in the case m=2, but not for case m>2. (Source: Efron, *et al*., 2004).

2.1.5 Genetic Algorithm

An interesting heuristic search algorithm well suited for the combinatorial optimization problem is the genetic algorithm (GA). Although GA is not widely known among statisticians, it has garnered some interest in this community. Chatterjee, Laudato, and Lynch gave an introductory review and showed how the GA can be applied to a number of classical problems in statistics.

The optimization strategy used by the GA is very simple. Start with a number of randomly generated candidates (the initial population). Using the Darwinian principle of "the survival of the fittest," the weaker (less optimal) candidates are gradually eliminated and the stronger ones are allowed to survive and generate offspring. This goes on for a number of generations until, in the end, good solutions are produced. In a typical setting, each individual ω is represented by a binary string, say of length p, which is treated as the genetic code (DNA) of ω; each position can be regarded as a single gene. Starting with a randomly generated population of size m, $\{\omega_1, \omega_2 \ldots, \omega_m\}$, a new generation is produced with three genetic operations: selection, reproduction, and mutation.

Selection. Each individual is evaluated by a fitness function, F, often the objective function for the underlying optimization problem, and assigned a score. When the goal is to maximize (minimize) F, those with high (low) scores are given higher likelihoods of surviving to the next generation (Table 2.1).

Reproduction. Two individuals are selected at random to produce a child. Typically, a cross-over position is chosen at random between 1 and p, say j; the child then inherits the

first j genes from the father and the rest p_j genes from the mother. The cross-over position in the illustration is between 3 and 4 (Table 2.2).

Mutation. At birth, an individual is allowed, with a certain small probability (called the mutation rate), to alter its genetic code at a randomly chosen position. In the typical setting where binary codes are used, this amounts to flipping a 0 to a 1 and vice versa (Table 2.3). Remark. For the feature selection problem, each ω_i represents a different subset of features, and hence a different model and the entire population $\{\omega_1, \omega_2 \ldots, \omega_m\}$ together specifies a total of m different models.

Table 2.1. An Illustration of Selection

	1	2	3	4	5	⋯	p		Score
ω_1	1	1	0	0	0	⋯	0		s_1
⋮				⋮					⋮
ω_i	0	0	1	1	1	⋯	1	$\overset{F}{\Longrightarrow}$	s_i
ω_{i+1}	1	0	0	1	0	⋯	0		s_{i+1}
⋮				⋮					⋮
ω_m	0	1	0	0	0	⋯	1		s_m

Table 2.2. An Illustration of Reproduction

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Father	*1*	*1*	*0*	0	0	1	0	0
Mother	0	0	1	*1*	*1*	*0*	*0*	*1*
Child	1	1	0	1	1	0	0	1

Table 2.3. An Illustration of Mutation

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Before	0	0	1	1	1	0	0	1
					↓			
After	0	0	1	1	0	0	0	1

A waited random selection is applied to the original population where the probability of a particular subset (chromosome) being selected is a function of its cost function response. Thus chromosomes with a good cost function response will have a greater chance of selection. Using this method two of the chromosomes are selected and 'mated', swapping sections of their respective gene sequences. This process produces two new 'child' chromosomes inheriting characteristics of their 'parents'. These child chromosomes are then subjected to random from a '1' to a '0' or vice versa. The probability of this change is normally very small. The process of selection followed by mutation is then repeated until n new chromosomes are created. The cost function is then evaluated for each of the chromosomes and the whole process repeats itself. The algorithm continues until a stopping criterion is reached. For example, this may be that a given cost function response is met, a certain number of generations has passed, or the chromosomes have converged to a similar configuration.

However, there are various ways of carrying out each step. A full explanation of the subject of GAs can be found in (Goldberg 1989; Holland 1992). Tutorial review of their applications in chemometrics and analytical chemistry are available in (Lucasius and Kateman 1993), while the use of GAs for feature selection in spectroscopy has been described, for instance, (Rimbaud 1995; Lucasius et al. 1994). Other studies of wavelength selection in spectroscopy include (Breiman 1993; Weisberg 1980; Dowell 1994; Brown1992; Horchner and Kalivas 1995; Kalivas et al. 1989; Messick 1994; Sutter and Kalivas 1988; Sutter and Kalivas 1992; Brereton and Elbergali 1994; Lindgren et al. 1995). A general schema of Genetic Algorithm process is shown in Fig. 2.4.

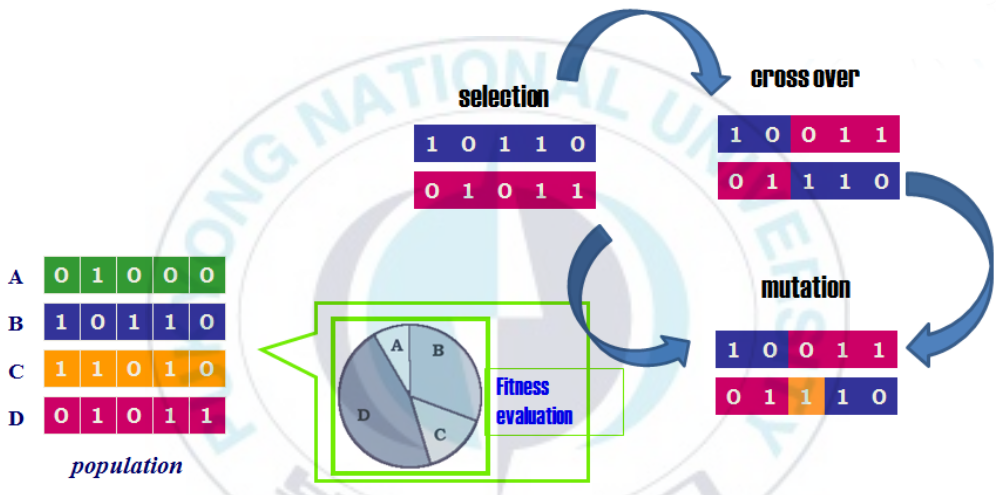

Figure 2.4. GA algorithm process

2.2 Support Vector Machines

2.2.1 SVM in Regression problems

A training data set is defined by D={($\mathbf{x}_1$,$\mathbf{y}_1$),…,($\mathbf{x}_n$,$\mathbf{y}_n$)} where $\boldsymbol{x}_i, \boldsymbol{y}_i \in R^N$ and n is the number of observations. Support Vector Machines (SVMs) are learning machines, which mean that a linear function of f($\mathbf{x}$) = $\mathbf{w}^T\Phi(\mathbf{x})$+b is used to solve the classification or regression problems in a higher dimensional version of x, Φ(x). To consider a regression problem in this study, the best line is defined to be a line which minimizes the following cost function:

