An Optimal Temporal and Feature Space Allocation in Supervised Data Mining

S. Tom Au, Guangqin Ma, and Rensheng Wang,

Abstract—**This paper presents an expository study of temporal data mining for prediction of a future response sequence via mining large number of highly correlated concurrent time series. In the study, we investigate a two dimensional search scheme over time domain weighting and feature space selection. The weighting of observation records over time domain is used to exploit the time dependency structure and feature space selection is enforced to avoid the over-fitting issue. For a specific temporal and spatial selection, its area under ROC curve (AUC) is used to evaluate the prediction performance over the training and testing data. By varying the weighting scheme and feature selection, AUC contour maps on both training and testing data are generated. The contour maps can suggest us to apply the optimal allocations with highest AUC for future responses prediction in training, testing , and possible validation data. Numerical results over two sets of temporal data with different applications have shown that the proposed scheme can improve the prediction performance of conventional data mining methods significantly.**

Index Terms—**temporal data mining, AUC, weighted logistic regression**

I. Introduction

TEMPORAL data mining is a research field of growing interest in which techniques and algorithms are applied on data collected over time [11] [15] [9]. Business market forecasting, financial or stock market prediction and medicine records pattern clustering etc have been some of the oldest and most studied applications [6] [2] [5]. As a financial industry example, INFORMS (institute of operations research and the management sciences) recently has organized a 2010 data mining contest to predict short term movements in stock prices by exploiting hundreds of related stock market time series. Some recent research works are also reported in [13] [3] [8]. In telecommunications, researchers are mining spatial and temporal data for detecting events that impact mobility networks [4]. In a similar scenario, people are interested in how to forecast the future network capacity demand for locations of interest by utilizing a set of spatial correlated and time dependent information. Many similar mining applications can be found in other fields, like weather forecasting, patients status prediction, etc. A common task for these different applications is how to predict a future response via numerous of related sequences over the past. This problem also raise the interests about how once can use the conventional statistical methods, such as logistic regression [10] [12], bootstrapping [7] etc, for predictive modeling with temporal and feature space data. An unique feature of prediction in temporal data mining is the time dependency structure inherent with the data sets. However, many aforementioned methods usually deal with non-time dependency, and simply treat as cross sectional data, i.e., the order of records does not affect the performance. If the time dependency structure is also considered in the mining scheme development, we are interested in how it can improve existing conventional prediction performance. Motivated by this problem, we propose a logistic regression based scheme which make varying weighting vectors over time domain and time series selection over feature space. Essentially, it is a weighted logistic regression (WLR) scheme with weighting searching over time and feature selection over space. For each fixed temporal and spatial combination, we use area under the curve (AUC) to evaluate the prediction performance over training data. After sweeping all the possible selections, an AUC contour map can be created to indicate the optimal allocations with the highest AUC. We then apply them to the test data for prediction. To validate our proposed scheme, we use two sets of temporal data with different applications. In both cases, our method outperforms the conventional logistic regression methods significantly.

II. Data Model

Let us assume $T \times p$ predictor matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \ \mathbf{x}_p]$ contains p time series, each with length $T \times 1$ and $\mathbf{y} = [y(1), y(2), \cdots, \ y(T)]^T$ as the response of a target time series. $\mathbf{y}$ is not necessary to be synchronized with $\mathbf{X}$, and it can be some time after, i.e., prediction of future movements with time series. Denote $x_1, x_2, \cdots, x_p$ as individual variable in $\mathbf{X}$ We have training data structure

$$\{y(t), \quad x_1(t), x_2(t), \cdots, x_p(t), \quad t = 1, 2, \cdots, T\};$$

where the time series $y(t) = \{0, 1\}$ is a binary response. Following that, we have a test set as $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \ \mathbf{u}_p]$ containing the same set of time series, but with starting time index at $t = T+1, T+2, \cdots$, and so on.

The problem of interest here is how the time index information can be embedded in the conventional regression models to improve the prediction performance over those disregarding the time order.

III. Study Design

The aforementioned problem is essentially to use a set of records to predict a target response. This falls into the conventional category of statistical methods for predictive modeling, such as logistic regression, bootstrapping etc. However, these methods usually ignore the time-related orders inherent with time series, and just simply treat every individual record as independent from each other, i.e., shuffling the records does not affect the performance. Here we use logistic regression model as a benchmark for our proposed scheme comparisons.

Manuscript received December 27, 2010; revised February 4, 2011.

