# Key course selection in academic warning with sparse regression

Min Yin*, Xijiong Xie and Shiliang Sun

Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, P. R. China
xjxie11@gmail.com,slsun@cs.ecnu.edu.cn

**Abstract.** Many colleges and universities are paying more attention to academic warning which warns large numbers of students who have unsatisfactory academic performance. Academic warning becomes a new part in the teaching management constitution but lacks of unified and scientific standards under the establishment of this stipulation at present. This paper solves the current setting of academic warning through well-known methods lasso and $\ell_1$-norm support vector regression with $\epsilon$-insensitive loss function which can select key courses based on the failed credits in one semester. The experiments are made on our collected academic warning datasets which are incomplete data. We impute them with one nearest neighbor method. The experimental results show that sparse regression is effective for colleges and universities to remind the students of key courses.

**Key words:** Academic warning, Lasso, $\ell_1$-norm support vector regression, Sparse regression

## 1 Introduction

With the continuous improvement of the whole education system, a large number of colleges and universities adopt academic warning systems in student academic management. Academic warning has been one of computational education science and problems [1]. Academic warning can monitor and supervise the students' study and promote them to learn consciously. As we know, different courses have correspondingly different credits and the score of each course decides whether the student can obtain the relevant credit. Whether a student is warned can be based on the sum of the credits of all failed courses. In general, each college or university warns the student by setting a credit line reasonably. If the setting can be obtained through machine learning methods [2][3][4], it will improve the performance of academic warning. Simultaneously, key course selection is also important. Key course selection is easily accomplished by traditional statistical methods. However, some hidden information in data may be ignored. For example, the failed student numbers of some courses are few but their scores are low. Machine learning methods can take advantage of this information.

* The first author and the second author contributed equally to this work.

A vastly popular and successful approach in statistical modeling is to use regularization penalties in model fitting. The use of $\ell_1$ regularization for statistical inference has become very popular over the last two decades. Tibshirani [5] proposed the least absolute shrinkage and selection operator (lasso) technique which uses an $\ell_1$-penalized likelihood for linear regression with independent Gaussian noise, which involves minimizing the usual sum of squared error loss with $\ell_1$ regularization. The lasso has become the standard tool for sparse regression for which its $\ell_1$ penalty leads to sparse solutions. That is, there are few nonzero estimates. Sparse models are more interpretable and often preferred in the natural and social sciences. Much of the early effort has been dedicated to solving the optimization problem efficiently [6][7][8]. Support vector machine is the most popular algorithm for classification and regression. There is an important model called $\ell_1$-norm support vector regression (SVR) [9][10][11], which is used to identify the critical features for regression. It can deal with the case where a lot of noisy and redundant features are present. $\ell_1$-norm SVR can be regarded as a linear programming (LP) problem. In fact, the $\ell_1$-norm SVR with Gaussian loss function is equivalent to a lasso problem. In this paper, we select key courses through well-known methods lasso and $\ell_1$-norm support vector regression with $\epsilon$-insensitive loss function on our collected academic warning datasets. The experimental results show that sparse regression can provide a new universal method for colleges and universities to remind the students of key courses.

The structure of the paper is organized as follows. In Section 2, we introduce two models $\ell_1$-norm SVR and lasso. Then data collection is also depicted in this section. After reporting experimental results in Section 3, we give conclusions in Section 4.

## 2 Model and Data Collection

In this section, we introduce two models $\ell_1$-norm SVR and lasso. Then we introduce our collected academic warning datasets.