$$\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 + C\sum_{i=1}^{n}(\vartheta_i) \qquad (2-6)$$

subject to:

$$\begin{cases} \mathrm{t_i} - (\mathrm{w}^T \cdot \Phi(\mathbf{x}) + \mathrm{b}) \leq \varepsilon + \vartheta_i \\ (\mathrm{w}^T \cdot \Phi(\mathbf{x}) + \mathrm{b}) - \mathrm{t_i} \leq \varepsilon + \vartheta_i^* \\ \vartheta_i \ , \vartheta_i^* \geq 0 \end{cases} \qquad (2-7)$$

where ϑ_i is the corresponding error at the ith point, C is the penalty parameter up to ε and $\Phi(\mathrm{x_i})$ maps x_i into a higher-dimensional space (Cristianini and J. S. Taylor 2000).

2.2.2 SVM in Classification Problems

A training data set is defined by D={($\mathbf{x}_1$,y_1),…,($\mathbf{x}_n$,y_n)} where $\boldsymbol{x}_i \in R^N$, $y_i \in \{-1,1\}$and n is the number of training data points. Support Vector Machines (SVMs) are learning machines, which mean that a linear function of f($\mathbf{x}$) = $\mathbf{w}^T\Phi(\mathbf{x})$+b is used to solve the classification

problems in a higher dimensional version of **x**, Φ(**x**). The best line is defined to be a line which minimizes the following cost function:

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n}\vartheta_i$$

where,

$$\begin{cases}\boldsymbol{w}^T\Phi(\mathrm{x_i}) + \mathrm{b} \geq +1 - \vartheta_i, & if\ y_i = +1\\ \boldsymbol{w}^T\Phi(\mathrm{x_i}) + \mathrm{b} \leq -1 + \vartheta_i, & if\ y_i = -1\end{cases} \tag{2-8}$$

where ϑ_i is the corresponding error at the ith point, C is the penalty parameter and $\Phi(\mathrm{x_i})$ maps x_i into a higher-dimensional space.

The minimization of (8-2) is a standard problem in optimization theory: minimization with constraints. This can be solved by applying the Lagrangian theory. With the help of Lagrange theory, the dual formulation becomes:

$$\begin{aligned}\max \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j)\\ s.t. \quad & \sum_{i=1}^{n}\alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad \forall i\end{aligned} \tag{2-9}$$

According to α_i Lagrange multipliers, the decision function is written as following:

$$y = \mathrm{sgn}[\sum_{i=1}^{n}\alpha_i\, y_i K(x_i, x) + b] \tag{2-10}$$

where $K(x_i, x)$ is named the kernel function. In this study, the following Radial Basis Function (RBF) was used:

$$K(x_i, x_j) = \exp(-\gamma\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2) \tag{11-2}$$

where γ is the kernel parameter (Seeja and Shweta 2011)

2.3 Bootstrapping

Bootstrapping introduced by Bradley Efron in 1979, is a resampling technique used to obtain estimates of summary statistics (Efron 1979).

Adèr et al. recommend the bootstrap procedure for the following situations:

- When the theoretical distribution of a statistic of interest is complicated or unknown.
- When the sample size is insufficient for straightforward statistical inference.
- When power calculations have to be performed, and a small pilot sample is available.

In this study regarding limited number of sample size, which satisfies second Ader's recommendations, bootstrapping can be an efficient method to evaluate the model validity.

Suppose a random sample $x = (x_1, \dots, x_n)$ from an unknown probability distribution F. A bootstrap sample is defined to be a random sample of size n drawn from $\hat{F}$ say $x^* = (x_1^*, \dots, x_n^*)$, where $\hat{F}$ is the empirical distribution defined by

$$F(x) = \frac{\text{number of values in x equals to x}}{n}$$

For each bootstrap sample x^* there is a bootstrap replicate of $\hat{\theta}$,

$$\hat{\theta}^* = s(x^*)$$

which θ is the parameter of interest we wish to estimate on the basis of x; For this purpose, we calculate an estimate $\hat{\theta} = s(x)$ from x.

The bootstrap estimate of $SE_{\hat{F}}(\hat{\theta})$ is defined by

$$SE_{\hat{F}}(\hat{\theta}^*)$$

This is called the ideal bootstrap estimate of the standard error of $s(x)$.

In this study, we used boot package in R to apply bootstrapping to the dataset.

CHAPTER 3. PROBLEM DESCRIPTION

3.1. Case study I: Multivariate Calibration for Soil Carbonate Content Prediction

3.1.1 Sampling

Carbonates are a key component of soils influenced by calcareous parent material, as it influences both chemical and physical soil properties and thus fertility and productivity (Schlichting et al.1995), as well as soil organic carbon (SOC) cycling. Carbonate determination is of increasing practical relevance, e.g. for The Intergovernmental Panel on Climate Change (IPCC) greenhouse gas inventories and national reporting schemes (Kamogawa et al. 2001). Common approaches aiming at SOC quantification are based on direct determination of SOC or determination of total carbon (TC) concentrations and subsequent subtraction of soil inorganic carbon (SIC) (Bruckman et al. 2011). Simple methods, such as the well-known Walkely-Black Method (Cools and Vos 2010) or modifications of it and those based on loss-on ignition (LOI) are sufficiently accurate for certain research questions.

