IE Interfaces
Vol. 24, No. 2, pp. 151-155 , June 2011.

# Signomial Classification Method with $L_0$-regularization

Kyung sik Lee[†]

Department of Industrial and Management Engineering, Hankuk University of Foreign Studies

# $L_0$-정규화를 이용한 Signomial 분류 기법

이경식

한국외국어대학교 산업경영공학과

In this study, we propose a signomial classification method with $L_0$-regularization ($SC_0$), which seeks a sparse signomial function by solving a mixed-integer program to minimize the weighted sum of the $L_0$-norm of the coefficient vector of the resulting function and the $L_1$-norm of loss caused by the function. $SC_0$ gives an explicit description of the resulting function with a small number of terms in the original input space, which can be used for prediction purposes as well as interpretation purposes. We present a practical implementation of $SC_0$ based on the mixed-integer programming and the column generation procedure previously proposed for the signomial classification method with $L_1$-regularization. Computational study shows that $SC_0$ gives competitive performance compared to other widely used learning methods for classification.

***Keywords***: classification, signomial function, $L_0$-regularization, mixed-integer program, column generation

## 1. Introduction

The supervised learning is a machine learning technique to deduce a hidden function by investigating a number of data each of which can be represented as a pair of an input vector and the corresponding output value, which is to be used to predict the output values for new (unforeseen) input vectors (Lee *et al.*, 2010).

The classification problem is one of the most important topics in the development of the supervised learning methods, which has been studied by many researchers, so that a number of popular classification methods including the logistic regression (Hosmer and Lemeshow, 2000), CART (classification and regression tree) (Brieman *et al.*, 1984; Kim and Loh, 2001), and SVM (support vector machine) (Burges, 1998; Gunn, 1998; Vapnik, 1995) have been proposed and widely used.

Due to its simplicity and flexibility, SVM has been received much attention and has been applied to many practical applications. SVM is originally developed for two-class classification and separates observations of different classes by an affine function (a hyperplane) with a maximal margin (Vapnik, 1995). SVMs can explore the nonlinearity of data patterns in a relatively flexible manner by means of the so-called kernel trick, which make it possible to find a separating hyperplane in a higher dimensional feature space induced by kernel functions. For the details of SVMs, we refer the readers to Burges (1998) and Gunn (1998).

A classifier produced by SVMs is in the form of a kernel expansion, which is a sum of a linear combination of (pre-specified) kernel functions and a constant term. Even though there have been lots of

This research was supported by Hankuk University of Foreign Studies Research Fund.

[†] Corresponding author : Kyungsik Lee, San 89, Wangsan-ri, Mohyun-myun, Yongin-shi, Kyonggi-do, 449-791 Korea,
Fax : +82-30-330-4092, E-mail : globaloptima@hufs.ac.kr

Received 06 February 2011, Revised 21 February 2011, Accepted 21 February 2011.

effort on improving the generalization performance and the calculation time involved in the prediction by developing noble and computationally efficient SVMs to get a sparser kernel expansion, SVM has some drawbacks as noted by Lee *et al*. (2010) and Jeong *et al*. (2010) : (i) Even though exploring high dimensional feature space is possible by the kernel trick, the resulting function is restricted to some surface in the high dimensional space because the weights between higher terms are fixed by the kernel parameters; (ii) it is not easy to get an explicit function description in the original input space if we use a nonlinear kernel. Thus we can use the resulting function (classifier) for prediction purposes but less easily for interpretation purposes; (iii) The generalization performance of the resulting function may heavily depend on the choice of kernel function and its parameters.

To handle those drawbacks of SVMs mentioned above, Lee *et al*. (2010) proposed the sparse signomial classification and regression (SSCR) model. SSCR seeks a sparse signomial function by solving a linear program to minimize the weighted sum of the $L_1$-norm of the coefficient vector of the function and the $L_1$-norm of violation (or loss) caused by the function. SSCR gives an explicit description of the resulting function in the original input space while exploring the nonlinearity of the given data without using nonlinear kernels, and shows competitive and even better prediction performance compared to other widely used classification methods including SVMs (Burges, 1998; Gunn, 1998).