### 2.1 $\ell_1$-norm SVR

We give a brief outline of $\ell_1$-norm SVR. Suppose $f(\bar{x})$: $\mathbb{R}^d \to \mathbb{R}$ that transforms the input vector $\bar{x} \in \mathbb{R}^d$ to a real number $f(\bar{x})$. SVR aims to estimate $f(\bar{x})$ by observing $n$ training examples. Here we consider the linear case, $f(\bar{x}) = \bar{w}^\top \bar{x} + b$, where $\bar{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$. Define that $x = \begin{pmatrix} \bar{x} \\ 1 \end{pmatrix}$, $w = \begin{pmatrix} \bar{w} \\ b \end{pmatrix}$. Then above formulation can be written in the homogeneous form $f(x) = w^\top x$. The optimization objective of $\ell_1$-norm SVR is

$$\begin{aligned} \min \quad & \|w\|_1 \\ \text{s.t.} \quad & y_i = w^\top x_i, \ i = 1, 2, \ldots, n. \end{aligned} \tag{1}$$

In practical applications. the output $y_i$ may be corrupted by some noise. A number of loss functions $L(x, y, f(x))$ and a penalty term $e^\top q$ can be introduced

into (1) to deal with the noise. The formulation (1) can be updated to

$$\begin{aligned} \min \ & \|w\|_1 + Ce^\top q \\ \text{s.t.} \ & L(x_i, y_i, f(x_i)) \le q_i, \ i = 1, 2, \ldots, n, \ q_i \ge 0, \end{aligned} \tag{2}$$

where $C$ is non-negative constant controlling the tradeoff between the norm regularization and penalty. Different loss functions are suitable for different problems. The $\epsilon$-insensitive loss function

$$L(x, y, f(x)) = \max\{|y - f(x)| - \epsilon, 0\} \tag{3}$$

is one of the most commonly used loss functions, where $\epsilon$ is a parameter which needs to be set in advance. The optimization objective is specified as

$$\begin{aligned} \min \ & \|w\|_1 + Ce^\top q_i \\ \text{s.t.} \ & |y_i - w^\top x_i| \le q_i, \ i = 1, 2, \ldots, n, \ q_i \ge 0. \end{aligned} \tag{4}$$

This formulation can be solved as an LP problem. The Gaussian loss function

$$L(x, y, f(x)) = \frac{1}{2}(y - f(x))^2 \tag{5}$$

is appropriate to deal with the Gaussian-type noise. Then the optimization objective is specified as

$$\begin{aligned} \min \ & \|w\|_1 + Ce^\top q_i \\ \text{s.t.} \ & (y_i - w^\top x_i)^2 \le q_i, \ i = 1, 2, \ldots, n, \ q_i \ge 0. \end{aligned} \tag{6}$$

The optimization objective cannot be regard as an LP problem but there exist many efficient methods for solving it. One noteworthy fact is that with the Gaussian noise, the notion of support vector is then meaningless.

### 2.2 Lasso

Given an input $X \in \mathbb{R}^{n \times d}$, each row of $X$ represents an example, and an output $y \in R^n$, the lasso is least square regression with $\ell_1$-norm regularization, which attempts to solve the following optimization problem

$$\min \ \|w\|_1 + C\|y - Xw\|_2^2 \tag{7}$$

where $w$ is sparse, which means that most elements of $w$ are zero. In fact, the $\ell_1$-norm SVR with the Gaussian loss function is equivalent to a lasso problem.

### 2.3 Data Collection and Imputation

Academic warning datasets are collected by ourselves from a certain university. They contain 28 datasets which represent the scores of students of class 1 and class 2 of grade 2010 and grade 2011 in the seven semesters (one row represents

one student and one column represents one course). The number of students are 47, 51, 23 and 52 in four classes. The label represents the failed credits of one student in one semester. However, there are a number of miss values in these datasets. We use one nearest neighbor (1NN) to impute these missing values. The 1NN method replaces the missing value in the data matrix with the corresponding value from the nearest row. That is to say, it can identify the most similar score to the current one with a missing value, and use the score as a guess for the missing one. Dataset 201011 represents the scores of students of class 1 of grade 2010 in the first semester. Dataset 201012 represents the scores of students of class 1 of grade 2010 in the second semester. The names of other datasets are analogous to the above ones.