However, there are various sources of errors, ranging from the quality of organic matter (e.g. humification status), the associated inorganic matter (type and quantity of clay minerals) as well as both, SOC and SIC concentrations and their mass ratio, which cannot be tolerated for a number of applications. Since a rising number of laboratories are equipped with elemental analyzers, dry combustion methods with subsequent IR-analysis of combustion gases are

preferably chosen if higher accuracies are desired. As it gives precise estimates on TC, the challenge of accurately measuring SIC still remains. The gas-volumetric Scheibler method (Huber 2004) is internationally recognized and commonly used, despite its relative low analytical precision and sensitivity towards the actual type of carbonate present in the respective sample material.

A number of spectroscopic approaches were developed in the recent years to overcome wet-chemical shortcomings and to provide a standardized laboratory protocol. These methods are based on spectroscopy (Fourier transform infrared (FT-IR) or Raman) or X-ray diffractometry (XRD) which have both strengths and weaknesses. While FT-IR and Raman are unable to clearly separate different types of carbonates, XRD has problems with sensitivity associated with very small concentrations in the sample material. Check (Kamogawa 2001) as a good overview for further reading.

Our case study is based on a dataset of FT-IR and XRD spectra previously collected and used in (Bruckman, K. Wriessnig 2013), which developed a model based on Partial Least Squares Regression (PLSR) using areas of certain peaks related to soil carbonates. The number of features notably far exceeds the number of observations (samples) in our case study. Consequently, this high dimensional dataset with $p \gg n$ requires distinct methods for feature selection.

3.1.2 FT-IR & XRD data

Samples for this study are Fourier transform infrared (FT-IR) and XRD data from a study on soil carbonate determination. For more details regarding soil sample such as sampling

protocol, please refer (Bruckman and Wriessnig 2013). A FT-IR spectrum and an XRD diffractogram of each soil sample were used for this study: A FT-IR spectrum consists of 7,466 wavenumbers and corresponding transmittance values (%), and an XRD diffractogram consists of 3,890 diffraction angles (2θ degrees) and corresponding intensity values. Fig. 3.1 and Fig. 3.2 represent a FT-IR spectrum and an XRD diffractogram for one sample.

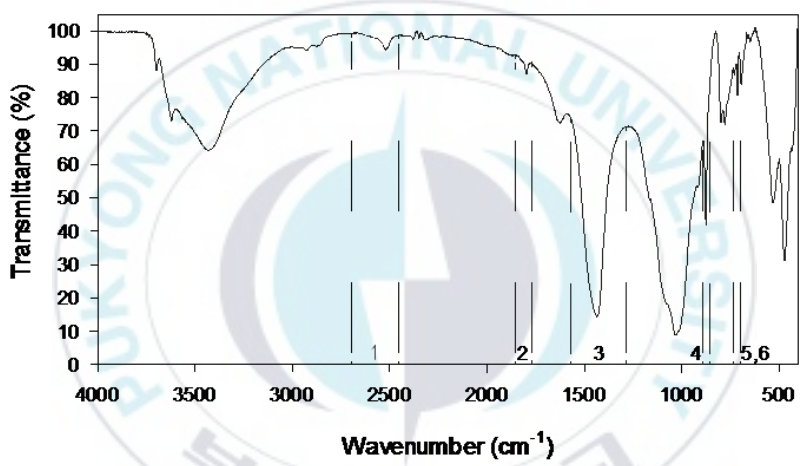

Figure 3.1. Pure carbonate and bulk soil infrared spectra: FT-IR (Source: Bruckman and Wriessnig, *et al*., 2013)

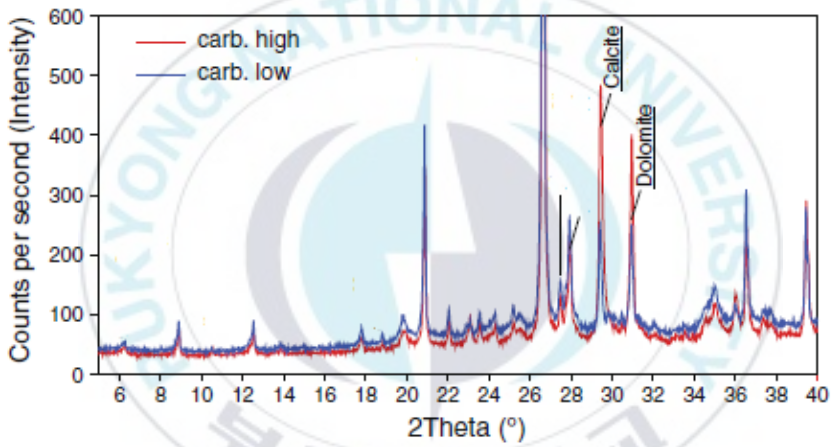

Figure 3.2. X-ray diffractogram of the same carb: XRD (Source: Bruckman and Wriessnig, *et al.*, 2013)

Two different cases are considered to compare the feature selection techniques: FT-IR (Case I) and FT-IR+XRD (Case II). Including XRD data to our main dataset (FT-IR) can be a good case study to check robustness of feature selection methods. Robustness is defined as "the ability of a system to resist changes without adapting its initial stable configuration" (Huber 2004); so if the created methods are not robust enough, adding additional features must affect the prediction error and efficiency of system. Finally the purpose is to check which feature selection technique can choose relevant peaks as important features and which algorithm has more relevant peaks among selected features to predict soil carbonate more efficiently and precisely.

3.1.3 Performance evaluation of feature selection methods

The challenge in this study is to find out important wavenumbers that affect soil carbonate determination. For this reason, different feature selection methods have been applied to the data in Case I and Case II and compared qualitatively as well as quantitatively. As a qualitative measure, the number of true features selected was counted considering relevant peaks in FT-IR and XRD data. As mentioned above, peaks in FT-IR and XRD spectra that are relevant to soil carbonates are already known through literature.

Considering the relevant peaks as true features which are included in the selected subsets, the challenge will be how many true features will be selected in each subset extracted by different feature selection methods. For this reason, comparing the number of true features

selected by different feature selection methods can be a good way of comparing them qualitatively as well as quantitatively.