A. Tom, G. Ma and R.Wang are with AT&T Research labs, Florham Park, NJ, 30332 USA e-mail: {sau,gma,rw218j}@att.com.

A. *Conventional Logistic Regression*

The conventional logistic regression is to predict individual vector $\mathbf{y}$ from $\mathbf{X}$ with probability predictions. Logistic regression is a member of the family of methods called generalized linear models (GLM). Such models include a linear part followed by a "logistic function". Mathematically, we will have estimated probability vector $\mathbf{p}$

$$\mathbf{p} \equiv P_r(\mathbf{y}|\mathbf{X}) = \text{Logit}(\beta_0 + \mathbf{X}\boldsymbol{\beta}) \qquad (1)$$

where β_0 and $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_p]^T$ are obtained from logistic regression, for given observations $\mathbf{X}$ with response $\mathbf{y}$.

B. *Proposed Scheme*

We note that the conventional logistic regression does not take advantage of the time index information from the time series order. Since the test set data $\mathbf{U}$ are exactly the following time series records after $\mathbf{X}$, the more recent records in $\mathbf{X}$ should reveal more reliable information for the adjacent response prediction with $\mathbf{U}$. In other words, $\mathbf{w}$ has monotonically increasing values over time. Based on this assumption, we impose an extra time importance weighting vector $\mathbf{w}$ over the response vector $\mathbf{y}$ and try to enhance the forecasting reliability. That is, the refined logistic regression should be a weighted logistic regression

$$\mathbf{p} \equiv P_r([\mathbf{y},\ \mathbf{w}]|\mathbf{X}) = \text{Logit}(\beta_0 + \mathbf{X}\boldsymbol{\beta}) \qquad (2)$$

where $\mathbf{w}$ is a vector containing weighting factors over time horizon. In fact, having a WLR is not new in the literature [14]. However, where and how to put the optimal weighting factors is more important.

1) Weighting Scheme: To generate weighting factors $\mathbf{w} = [w(1), w(2), \cdots,\ w(T)]^T$, we use piecewise function to map the time index (1,2, $\cdots$, T) into three weighting areas, where we need to set up two thresholds. The first and the third areas are flat line with constant weighting values, while the second area is transition area with monotonically increasing values. Mathematically, we have

$$w(t) = \begin{cases} 0, & t \leq T_1; \\ \frac{e^{f(t)}}{1+e^{f(t)}}, & T_1 < t \leq T_2; \\ 1, & T_2 < t \leq T; \end{cases}$$

where $f(t) = 10\frac{t-(T_1+T_2)/2}{T_2-T_1}$ to make sure $-5 \leq f(t) \leq 5$ and $\frac{e^{f(t)}}{1+e^{f(t)}}$ has smooth curve range $(0, 1)$. T_1 and T_2 are selected as percentage points of the time index. Generally, we select the transition area between T_1 and T_2 as 40% of total time index. The remaining 1 and 0 points in $\mathbf{w}$ are around 60%. We slide the cutting point T_2 to generate and identify different weighting vectors. For example, when $T_2 = 50\% \times T$, then, the first 10% are zeros and the last 50% are all ones, and between them are transition values between $(0, 1)$. However, when $T_2 = 20\% \times T$ is less than the 40% transition length, we will make truncation, i.e., we only have the first 20% transition points and the remaining 80% are all ones.

For extreme situation, $T_2 = 0$, we have $\mathbf{w} = [1, 1, \cdots, 1]^T$ equals all ones, which is identical to the case where there is no weighting vector imposed on the logistic regression. As such, we can always use the $T_2 = 0$ weighting scheme (i.e., conventional no weighting) to compare with other weighting schemes.

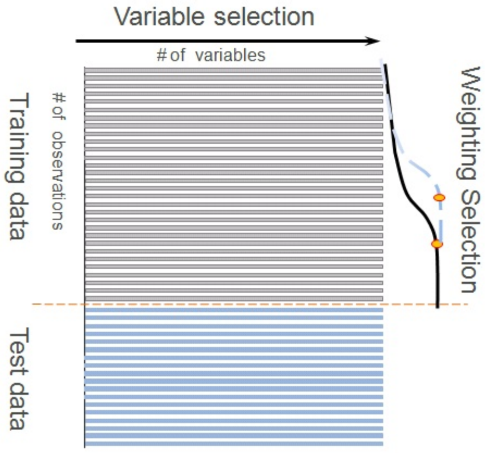


Fig. 1. Weighted logistic regression (WLR) is implemented over two dimensional tuning: One is to change the number of variable that will be included in the WLR; and the other is to use a sliding window to change the weighting vector over the number of observations that will be imposed on the WLR.