## 3 Experiments and Results

We use lasso and $\ell_1$-norm SVR with the $\epsilon$-insensitive loss function to select features which have the most information. The nonzero values of $w$ represent key courses and the absolute value of element in $w$ represents the importance of course. If the absolute value of element in $w$ is bigger, the corresponding course is more important. In the lasso method, we use the lasso function in matlab and select number of non-zero coefficients from small to large in the range of integers until $w$ emerges. In the $\ell_1$-norm SVR, we use the code in reference [11] and select optimal parameter by grid search strategy until the sparseness of $w$ is smallest. The experimental results are list in Fig.1-Fig.28 (the horizontal coordinate represents courses and the vertical coordinate represents vaules of elements in $w$). The results of the lasso are represented by the white bar and the results of $\ell_1$-norm SVR are represented by the green bar. For simplicity, we analyze the results of the first four datasets in detail.

The dataset 201011 contains C programming, sports, military theory, psychology, ideological and moral cultivation and legal basis, linear algebra, English, introduction to computer science, experiments of introduction to computer and advanced mathematic. In this semester, we use lasso to select key courses. The results are advanced mathematic, English and C programming in descending order. Then we use $\ell_1$-norm support vector regression to select key courses. The results are advanced mathematic, introduction to computer science, C programming, ideological and moral cultivation and legal basis, military theory and experiments of introduction to computer. The true importance descending order of courses is C programming, experiments of introduction to computer, advanced mathematic, psychology, linear algebra according to the failed student number of courses. We can find that the lasso method finds out key courses advanced mathematic and C programming while $\ell_1$-norm support vector regression finds out key courses advanced mathematic, C programming and experiments of introduction to computer.

The dataset 201012 contains outline of modern Chinese history, sports, public elective course, college English, teaching media and technology, education, module one, life sciences, computer programming practice, object-oriented pro-