For soil carbonate content prediction, we will apply SVM as a regression technique to all subsets obtained by LARS, Lasso and GA for Case I and Case II. To compare the accuracy of the methods, criterions such as RMSEP (Root Mean Squared Error of Prediction) (Armstrong and Collopy1992), MSE (Mean Squared Error) (Erdogmus and J.C. Principe 2002) and Percentage of Error (PE) (Swanson 2011) are considered; RMSEP is defined as:

$$\mathrm{RMSEP} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{y_i} - \mathrm{y_i'})^2$$

Where y is the actual target value, y′the predicted target value and n is the number of samples. Also MSE is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{y_i} - \mathrm{y_i'})^2$$

and PE as follows:

$$PE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i - y_i')}{y_i}\right|$$

where y and y′ are defined same in RMSEP.

Then the lower values of RMSEP, MSE and PE for each method, the more efficient carbonate content prediction. Moreover, as the sample size is very small (41 soil samples), confidence intervals of the criterions need to be considered which is obtained by bootstrapping techniques (Efron 1979).

3.2. Case Study II: Diagnosing Prostate Cancer using DNA Microarray Gene Expression Data for

Deoxyribonucleic acid (DNA) micro array technology provides tools for studying the expression levels of a large number of distinct genes simultaneously (J. Chen and H. Chen 2003). Micro array technology allows biologists to simultaneously measure the expressions of thousands of genes in a single experiment (Lee and C.H. Chao 2003; Seeja and Shweta 2011; Yang and Thorne 2003).

Gene expression data is widely used in disease analysis and cancer diagnosis (Yang et al. 2008). Gene expression data from DNA micro arrays are characterized by many measured features (genes) on only a few observations (experiments) although both the number of experiments and genes per experiment are growing rapidly (Zheng at al. 2008; Nguyen, D.M. Rocke 2002). Gene expression data from DNA micro array can be characterized by many features (genes), but with only a few observations (experiments). Prediction, classification, and clustering techniques are being used for analysis and interpretation of the data (Nguyen and D.M. Rocke 2001). An important application of gene expression micro array data is classification of biological samples or prediction of clinical and other outcomes (Dai et al. 2006). Micro array technology is to classify the tissue samples using their gene expression profiles as one of the several types (or subtypes) of cancer. Compared with the standard histopathological tests, the gene expression profiles measured through micro array technology provide accurate, reliable and objective cancer classification.

In this study, 148 prostate samples, with various amounts of tumor, stroma, BPH and atrophic gland, were used for this study from Vaccine Research Institute of San Diego (Sharma and K. Paliwal 2008).

Prostate cancer gene expression profiles were studied in this project. A total RNA from 148 prostate samples with various amount of different cell types were hybridized to Affymetrix U133A arrays. The percentage of different cell types vary considerably among samples and were determined by pathologist.

Dataset reference is “Dataset GSE8218” which is related to GEO (Gene Expression Omnibus) and is available in The National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/geo/query/acc.cgiacc=GSE8218.

In GSE8218 dataset, RNA from 148 prostate samples is considered. The total number of gene expressions is 22,283 which need to be extracted from GEO and then be normalized in order to change it to data arrays for being able of numeric analysis.

In the dataset, tumor cell percentage (the index which significantly relates to overall survival) (p) considers to distinguish and divide patients; so for 71 patients, p=0, 56 patients p= 0.1%~0.8% and 12 patients were not reported; so 71 person are considered as healthy and 65 are considered as non-healthy (measures are classified as 0 (no cancer) and 1(cancer)) and finally 136 samples will be included in this study.

3.2.1. Performance evaluation of feature selection methods

In this case, since it’s a medical case study related to cancer detection, it is the main challenge to find true samples related to each category and the most important features impact them. So among more than 22,000 features that are considered as factors affect prostate cancer in 148 samples, the aim is to find out most effective gene expressions that

can impact prostate cancer; so using LARS, LASSO and GA as feature selection methods in high dimensional data, a dimension reduction as well as feature selection can be gained for this study. Also to show a comparison of these methods with classical methods, results from forward selection and stepwise selection also are presented.

Then, to compare efficiency of each method, SVM is applied as the modeling algorithm and finally the best method is chosen based on the highest area under the curve (AUC) in ROC curve (Receiver operating characteristic curve) which is a graphical plot that illustrates the performance of a binary classifier (here defines as 0 for no cancer and 1 for cancer in target variable).

In ROC, an area of 1 represents a perfect test; so the more AUC is closer to 1, the more accurate is the feature selection methods applied to gene expression data.

CHAPTER 4. RESULTS AND CONCLUSION

Results obtained from each method including quantitative results and qualitative results are presenting in chapter 4.

4.1 Results for Case Study I: FT-IR & XRD Data

In the first case study, considering FT-IR and XRD datasets, total 7,466 features (i.e., wavenumbers) in Case I and 11,356 features (7,466 wavenumbers in a FT-IR spectrum plus 3,890 diffraction angles in an XRD diffractogram) in Case II were used to apply feature selection techniques. For Case I and II, a training set of 30 samples was obtained from total 41 samples by random selection and the remaining 11 samples were used as a testing set. Therefore, for Case I the training set is a 7466×30 matrix and the testing set is a 7466×11 matrix; as well as 11356×30 as training set and 11356×11 as testing set for Case II.

As discussed before, the challenge for this case study is to find the most important features affect soil carbonate content prediction and at the same time to check if our methods can select the true features (features based on peaks) or not. Five feature selection methods were applied to Case I and Case II and each method selected a special number of features; some methods chose more true features that others and some methods only a few true features. Comparing methods would be regarding the true features each method includes in the subset selection and also by the prediction error after modeling using subset selected.

Table 4.1. shows the results obtained for features selection algorithms, forward selection and stepwise selection as traditional methods, LARS and Lasso as new and algorithms in case of high dimensional data and GA as an innovative and well known feature selection algorithm which is used widely in various fields of research to compare with performance of LARS & Lasso.