2) Procedures for proposed scheme: In this paper, we propose that, the weighting vector should be chosen to slide over the time index and meanwhile the time series in predictor matrix $\mathbf{X}$ should be selected over different variables. For every selected combination with the weighting vector location and the number of variables, we have area under curve (AUC) to evaluate the prediction performance. In total, from the two dimensional tuning, an AUC contour map can be created to indicate which parameter combination shows the highest AUC performance. We then can apply the same set-up to the test data. If multiple areas in the contour map stand out, we can use the linear combination of them for testing.

In summary, we have below 5-step procedures for our proposed scheme:

1) Use standard logistic regression as criterion to rearrange the orders of the time series in $\mathbf{X}$. The first column of the re-arranged data matrix $\mathbf{X}$ can best predict the data in $\mathbf{y}$ and the second one has the second best prediction, and so on and so forth for the remaining ones.
2) Set loop 1 over all p variables: `for` i `in the set of` $\{1, 2, \cdots, p\}$ and select the first i columns of $\mathbf{X}$ as
$$\mathbf{X}_{\text{sel}} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i]$$
3) Given $\mathbf{X}_{\text{sel}}$, set loop 2 `for` j `in the set of` $\{0,\ 1,\ 2,\ \cdots,\ 20\}$, compute different weighting factors, with $T_2 = j \times 5\% \times T$;
4) For given $\mathbf{X}_{\text{sel}}$ and T_2, use equation (2) to compute `AUC` in WLR.
$$\mathbf{AUC}[i, j] = \texttt{AUC};$$
5) At the end of loop 1 and loop2, use two-dimensional AUC map to find the optimal combination of the number of time series from $\mathbf{X}$ and weighting factor $\mathbf{w}$ for $\mathbf{y}$ prediction. Use the same setup with testing data $\mathbf{U}$.

3) Measuring Scenarios: In the 5-step procedures, we assume that the optimal performance in the training data contour map will be the same as that in the test data set. Note that we are dealing with time-series data set which are varying over time. One might concern how the consistent performance can be achieved between training and testing. To clarify this concern, we use figure 2 to show how the proposed scheme is implemented in two different scenarios, where a red dash line is used to separate the total known time series records for training and those records for prediction.

In Scenario 1, the training data is sliced as pre-training and validation data sections. We use the aforementioned 5-step procedures to treat pre-training part as training while validation part as testing. Since both are within the training data (we know the response $\mathbf{y}$), we can easily evaluate both contour maps in pre-training and validation parts. We expect that they should be consistent. After we are confident with this result, we proceed to use the whole training data and testing data to repeat the 5-procedures again. Again, the training and testing results should be consistent on the AUC contour map.

In Scenario 1, the training data is sliced as pre-training and validation data sections. We use the aforementioned 5-step procedures to treat pre-training part as training while validation part as testing. Since both are within the training data (we know the response $\mathbf{y}$), we can easily evaluate both contour maps in pre-training and validation parts. We expect that they should be consistent. After we are confident with this result, we proceed to use the whole training data and testing data to repeat the 5-procedures again. Again, the training and testing results should be consistent on the AUC contour map.

In Scenario 2, we take the most recent part as the validation part. When implementing 5-step procedures, unlike scenario 1, we use the whole set of records of $\mathbf{X}$ for training. We apply the same combination setup to both validation data and testing data. In total, we use WLR once to generate three contour maps and expect all of them should be consistent.

Comparing with these two scenarios, we can see that scenario 1 separate the testing part from training data in two WLR implementations, while the scenario 2 contains validation part in training. On the other hand, scenario 1 use two WLR, where the performance in the first WLR implement does not necessarily be consistent with the second WLR since the second WLR contains more recent data which is different from the first WLR. Meanwhile, one can see that scenario 2 contains the most recent data and only has one WLR.

IV. Numerical Results

We like to compare the proposed optimal allocated WLR in (2) with the normal method in (1). The criterion is the area under ROC curve (AUC) with probabilities prediction. The AUC contour map is generated by SAS graphics program [1]. As we explain in the subsection "Weighting Scheme", when threshold T_2 is selected as $T_2 = 0\% \times T$, the weighting vector $\mathbf{w}$ contains all ones, i.e., no weighting. Under this situation, WLR in (2) reduces to be normal logistic regression in (1).