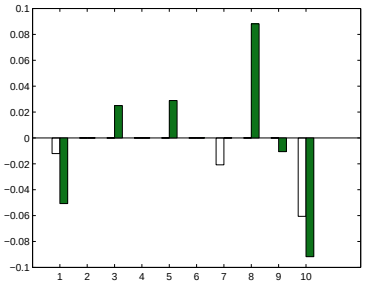


**Fig. 1.** Results of dataset 201011

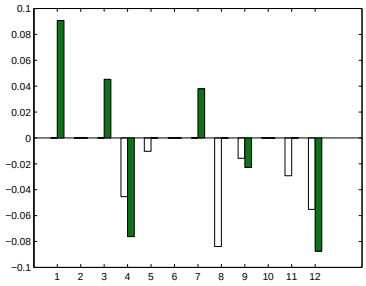


**Fig. 2.** Results of dataset 201012

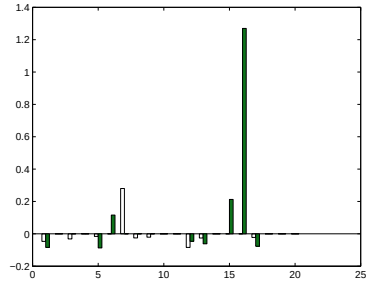


**Fig. 3.** Results of dataset 201013

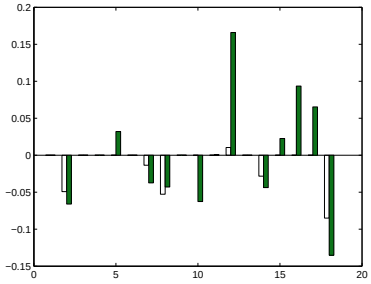


**Fig. 4.** Results of dataset 201014

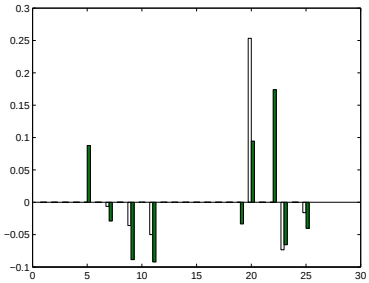


**Fig. 5.** Results of dataset 201015

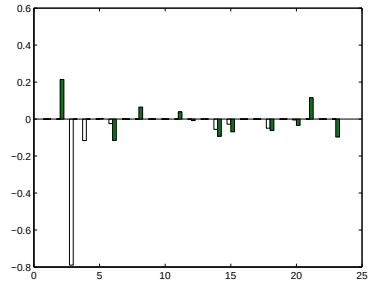


**Fig. 6.** Results of dataset 201016

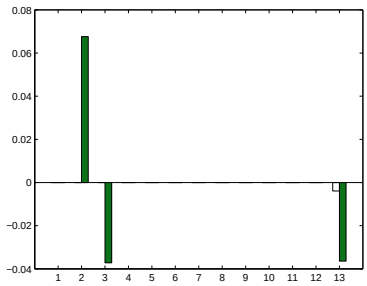


**Fig. 7.** Results of dataset 201017

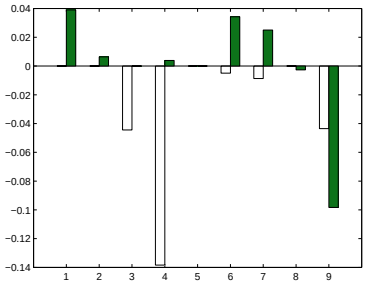

**Fig. 8.** Results of dataset 201021

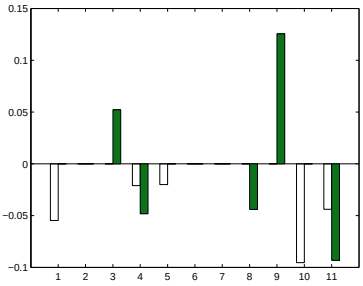

**Fig. 9.** Results of dataset 201022

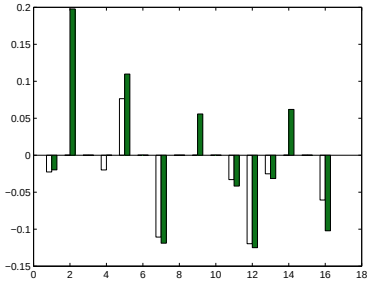

**Fig. 10.** Results of dataset 201023

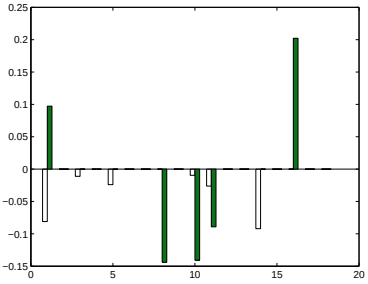

**Fig. 11.** Results of dataset 201024

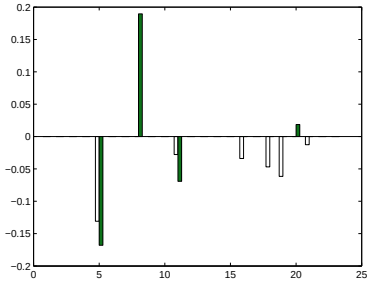

**Fig. 12.** Results of dataset 201025

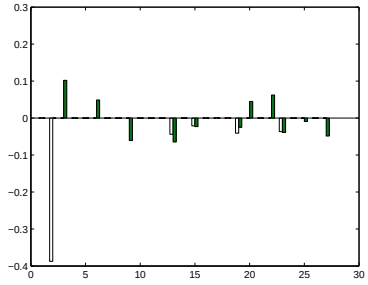

**Fig. 13.** Results of dataset 201026

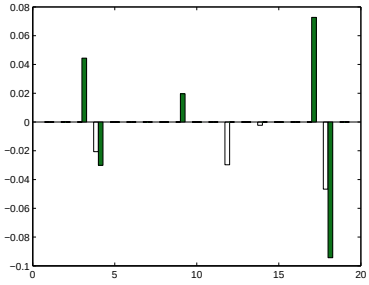

**Fig. 14.** Results of dataset 201027

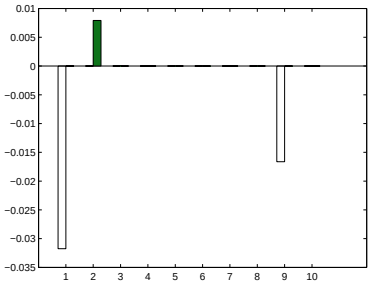

**Fig. 15.** Results of dataset 201111

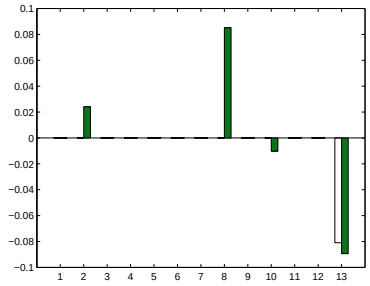

**Fig. 16.** Results of dataset 201112

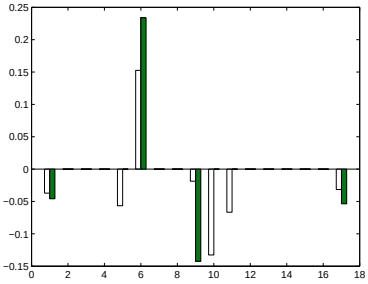

**Fig. 17.** Results of dataset 201113

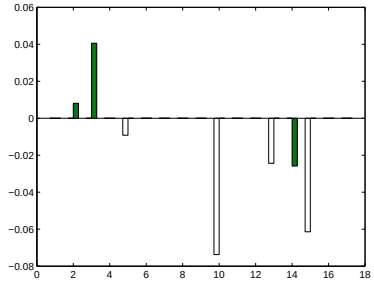

**Fig. 18.** Results of dataset 201114

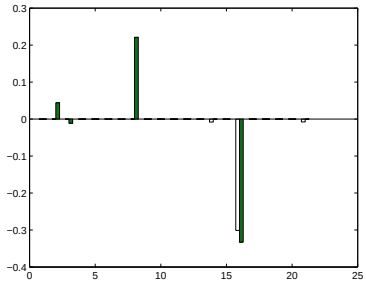

**Fig. 19.** Results of dataset 201115

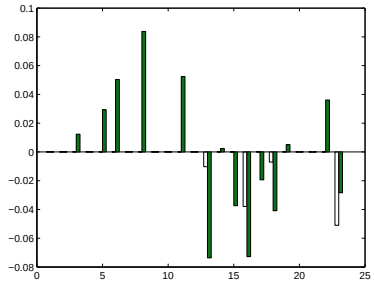

**Fig. 20.** Results of dataset 201116

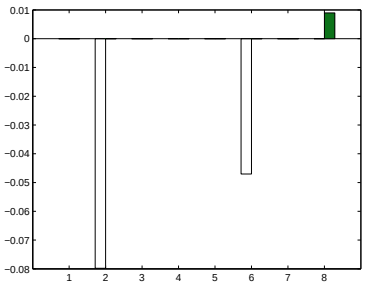

**Fig. 21.** Results of dataset 201117

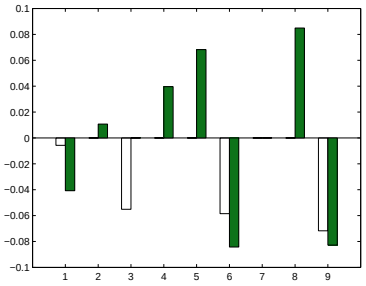