The number of terms of the resulting classifier is closely related to the prediction performance of the classifier, and to have a smaller number of terms (a sparser classifier) may be desirable for the better prediction performance (Vapnik, 1995). To get a sparse classifier, SSCR used $L_1$-regularization as mentioned above, which is in fact a relaxation of $L_0$-regularization.

In this study, we propose a signomial classification method with $L_0$-regularization (SC$_0$) based on SSCR (Lee *et al*., 2010), which seeks a sparser signomial function than those obtained by SSCR by solving a mixed-integer program to minimize the weighted sum of the $L_0$-norm of the coefficient vector of the resulting function and the $L_1$-norm of loss caused by the function. We present a practical implementation of SC$_0$ based on the mixed-integer programming and the column generation procedure previously proposed for SSCR.

In the next section, we present the detailed description of SC$_0$ with a practical implementation of it. The performance of SC$_0$ is evaluated in a computational study given in section 3. Finally, concluding remarks are given in section 4.

## 2. An implementation of SC$_0$

Let $x = (x_1, \cdots, x_n)$ be a positive real vector, and let $d = (d_1, \cdots, d_n)$ be a real vector. Define a real-valued function of $x$ for a given $d$, $g_d(x) := \prod_{i \in N} x_i^{d_i}$, where $N = \{1, 2, \cdots, n\}$. Note that the components of $d$ can be negative and/or fractional. Then, a signomial function of $x$ is defined as

$$f(x) = \sum_{d \in D} w_d g_d(x) + b \qquad (1)$$

where $b \in R$, $D$ is a finite subset of $R^n$ such that $0 \not\in D$, and $w_d \in R$ for all $d \in D$. For a given signomial function of the form (1), $L_1$-norm ($L_0$-norm) of the coefficient vector $w \in R^{|D|}$ is denoted by $\| w \|_1 (\| w \|_0)$. Note that $\| w \|_0$ is defined as the number of non-zero elements of the coefficient vector.

For given two subsets in $R^n_{++}$ (the set of positive real vectors of $n$ dimension), $\{x^j\}_{j \in M_0}$ and $\{y^j\}_{j \in M_1}$, the signomial classifiers obtained by SSCR for the two sets are defined as the solution of the following optimization problem (Lee *et al*., 2010).

$$\begin{aligned} \text{minimize } & \| w \|_1 + C(\| u \|_1 + \| v \|_1) \qquad (2) \\ \text{subject to } & \sum_{d \in D} w_d g_d(x^j) + b \geq 1 - u_j,\ j \in M_0, \\ & \sum_{d \in D} w_d g_d(y^j) + b \leq -1 + v_j,\ j \in M_1, \\ & w \in R^{|D|},\ b \in R,\ u \in R_+^{|M_0|},\ v \in R_+^{|M_1|}. \end{aligned}$$

In the above problem (2), the set $D$ is defined by four parameters, $d_{max}$, $d_{min}$, $T$, and $S$, as follows:

$$D = \Big\{ d \in R^n \,|\, d_{min} \leq d_i \leq d_{max},\ i \in N, \sum_{i \in N} |d_i| \leq S,\ Td \in Z^n \Big\} \qquad (3)$$

The problem (2) can be readily reformulated as a linear program. However, depending on the four parameters of the set $D$, the number of elements of $D$ can be huge, which means there may be huge number of variables in the reformulated linear program. To overcome this issue Lee *et al*. (2010)

proposed an algorithm to solve the linear program based on the column generation technique (Chvatal, 1983) to handle this issue. For the details of SSCR, we refer the readers to Lee *et al.* (2010). Now, we present the detailed description of SC$_0$.