The time series used for training and test are from some anonymous industrial daily sales data. We apply both scenarios for comparisons. Figure 3 demonstrates all four AUC contour maps from four separate steps in Scenario 1,where Step 1 and Step 2 results are from the first WLR and Step 3 and Step 4 are from the second one. We first notice that in both WLR, the pair of AUC maps display similar, i.e., consistent performance pattern. Then, we note that, from the first WLR to the scond one, the AUC map patterns change a little bit. This is due to extra time series records included in the second WLR. One can conclude that the AUC map pattern can vary over time but the training and test have the comparable results. Most importantly, AUC maps exhibit that the best performance does not happen at $T_2 = 0\% \times T$, which indicates that WLR in (2) outperforms the logistic regression without weighting in (1).

The corresponding scenario 2 results are shown in Figure 4. Note that there is only one WLR involved and all three of them are displaying similar performance pattern. The drawback in this scenario is that we repeatedly use the same data for training and validation. Similar to previous scenario, the WLR in (2) outperforms the logistic regression without weighting in (1).

To show this proposed scheme can be applied for different application fields, we implement it with the popular S&P 500 stock market daily prices to predict the following day index price movement. The comparison of training and testing data is presented in Figure 5. We note here that we are interested in how to generate new features or variable time series for the best prediction mining scheme. Instead, we are look for an extra performance enhancement via our proposed optimal allocation search for a given set of feature space data. Obviously, both contour maps show the extra performance improvements can be achieved by an optimal allocation search. We have the same area in both maps which shows the highest AUC values. That is, the optimal allocation area in training is validated in the testing data.

In sum, for either scenario, a two dimensional AUC map from training data manifests the optimal combinations of the weighting vector and the selected time series features, which can be utilized to apply for the best test data prediction.

V. Conclusions

In this study, we investigate a classic predictive modeling for temporal data mining, i.e., utilizing a number of past time sequences to predict a related future response movement. We propose a time-domain weighting and feature space selection search scheme which sweeps all the allocation combinations simultaneously. After this two dimensional tuning, an AUC contour map can be created to reveal where the optimal allocations lie on. We then apply them to test data for performance enhancement. Two different implementation scenarios are considered and both sets of the results show WLR with optimal allocations outperforms the conventional logistic regression without weighting significantly.

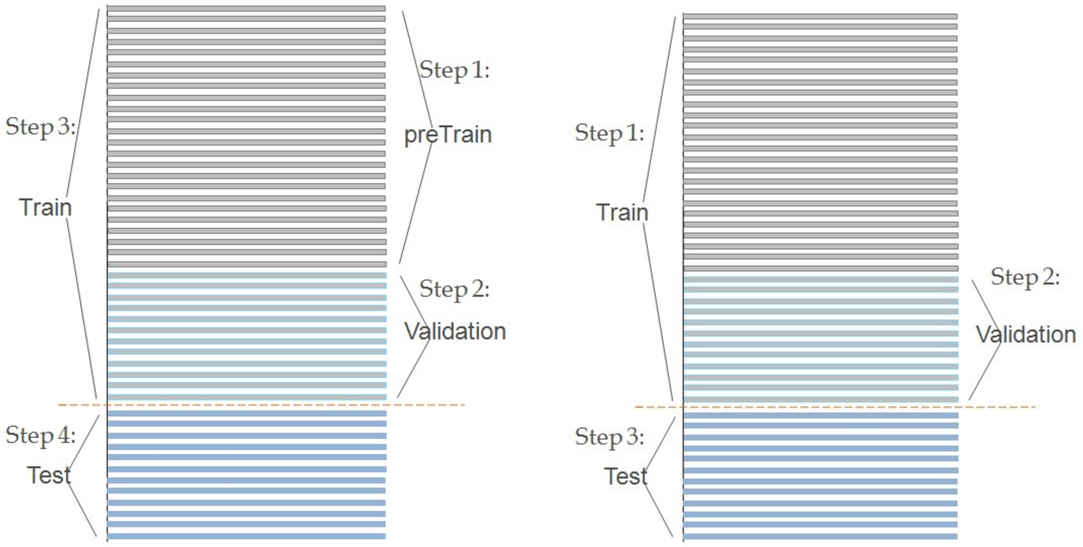
References

[1] SAS Institute Inc. SAS-Graphics. *Cary, NC, USA*, 2001-2003.