**Fig. 22.** Results of dataset 201121

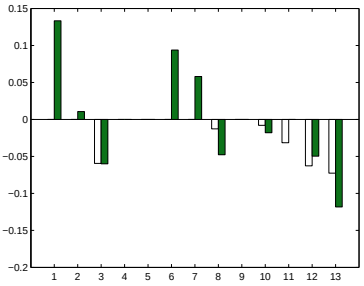


**Fig. 23.** Results of dataset 201122

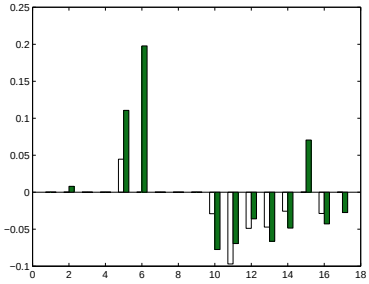


**Fig. 24.** Results of dataset 201123

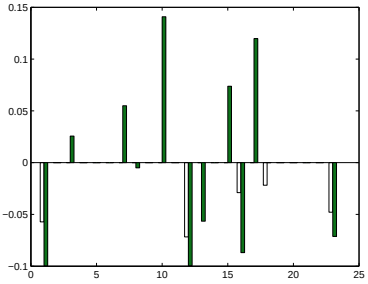


**Fig. 25.** Results of dataset 201124

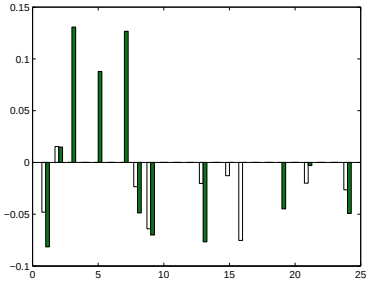


**Fig. 26.** Results of dataset 201125

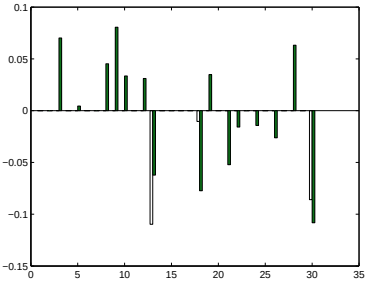


**Fig. 27.** Results of dataset 201126

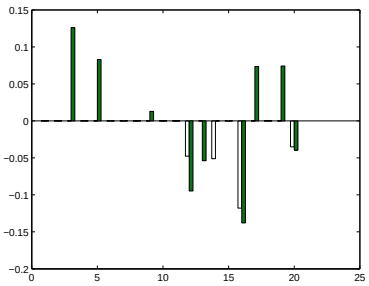


**Fig. 28.** Results of dataset 201127

gramming (based on C++), object-oriented programming (based on java) and advanced mathematic. We use lasso to select key courses. The results are life sciences, advanced mathematic, college English, object-oriented programming (based on java), computer programming practice and teaching media and technology in descending order. Then we use $\ell_1$-norm support vector regression to select key courses. The results are outline of modern Chinese history, advanced mathematic, college English, public elective course, module one, computer programming practice in descending order. The true importance descending order of courses is advanced mathematic, computer programming practice, object-oriented programming (based on java), object-oriented programming (based on C++) and college English according to the failed student number of courses. We can find that the lasso method finds out key courses advanced mathematic, object-oriented programming (based on java), college English while $\ell_1$-norm support vector regression finds out key courses advanced mathematic, computer programming practice, college English.