Table 4.1. Results obtained for each feature selection methods in Case study I

	Case I					**Case II**				
	Lasso	LARS	GA	Forward selection	Stepwise selection	Lasso	LARS	GA	Forward selection	Stepwise selection
RMSEP	4.08	3.77	4.59	8.06	8.10	5.78	5.18	6.34	9.13	9.18
MSE	5.71	5.09	6.73	7.24	7.42	7.34	7.07	8.9	8.34	8.47
PE	23.16	23.48	30.09	34.20	32.41	31.42	29.74	34.43	38.42	39.74
#of features selected	47	44	62	147	159	67	65	78	158	160
#of true features selected	31	39	30	4	7	12	16	15	3	5

In Table 4.1. we can see the results obtained for each feature selection method for Case I and Case II. Classical methods, forward selection and stepwise selection, as well as innovative and new methods, Lasso, LARS and GA, were applied to data and results are included here. To compare the efficiency of each method, three factors were considered in this table: 1- number of features selected in each method, 2- number of true features selected in each method and 3- calculating prediction error after SVM modeling on all subset selected by different methods. Prediction error was calculated mathematically based on RMSEP, MSE and PE, the famous criteria for comparing results of modeling.

Case I:

As it is shown in the Table 4.1., in Case I, RMSEP got its lowest value in LARS method (3.77), also for Lasso and GA, RMSEP is 4.08 and 4.59; which does not have significant difference but we can see in LARS the lowest RMSEP and so we can result in LARS is the best feature selection method among new methods for Case I. Looking at the classical methods, RMSEPs are 8.10 for stepwise selection and 8.06 for forward selection which has a significant difference with LARS, Lasso and GA. Although results show a little better performance by forward selection comparing stepwise selection, but we can conclude that LARS is the best method regarding RMSEP.

Regarding MSE and PE, also as explained above by looking at Table 4.1., values for LARS are 5.09 and 23.48; comparing to other methods even GA and Lasso which got higher MSE and PE, so considering this criteria also LARS is the best method among our feature selection methods here.

The important element here is the number of true features selected by each method. For classical methods we see a few true features selected by the method; such that in forward

selection only 4 features is among selected subset and in stepwise selection only 7. Even if stepwise selected shows a better performance in this issue, but still comparing other methods they chose few true variables. Number of true variables selected in LARS is 39, in Lasso 31 and in GA is 30. As more the number of true variables in subset selected, as more accurate the prediction error and finally the better the carbonate content prediction. Thus as the number of features selected by LARS is quite good and more than others, then LARS shows a good performance in extracting true features comparing others. Also Lasso as in innovative feature selection method, does not have a significant difference with LARS and as it's the purpose of this study, these two methods are efficient feature selection methods for high dimensional data such as this case study.

Case II:

As it is shown in the Table 4.1., in Case II, RMSEP got its lowest value in LARS method (5.18), also for Lasso and GA, RMSEP is 5.78 and 6.34; which does not have significant difference but we can see in LARS the lowest RMSEP same as Case I, and so we can result in LARS is the best feature selection method among new methods for Case II. Looking at the classical methods, RMSEPs are 9.18 for stepwise selection and 9.13 for forward selection which has a significant difference with LARS, Lasso and GA. Although results show a little better performance by forward selection comparing stepwise selection, but we can conclude that LARS is the best method regarding RMSEP.

Regarding MSE and PE, also as explained above by looking at Table 4.1., values for LARS are 7.07 and 29.74; comparing to other methods even GA and Lasso which got higher MSE and PE, so considering this criteria also LARS is the best method among our feature selection methods here.

The important element here is the number of true features selected by each method. For classical methods we see a few true features selected by the method; such that in forward selection only 3 features is among selected subset and in stepwise selection only 5. Even if stepwise selected shows a better performance in this issue, but still comparing other methods they chose few true variables. Number of true variables selected in LARS is 16, in Lasso 12 and in GA is 15. As more the number of true variables in subset selected, as more accurate the prediction error and finally the better the carbonate content prediction. Thus as the number of features selected by LARS is quite good and more than others, then LARS shows a good performance in extracting true features comparing others. Also Lasso as in innovative feature selection method, does not have a significant difference with LARS and as it's the purpose of this study, these two methods are efficient feature selection methods for high dimensional data such as this case study.

As a conclusion for Table 4.1., as explained above for Case I and Case II, It is obviously shown that performance of LARS and Lasso are better than that of GA and LARS is the best among them; so the most efficient feature selection algorithm in this case is LARS.

To show the robustness of our model, we used Case II and it is obviously confirmed from the results that by adding a not relevant dataset to our main dataset, the efficiency of the model shouldn't change significantly, even though less efficient results were obtained.

So we can see in Table 4.1. Case II, that results had a small change comparing Case I in almost all criteria as well as number of selected variables. This shows even the robustness of our model represents the efficiency in modeling and feature selection by using appropriate methods such as LARS and Lasso.

As it was explained in Booststrap Section, in case that there are few samples comparing number of variables, we have to use some resampling methods in order to overcome this problem. It's a method which provides a confidence interval so the problem of few samples for training and testing set can be considered. So instead of one value for errors, a confidence interval for errors can be presented for more considerable result.

Results of 95% confidence interval obtained by bootstrapping in order to check the accuracy of the methods concluded in Tables 4.2.1~4.2.6.

Table 4.2. 1. Confidence interval of RMSEP using bootstrapping (Case I)

	FT-IR					
	LARS		**Lasso**		**GA**	
	bias	std. error	bias	std. error	bias	std. error
Bootstrap statistics	-0.65	0.08	-0.68	0.10	-0.73	0.17
Confidence interval	(3.77 , 5.48)		(3.08 , 4.42)		(4.44 , 9.12)	

Table 4.2.2. Confidence interval of MSE using bootstrapping (Case I)

	FT-IR					
	LARS		**Lasso**		**GA**	
	bias	std. error	bias	std. error	bias	std. error
Bootstrap statistics	-0.63	0.08	-0.68	0.09	-0.70	0.15
Confidence interval	(3.62 , 5.46)		(3.09 , 4.40)		(4.47 , 8.94)	

Table 4.2.3. Confidence interval of PE using bootstrapping (Case I)

	FT-IR					
	LARS		**Lasso**		**GA**	
	bias	std. error	bias	std. error	bias	std. error
Bootstrap statistics	-0.67	0.09	-0.70	0.11	-0.74	0.19
Confidence interval	(5.00 , 5.49)		(3.10 , 4.49)		(4.54 , 9.17)	

Table 4.2.4. Confidence interval of RMSEP using bootstrapping (Case II)