[2] R. Agrawal, C. Faloutsos, and A. Swami. Efficient Similarity Search in Sequence Databases. *Proc. 4th Int. Conf. Foundations of Data Organizations and Algorithms (FODO 93)*, Chicago, IL USA:69–84, 1993.

[3] G. Atsalakis and K. P. Valavanis. Surveying Stock Market Forecasting Techniques Part II: Soft Computing Methods. *Expert Systems with Applications*, 2009.

[4] S. Au, R. Duan, H. Kim, and G. Ma. Spatial-temporal events detection in mobility network. In *The 10th IEEE International Conference on Data Mining*. ICDM 2010, Sydney, Australia, December 13-17 2010.



(a) Scenario 1.

(b) Scenario 2.

Fig. 2. Two different Scenarios. (a) Implement WLR twice. (b) Implement WLR once.

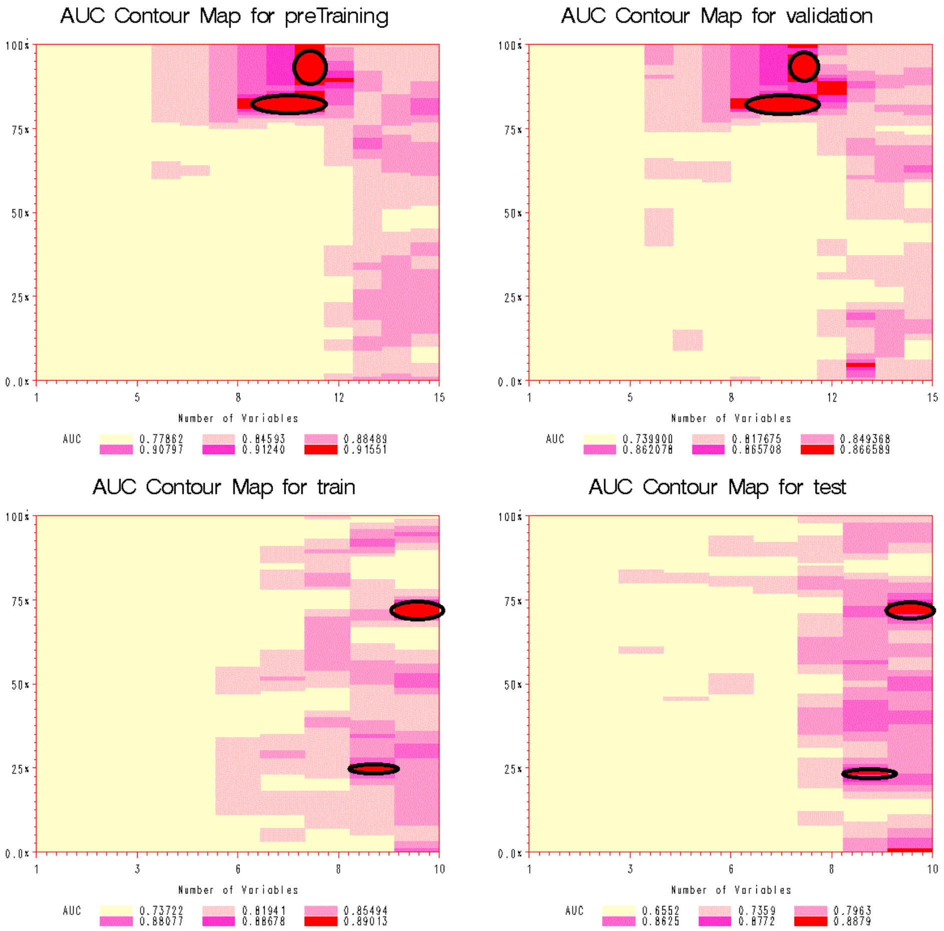


Fig. 3. Comparison of AUC within scenario 1 for anonymous industrial daily data.