The dataset 201013 contains C programming, sports, information technology curriculum and teaching theory, confucianism and modern society, public elective course, the outline of history, classical Chinese culture, college English level 4, college Chinese, psychology, teachers spoken language, digital logic and experiment, data structure, module one, an introduction to Mao Zedong thought and the theory system of socialism with Chinese characteristics, material science, discrete mathematics, linear algebra, basic principle of Marxism and advanced mathematic. We use lasso to select key courses. The results are classical Chinese culture, digital logic and experiment, C programming, information technology curriculum and teaching theory, public elective course, data structure, college English level 4 and discrete mathematics in descending order. Then we use $\ell_1$-norm support vector regression to select key course. The results are material science, an introduction to Mao Zedong thought and the theory system of socialism with Chinese characteristics, the outline of history, public elective course, C programming, discrete mathematics, data structure, digital logic and experiment in descending order. The true importance descending order of courses is C programming, discrete mathematics, public elective course, digital logic and experiment, data structure, information technology curriculum and teaching theory, college English level 4, module one and material science according to the failed student number of courses. We can find that the lasso method finds out key courses C programming, discrete mathematics, public elective course, digital logic and experiment, data structure, information technology curriculum and teaching theory, college English level 4 while $\ell_1$-norm support vector regression finds out key courses C programming, discrete mathematics, public elective course, digital logic and experiment, data structure, material science.