	FT-IR+XRD					
	LARS		**Lasso**		**GA**	
	bias	std. error	bias	std. error	bias	std. error
Bootstrap statistics	-1.13	2.45	-2.08	2.31	-3.07	4.10
Confidence interval	(4.82 , 5.88)		(4.08 , 6.42)		(2.44 , 9.12)	

Table 4.2.5. Confidence interval of MSE using bootstrapping (Case II)

	FT-IR+XRD					
	LARS		**Lasso**		**GA**	
	bias	std. error	bias	std. error	bias	std. error
Bootstrap statistics	-1.12	2.41	-2.10	2.31	-3.13	4.11
Confidence interval	(4.82 , 5.90)		(4.09 , 6.44)		(2.54 , 9.22)	

Table 4.2.6. Confidence interval of PE using bootstrapping (Case II)

	FT-IR+XRD					
	LARS		**Lasso**		**GA**	
	bias	std. error	bias	std. error	bias	std. error
Bootstrap statistics	-1.19	2.49	-2.11	2.37	-3.17	4.20
Confidence interval	(4.90 , 5.89)		(4.13 , 6.48)		(2.54 , 9.20)	

Results of bootstrapping and confidence intervals for different criteria are included in Table 4.2.1. ~ Table 4.2.6. Bootstrapping method was performed for the data in Case I and Case II for RMSEP, MSE and PE. Table 4.2.1. represents Confidence interval of RMSEP using bootstrapping (Case I). Bootstrap statistics for Case I in LARS is -0.65 and 0.08, and confidence interval is (5.03, 5.48). The statistics here are same as prediction error and the lower value for these statistics the better result in accuracy of model. Confidence interval for Lasso in this case is (3.08, 4.42) and for GA (4.44, 9.12). Comparing these intervals, we can see the one for LARS is less wide in comparison with Lasso and GA; this shows that our prediction error is located in a narrow interval and so the accuracy is higher than the ones in wider intervals.

Results in Table 4.2.1 confirm the conclusion obtained from Table 4.1. Bias parameter and std. error for LARS algorithm is the least among other methods and the confidence interval obtained has the smallest range, indicating more precise in detecting important variables among other methods. Also results of Lasso shows better performance comparing GA; which shows the efficiency and accuracy of LARS and Lasso in finding the most affective variables among too many variables. All the other tables above have same analysis and shows LARS as the best feature selection method.

4.2 Results for Case Study II: Gene Expression Data

Case study II which includes gene expression data representing a classification problem, such that in target variable, we have two classes of “0” and “1”; representing “without cancer”

and “with cancer”. So the aim is to classify new data based on the data in hand into these two groups and then find out the cancer cases among the sample size. As more the true classification happens, the more accurate our prediction modeling will be. So different feature selection methods were applied to gene expression data and subset selected by each method was considered to use for modeling with SVM. Since in this case, we have binary (0 & 1) feature as the target value, so we use ROC curve in order to find evaluation of each method and choose the best method related to the highest area under the curve in ROC curve for each method.

ROC curve for LARS, Lasso and GA are concluded in Figure 4.1.

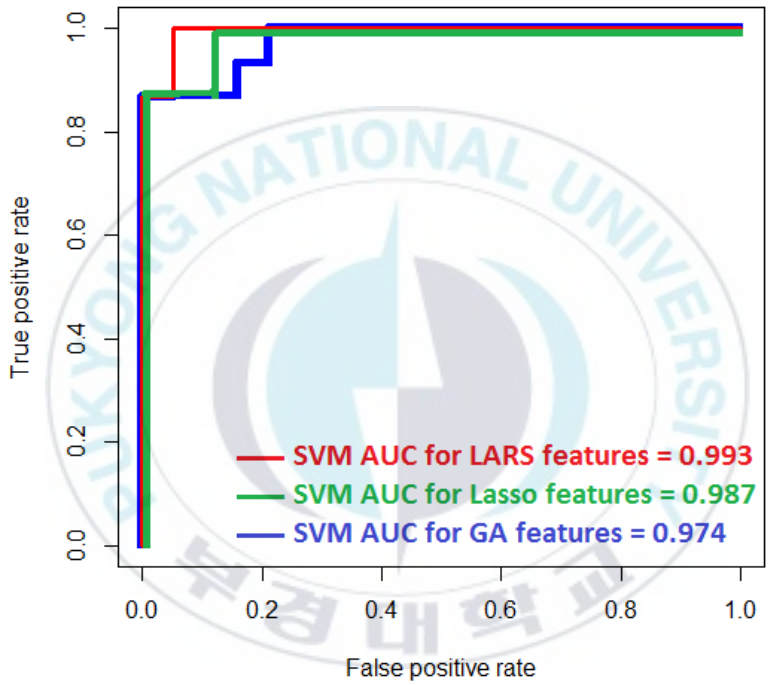

Figure 4.1 ROC Curve for case study II

In a ROC curve the true positive rate is plotted in function of the false positive rate. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups. The accuracy of the test depends on how well the test separates the group, so as we mentioned before, the more AUC is closer to one, the more accurate is the method. So as we see, AUC for LARS is 0.993 which shows a high accuracy in distinguishing cancer and no cancer group. Lasso and GA also shows good results, but among all, LARS could achieve the best accuracy in selecting most effective features.