### 2.1 A mixed-integer program for SC$_0$

The signomial classifiers with $L_0$-regularization for the given two sets $\{x^j\}_{j\in M_0}$ and $\{y^j\}_{j\in M_1}$ in $R^n_{++}$, are defined as the solution of the following optimization problem.

$$\begin{aligned} \text{minimize } & \| w \|_0 + C( \| u \|_1 + \| v \|_1) \qquad (4) \\ \text{subject to } & \sum_{d\in D} w_d g_d(x^j) + b \ge 1 - u_j,\ j\in M_0, \\ & \sum_{d\in D} w_d g_d(y^j) + b \le -1 + v_j,\ j\in M_1, \\ & w\in R^{|D|},\ b\in R,\ u\in R_+^{|M_0|},\ v\in R_+^{|M_1|}, \end{aligned}$$

where the set $D$ is defined as (3).

Unlike problem (2) for SSCR, the above problem (4) can not be transformed into a linear program.

For a sufficiently large positive real number $U$, the above problem (4) can be transformed into the following mixed-integer program by introducing a binary variable $z_d$ for each $d \in D$, which we call MP :

$$\begin{aligned} \text{minimize } & \sum_{d\in D} z_d + C( \sum_{j\in M_0} u_j + \sum_{j\in M_1} v_j) \qquad (5) \\ \text{subject to } & \sum_{d\in D} w_d g_d(x^j) + b \ge 1 - u_j, j\in M_0, \\ & \sum_{d\in D} w_d g_d(y^j) + b \le -1 + v_j, j\in M_1, \\ & -Uz_d \le w_d \le Uz_d, d \in D, \\ & w\in R^{|D|}, b\in R, u\in R_+^{|M_0|}, v\in R_+^{|M_1|}, \\ & z \in B^{|D|} \end{aligned}$$

### 2.2 An Implementation of SC$_0$

Even though the exponents set $D$ is finite, it can be exponentially large, which makes it practically intractable to solve MP after enumerating all the elements of $D$. In this section, we propose a practical implementation of SC$_0$ based on the generation method for the problem (2) devised in Lee *et al.* (2010).

By using the fact that the linear program corresponding to the problem (1), LP, is a relaxation of MP, we first solve LP rather than the LP-relaxation of MP. To solve LP, we used the algorithm devised by Lee *et al.* (2010). Then, we solve MP which is formulated by using the generated columns (a subset of $D$) in the process of solving LP rather than the entire set $D$. To formulate MP, We also specify the value of $U$ to the maximum of absolute values of $w_{d,} d \in D$. This approach, of course, can not guarantee an optimal solution to MP. However, as proved by Lee *et al.* (2010), the column generation problem for LP is NP-hard so that it may not be practical to try to solve MP to optimality. In the following, we give the details of our implementation of SC$_0$.

```
Procedure SC0($C_1, C_2$) :
Given  : Positive real numbers $C_1, C_2$;
Step1  : Solve LP with $C = C_1$;
Step2  : Let $\tilde{D}$ be the set of generated exponents;
         Let $\tilde{w}$ be the solution of LP;
Step3  : Solve MP with $D := \tilde{D}$, $C = C_2$,
         and $U = \max_{d\in\tilde{D}}\{|\tilde{w_d}|\}$;
```

## 3. Experimental Results

In this section, we report on the performance of SC0 on four widely circulated real data sets, 'Breast Cancer,' 'BUPA Liver,' 'Diabetes,' and 'Heart Disease,' from the UCI Machine Learning Repository (Murphy and Aha, 1992). <Table 1> shows the characteristics of those data sets including the name of each data set, the number of observations and variables of each data set, and the associated task with each data set.

**Table 1.** The characteristics of data sets

| Data Set | #Observations | #Variables |
|---|---|---|
| Breast Cancer | 683 | 9 |
| BUPA Liver | 345 | 6 |
| Diabetes | 768 | 8 |
| Heart Disease | 297 | 13 |

The 'Heart Disease' data set originally consists of 303 observations but 6 of them have some missing values. We discarded those observations and used the remaining 297 observations.