The dataset 201014 contains web application technology, windows application design, sports, introduction to information system security, public elective course, college English, college English (advanced), operating system, educational probation, probability theory and mathematical statistics, module one, material science, algorithm analysis and design, computer composition and struc-

ture, computer aided education, object-oriented programming (based on C++), object-oriented programming (based on java), and advanced mathematic. We use lasso to select key courses. The results are advanced mathematic, operating system, windows application design, computer composition and structure, college English (advanced) and material science in descending order. We use $\ell_1$-norm support vector regression to select key courses. The results are material science, advanced mathematic, object-oriented programming (based on C++), windows application design, object-oriented programming (based on java), probability theory and mathematical statistics, computer composition and structure, operating system, college English (advanced), public elective course, computer aided education and module one in descending order. The true importance descending order of courses is windows application design, computer composition and structure, operating system, probability theory and mathematical statistics, advanced mathematic, public elective course, college English (advanced), algorithm analysis and design, sports and object-oriented programming (based on java) according to the failed student number of courses. We can find that the lasso method finds out key courses windows application design, computer composition and structure, operating system, advanced mathematic, college English (advanced) while $\ell_1$-norm support vector regression finds out key courses windows application design, computer composition and structure, operating system, public elective course, college English (advanced), object-oriented programming (based on java).

Above all, we can conclude that the two methods can obtain comparatively front key courses. In addition, the two methods can find out other courses for which many students obtained low scores but they cannot be found by simple statistics. This is a special advantage of machine learning methods that can explore the hidden information of data. From all figures, $\ell_1$-norm SVR with $\epsilon$-insensitive loss function can select more key courses compared with the lasso method in most cases.

## 4 Conclusion

In this paper, we use lasso and $\ell_1$-norm SVR to select key courses on our collected academic warning datasets. The experimental results show that the two methods can obtain comparatively accurate key courses.

**Acknowledgment**

The corresponding author Shiliang Sun would like to thank support by the National Natural Science Foundation of China under Project 61370175, and Shanghai Knowledge Service Platform Project (No. ZF1213).

## References


1. Sun, S.: Computational education science and ten research directions. Communications of the Chinese Association for Artificial Intelligence, vol. 5, pp. 15-16 (2015)
2. Dai, J., Li, M., Li, W., Xia, T., Zhang, Z.: Application of Monte Carlo simulation in college and university academic warning. Advanced Materials Research, vol. 955-959, pp. 1817-1824 (2014)
3. Dai, J., Li, M., Li, W., Xia, T., Zhang, Z.: Setting of academic warning based on multivariate copula functions. Applied Mechanics & Materials, vol. 571-572, pp. 156-163 (2014)
4. Taylor, J., Lawrence, J.: Making students AWARE: an online strategy for students given academic warning. Studies in Learning Evaluation Innovation & Development, vol. 4, pp. 39-52 (2007)
5. Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective, Journal of the Royal Statistical Society, vol. 73, pp. 273-282 (2011)
6. Meinshausen, N., Bhlmann, P.: High-dimensional graphs and variable selection with the Lasso, Annals of Statistics, vol. 34, pp. 1436-1462 (2006)
7. Vidaurre, D., Bielza, C., Larrañaga, P.: A survey of $L1$ regression. International Statistical Review, vol. 81, pp. 361-387 (2013)
8. Sun S., Huang, R., Gao Y.: Network-scale traffic modeling and forecasting with graphical lasso and neural networks. Journal of Transportation Engineering, vol. 138, pp. 1358-1367 (2012)
9. Shawe-Taylor, J., Sun, S.: A review of optimization methodologies in support vector machines. Neurocomputing, vol. 74, pp. 3609-3618 (2011)
10. Shawe-Taylor, J., Sun, S.: Kernel methods and support vector machines. Book Chapter for E-Reference Signal Processing, Elsevier, DOI:10.1016/B978-0-12-396502-8.00026-7 (2013)
11. Zhang Q., Hu X., Zhang B.: Comparison of $\ell_1$-norm SVR and sparse coding algorithms for linear regression. IEEE Transactions on Neural Networks and Learning Systems, vol. 26, pp. 1828-1833 (2015)