In gene expression data, as the purpose of study is to classify “cancer” sample and “no cancer” one, so it’s very important to predict truly and group them in the correct one; so the accurate modeling and method will be the one which predicts all real cancer values into the cancer group and that’s also very challenging with high risk not to predict real “cancer” ones to “no cancer”; so an accurate method is the one which predict “cancer” as “cancer” as many as possible and predict “cancer” as “no cancer” as few as possible. Results of applying LARS, Lasso and Ga also classical methods (forward selection and stepwise selection) for this case study are included in Table 4.3.1~Table 4.3.5:

Table 4.3.1. Results of applying LARS to Case Study II

		PREDICTED	
		Cancer	No Cancer
ACTUAL	Cancer	126	22
	No Cancer	32	116

Table 4.3.2. Results of applying Lasso to Case Study II

		PREDICTED	
		Cancer	No Cancer
ACTUAL	Cancer	110	38
	No Cancer	39	109

Table 4.3.3. Results of applying GA to Case Study II

		PREDICTED	
		Cancer	No Cancer
ACTUAL	Cancer	106	42
	No Cancer	48	100

Table 4.3.4. Results of applying forward selection to Case Study II

		PREDICTED	
		Cancer	No Cancer
ACTUAL	Cancer	70	78
	No Cancer	58	90

Table 4.3.5. Results of applying stepwise selection to Case Study II

		PREDICTED	
		Cancer	No Cancer
ACTUAL	Cancer	68	80
	No Cancer	53	95

As we see in the tables 4.3.1 ~ 4.3.5, the minimum number of wrong prediction (detecting cancer when actually no cancer and detecting no cancer when actually cancer) is achieved in LARS algorithm comparing Lasso and GA. Classical methods also show weak results regarding accurate prediction. So in Table 4.3.4 and Table 4.3.5 there are big numbers of cases who actually have cancer, but modeling categorized them as no cancer. Moreover, using classical methods, many cases which actually have no cancer are categorized as cancer group. So comparing LARS, Lasso and GA, results show that classical methods are not efficient ones regarding thousands of numbers of gene expression in this case study.

CHAPTER 5. CONCLUSION

This study included feature selection methods and comparison of some classical and innovative feature selection methods and applying them to regression problem and classification problem in two case studies. In spectroscopy data as Case Study I, feature selection methods were applied to FT-IR and XRD data as Case I and Case II and then subset selected by each method was obtained and used for SVM modeling. Beside feature selection problem, the accuracy of methods in case of extracting true variables among all variables was also considered. The results showed the best performance and highest accuracy for LARS in Case Study I.

DNA microarray gene expression data for diagnosing prostate cancer used as Case Study II. As a classification problem, the main goal was to allocate true variables to the right groups; means real “cancer” group would be predicted as “cancer” as many as possible and at the same time “cancer” group was predicted as “no cancer” as few as possible because of high risk. In this case study also different feature selection methods were applied to the data to extract the most important variables on cancer prediction. The result was what expected, LARS was the best algorithm among Lasso and GA and significantly classical methods to predict cancer. The results anyway show in two general case studies that LARS and Lasso as special feature selection methods for high dimensional data, can be applied to massive data and extracting the most important variables using these methods.

REFRENCES

Almuallim, H., & Dietterich, T. G.. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 69(1) (1994), 279-305.

Armstrong, J. S. and F. Collopy, Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons, International Journal of Forecasting, 8 (1992) 69–80.

Balabina, R. M., S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Analytica Chimica Acta 692 (2011) 63–72.

Blum, A. L., P. Langley, Selection of Relevant Features and Examples in Machine Learning, Artificial Intelligence, 97 (1997) 245-271.

Blum, A. L., R. L. Rivest, Training a 3-node neural networks is NP-complete, Neural Networks, 5 (1992)117-127.

Breiman, L. (1993) Better subset selection using the nonnegative garotte. Technical Report. University of California, Berkeley.

Brereton, R. G. and A.K. Elbergali, J. Chemometrics, 8 (1994) 423-437.

Brown, P. J., J. Chemometrics, 6 (1992) 151-161.

Bruckman, V., S. Yan, E. Hochbichler, G. Glatzel, Carbon pools and temporal dynamics along a rotation period in Quercus dominated high forest and coppice with standards stands, Forest Ecology and Management, 262 (2011) 1853-1862.

Bruckman, V. and K. Wriessnig, Improved soil carbonate determination by FT-IR and X-ray analysis, Environmental Chemistry Letters, 11 (2013) 65-70.

Cheng-San, Y., L. Chuang, C.Ke and C. Yang, "A hybrid Feature Selection Method for Micro array Classification", International Journal of Computer Science, 35 (2008) 3-10.

Cools, N., B. de Vos, Sampling and Analysis of Soil. Manual Part X., in: Manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis

of the effects of air pollution on forests, UNECE, ICP Forests, Hamburg, (2010) 208-221.

Cristianini, N.; J. S. Taylor, An Introduction to support vector machines and other kernel-based learning methods; Cambridge University Press: New York. 2000.

Dai, J. J., L. Lieu, and D. Rocke, "Dimension Reduction for Classification with Gene Expression Micro array Data", Statistical Applications in Genetics and Molecular Biology: 5(2006) 152-167.

Dowell, F. E., Neural network parameter effects on object classification and wavelength selection, Optical Engineering, 33 (1994) 2728-2732.

Domingos, P., M. Pazzani, On the Optimality of the Simple Bayesian Classifier under Zero-One loss, Machine Learning, 29 (1997) 103-130.

Donoho, D. and Johnstone, I., Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81 (1994) 425-455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D., Wavelet shrinkage; J. R. Statist. Soc. B, 57 (1995) 301-337.

Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S., Maximum entropy and the nearly black object (with discussion). J. R. Statist. Soc. B, 54 (1992) 41-81.

Donoho D. L. and E. Elad, Maximal sparsity representation via l1 Minimization, Proc. Nat. Aca. Sci., 100 (2003) 2197-2202.

Donoho, D. L., High-dimensional data analysis: the curses and blessings of dimensionality. Aide-Memoire of the lecture in AMS conference "Math challenges of 21st Centrury (2000). Available at http://statweb.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf.

Efroymson,M. A. (1960) "Multiple regression analysis," Mathematical Methods for Digital Computers, Ralston A. and Wilf,H. S., (eds.), Wiley, New York.

Efron, B., Bootstrap methods: Another look at the jackknife. Ann. Statist. 7 (1979)1–26.

Efron, B., I. Johnstone, T. Hastie and R. Tibshirani, Least angle regression, Annals of Statistics, 32 (2004) 407-499.

Erdogmus, D. and J.C. Principe, An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems, IEEE Trans. Signal Process, 50 (2002) 1780–1786.

Foley, T. G., D.d. Richter, C.S. Galik, Extending rotation age for carbon sequestration: A cross-protocol comparision of North American forest offsets, Forest Ecology and Management, 259 (2009) 201-209.

Frank, I. and Friedman, J. (1993) A statistical view of some chemometrics regression tools (with discussion). Technometrics, 35, 109-148.