For the test of SC$_0$, the procedure SC$_0$($C_1$, $C_2$) was implemented by using Xpress Mosel language and associated optimization library functions including the branch-and-bound procedure for solving mixed-integer programs (Xpress, 2010). The experiment was executed on a PC(2.5GHz CPU, 3GB RAM), and we set the time limit of the branch-and-bound procedure to 100 seconds. In the experiment, we specified the set $D$ defined as (3) with its parameters $d_{min}=-1$, $d_{max}=1$, $S=1$, and $T=100$ for all the data sets. The larger value of $C_1$ for LP in Step1 results in generating more signomial terms, which enables MP in Step 3 to choose less number of terms. In our computational test, the parameter $C_1$ for LP in Step1 was set to 100 for all the data sets. The other parameter $C_2$ for MP in Step 3 was varied such that $C_2=10^{-3}$, $10^{-2}$, $10^{-1}$, 1, $10^1$, $10^2$ and $10^3$. For each value of $C_2$, we performed five independent runs. We summarize the average training, validation and test accuracies in <Table 2> for the best value of $C_2$ with respect to the average validation accuracy.

**Table 2**. The performance of SC$_0$

| Data Set | Train | Validation | Test |
|---|---|---|---|
| Breast Cancer | 0.967 | 0.959 | 0.964 |
| BUPA Liver | 0.720 | 0.718 | 0.692 |
| Diabetes | 0.757 | 0.754 | 0.749 |
| Heart Disease | 0.808 | 0.784 | 0.810 |

The following <Table 3> shows the sparsity of the resulting classifiers obtained from SC$_0$ for the best value of $C_2$ for each data set. In the table, #GenTerms denotes the average number of generated terms (the number of elements in $\widetilde{D}$) in Step1, and #SelTerms denotes the average number of terms in the resulting classifiers.

**Table 3**. The sparsity of classifiers (SC$_0$)

| Data Set | #GenTerms | #SelTerms |
|---|---|---|
| Breast Cancer | 91.8 | 1.0 |
| BUPA Liver | 103 | 2.6 |
| Diabetes | 240.8 | 3.0 |
| Heart Disease | 105.2 | 2.0 |

Now, we present the performance of previously proposed classification methods on the same data sets including SSCR (Lee *et al*., 2010), SVM (Vapnik, 1995), logistic regression (Hosmer and Lemeshow, 2000), and CART (Brieman *et al*., 1984; Kim and Loh, 2001), and compare their performance to that of SC$_0$ in terms of test accuracy measured by the portion of correctly classified observations. These tests were performed by one of the authors of SSCR (Lee *et al*., 2010) to compare SSCR with SVM, logistic regression, and CART.

SSCR was implemented by using Xpress Mosel language and associated optimization library functions (Xpress, 2010). SVM was tested by using the LIBSVM package (Chang and Lin, 2001). We used the MATLAB toolbox (Matlab statistic toolbox, 2008) for testing CART, logistic regression.

For SVM, linear, polynomial(with degree 2 and 3) and RBF kernels were all tested with varying model parameter $P$ (similar parameter to $C$ in MP) such that $P=10^{-3}$, $10^{-2}$, $10^{-1}$, 1, $10^1$, $10^2$ and $10^3$. For RBF kernels, the kernel parameter $\sigma$ was varied such that $\sigma=10^{-3}$, $10^{-2}$, $10^{-1}$, 1, $10^1$, $10^2$ and $10^3$. For each combination of a kernel, $P$, and $\sigma$ among 70 combinations in all, five independent runs were also performed. The average test accuracies for the best combination in terms of the average validation accuracy are summarized in <Table 4>.

For CART, trees are pruned based on an optimal pruning scheme that first prunes branches giving less improvement in error cost. By varying the pruning level, five independent runs were also performed. The average test accuracies for the best pruning level in terms of the average validation accuracy are summarized in <Table 4>.

**Table 4**. Test accuracies of SVM and CART

| Data Set | SVM | CART |
|---|---|---|
| Breast Cancer | 0.960 | 0.945 |
| BUPA Liver | 0.701 | 0.614 |
| Diabetes | 0.773 | 0.706 |
| Heart Disease | 0.831 | 0.749 |

From <Table 2> and <Table 4>, we observe that the performance of SC$_0$ in terms of test accuracy is competitive to those of SVM and CART. In particlular, SC$_0$ shows consistently better performance than CART for all the data sets.

For the logistic regression, no special parameters are needed to be specified. As in the other cases, five independent runs were also performed, and the average test accuracies are summarized in <Table 5>.

**Table 5**. Test accuracies of Logistic Regression and SSCR

| Data Set | Logistic Regression | SSCR |
|---|---|---|
| Breast Cancer | 0.961 | 0.964 |
| BUPA Liver | 0.672 | 0.754 |
| Diabetes | 0.767 | 0.761 |
| Heart Disease | 0.824 | 0.834 |

From <Table 2> and <Table 5>, we also observe that the performance of SC$_0$ in terms of test accuracy is competitive to that of logistic regression. However, SSCR shows consistently better performance than $SC_0$ for all the data sets.

The following <Table 6> shows the sparsity of the resulting classifiers obtained from SSCR for the best value of $C$ for each data set. In the table, #GenTerms denotes the average number of generated terms, and #SelTerms denotes the average number of terms in the resulting classifiers.

**Table 6**. The sparsity of classifiers (SSCR)

| Data Set | #GenTerms | #SelTerms |
|---|---|---|
| Breast Cancer | 20.4 | 8.0 |
| BUPA Liver | 24.6 | 10.4 |
| Diabetes | 39.4 | 20.6 |
| Heart Disease | 54.6 | 13.6 |

Even though SSCR shows better test accuracies than $SC_0$, <Table 3> and <Table 6> show that classifiers obtained from $SC_0$ are much sparser than those obtained from SSCR.

## 4. Conclusion

A signomial classification method with $L_0$-regularization with a practical implementation is developed, and is compared to the existing classification methods. Experimental results show that the proposed $SC_0$ shows competitive prediction performance to the existing methods with smaller number of terms in the resulting classifiers. The average number of terms in the resulting classifiers is only 1~3. This is an encouraging result since the proposed approach gives an explicit function description in original input space, which can facilitate easier interpretation.

Topics for the future research may include devising more sophisticated implementation of $SC_0$ along with an extensive computational experiments to further enhance the proposed approach and to extend it to the case of the variable selection.

## Acknowledgment

The authors would like to thank Dr. Norman Kim for providing us with the computational results of SVM, logistic regression, CART, and SSCR.

## References

Brieman, L., Friedman, J., Olshen, R., and Stone, C. (1984), Classication and Regression Trees, Chapman and Hall.

Burges, C. J. C. (1998), A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2, 121-167.

Chang, C. C. and Lin, C. J. (2001), LIBSVM : a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chavatal, V. (1993), Linear Programming, W. H. Freeman and Company.

Gunn, S. R. (1998), Support Vector Machines for Classification and Regression, Technical Report of School of Electronics and Computer Science, University of Southampton.

Hosmer, T. and Lemeshow, S. (2000), Applied logistic regression, John Wiley and Sons.

Jeong, Y., Lee, C., Kim, N., and Lee, K. (2010), Remote health monitoring Parkinson's disease severity using signomial regression model, *IE Interfaces,* 23, 365-371.

Kim, H. and Loh, W. Y. (2001), Classification tree with unbiased multiway splits,*Journal of American Stattistical Association*, 96, 598-604.

Lee, K., Kim, N., and Jeong, M. (2010), Sparse Signomial Classification and Regression, RUTCOR Research Reports.

Matlab statistics toolbox (2008), http://www.mathworks.com.

Murphy, P. M. and Aha, D. W. (1992), UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.

Vapnik, V. N. (1995), The Nature of Statistical Learning Theory, Springer.

Xpress (2010), http://www.fico.com.

Kyungsik Lee
Professor, Department of Industrial and Management Engineering, Hankuk University of Foreign Studies
Ph.D and MS : Korea Advanced Institute of Science and Technology
BS : Seoul National University
Research Topics: Optimization theory and applications