Goldberg, D. E., Genetic Algorithms in search, optimization and machine learning, Addison-Wesley, 1989.

Hocking, R. R., "The Analysis and Selection of Features in Linear Regression," Biometrics, 32 (1976) 82-97.

Hoerl A. E., and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1970) 55-67.

Holland, J. H., Adaption in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, MIT Press, 1992.

Horchner, U. and J.H. Kalivas, J. Chemomettics, 9 (1995) 283-308.

Horchner U., and J.H. Kalivas, Anal. Chim. Acta, 311 (1995) 1-13.

Huber, P. J. and Robust Statistics, Wiley, 1981 (republished in paperback, 2004), page 1.

James J. Chen and Chun-Houh Chen, "Micro array Gene Expression", Encyclopedia of Biopharmaceutical Statistics, 2nd Edition, Marcel Dekker, Inc., (2003) 599-613.

Jouan-Rimbaud, D., D.L. Massart, R. Leardi and O.E. de Noord, Anal. Chem., 67 (1995) 42954301.

Kalivas, J. H., N. Roberts and J.M. Sutter, Anal. Chem., 61 (1989) 2024-2030.

Kamogawa, M. Y., A.R.A. Nogueira, M. Miyazawa, J. Artigas, J. Alonso, Determination of soil calcareous efficiency using flow system with pervaporative separation, Anal Chim Acta, 438 (2001) 273-279.

Kohavi R., and H. John,Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273-324.

Lee, Y. and Chia-Huang Chao, "A Data Mining Application to Leukemia Micro array Gene Expression Data Analysis", International Conference on Informatics, Cybernetics and Systems (ICICS), Kaohsiung, Taiwan, 2003.

Lindgren, F., P Geladi, A. Berglund, S. Jostrom and S. Wold, J. Chemometrics, 9 (1995) 331-342.

Liu, C., Two Tales of Variable Selection for High Dimensional Data: Screening and Model Building, Dissertation, , The Ohio State University 2012.

Lucasius, C. B., M.L.M. Beckers and G. Kateman, Anal. Chim. Acta, 286 (1994) 135-153.

Lucasius, C. B., and G. Kateman, Chemometrics Intell. Lab. Syst., 19 (1993) 1-13.

Lucasius, C. B., and G. Kateman, Chemometrics Intell. Lab. Syst, 25 (1993) 99-145.

Messick, N. J., U.B. Horchner, J.H. Kalivas, Abstracts Of Papers Of the American Chemical Society, 208 (1994) 136-CHED.

Nguyen, D. V. and D. M. Rocke, Classification of Acute Leukemia based on DNA Micro array Gene Expressions using Partial Least Squares, Kluwer Academic, Dordrecht, 2001.

Nguyen, D. V. and D. M. Rocke, "Tumor Classification by Partial Least Squares Using Micro array Gene Expression Data", Bioinformatics, 18 (2002) 39-50.

Novakovic, J., The Impact of Feature Selection on the Accuracy of Naïve Bayes Classifier, 18th Telecommunications Forum TELFOR, November 23-25, Serbia, Belgrade, (2010) 1114-1116.

Olson, D.L. and Delen, D., Advanced Data Mining Techniques, Springer, 2008.

Schlichting, E., H.-P. Blume, K. Stahr, Bodenkundliches Praktikum - eine Einführung in pedologisches Arbeiten für Ö kologen, insbesondere Land- und Forstwirte, und für Geowissenschaftler, 2., neubearb. Aufl. ed., Blackwell Wiss.-Verl., Berlin ; Wien u.a., 1995.

Seeja, S., and J. Shweta, "Microarray Data Classification Using Support Vector Machine", International Journal of Biometrics and Bioinformatics (IJBB), 5 (2011) 10-15.

Shah, R., The Lasso: Variable selection, prediction and estimation, Statistics Reading Group 2012.

Sharma, A. and K. K. Paliwal, "Cancer classification by gradient LDA technique using micro array gene expression data", Data & Knowledge Engineering, 66 (2008) 338-347.

Sutter, J. M., J.H. Kalivas, Abstracts Of Papers Of the American Chemical Society, 203 (1992) 24-COMP.

Sutter, J., J.H. Kalivas, Abstracts Gf Papers Of the American Chemical Society, 195 (1988) 193CHED.

Swanson, D. A., J. Tayman and T. M. Bryan, MAPE-R: A Rescaled measure of accuracy for cross-selectional forecasts, Journal of Population Research, 28 (2011) 225-243.

Tibshirani, R., The Lasso method for variable selection in the COX model, Statistics in Medicine, 16 (1997) 385-395.

Tibshirani, R., Regression shrinkage and selection via the Lasso, Journal Royal Statistical Society B, 58 (1996) 267-288.

Walkeley, A., A critical examination of a rapid method for determining organic carbon in soils - effect of variations in digestion conditions and of inorganic soil constituents, Soil Science, 63 (1947) 251-264.

Wang, X., J. Wang, J. Zhang, Comparisons of Three Methods for Organic and Inorganic Carbon in Calcareous Soils of Northwestern China, PLoS ONE, 7 (2012) e44334.

Weisberg, S., Applied Linear Regression. Wiley, New York. MR591462, 1980.

Wilkinson, L., & Dallal, G.E., Tests of significance in forward selection regression with an F-to enter stopping rule. Technometrics, 23 (1981) 377–380.

Yang, Y. and N. P. Thorne, "Normalization for Two-color cDNA Microarray Data", Science and Statistics: A Festschrift for Terry Speed, Vol. 40 (2003) 403-418.

Zeng L., and J. Xie, Regularization and variable selection for data with interdependent structures, Biometrics, 20 (2013) 234-134.

Zhao P. and B. Yu, On Model Selection Consistency of Lasso, Journal of Machine Learning Research 7 (2006) 2541-2563.

Zheng, C., B. Li, L. Zhang and H. Q. Wang, "Locally Linear Discriminant Embedding for Tumor Classification", In Proceedings of ICIC, (2008)1093-1100.

Zou, H., The Adaptive Lasso and Its Oracle Properties, The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association, 101:476 (2006) 1418-1429.