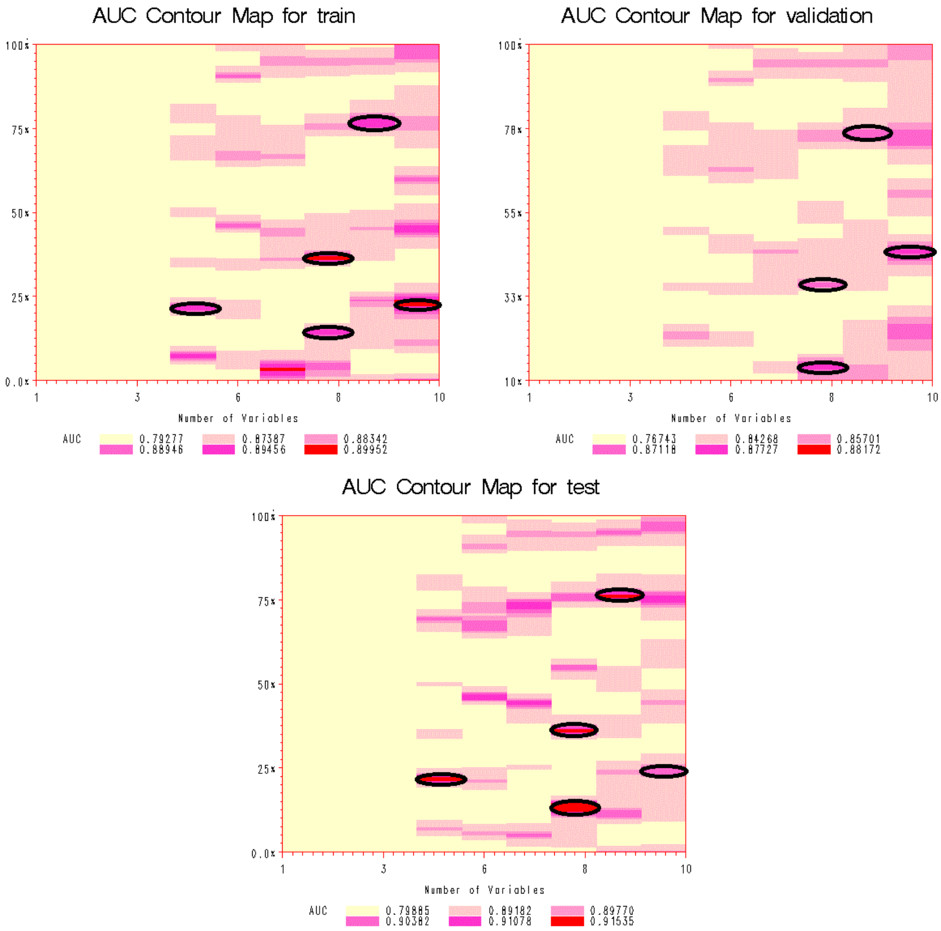


Fig. 4. Comparison of AUC within scenario 2 for anonymous industrial daily data.

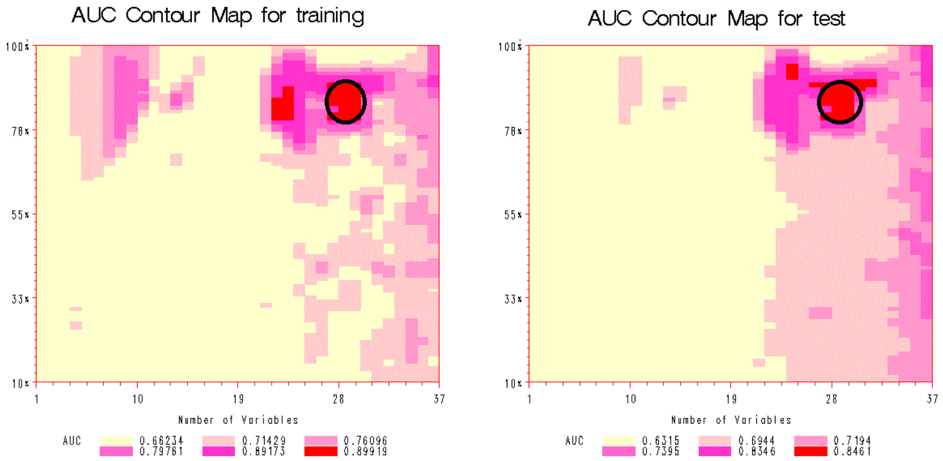


Fig. 5. Comparison of AUC within scenario 2 for stock market data.

[5] K.-P. Chan and A.-C. Fu. Efficient Time Series Matching by Wavelets. *Proc. 15th Int. Conf. Data Engineering*, Sydney, Australia:126–133, 1999.

[6] G. Das, D. Gunopulos, and H. Mannila. Finding Similar Time Series. *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining (KDD 97)*, Newport Beach, California, USA:88–100, 1997.

[7] J. Friedman and T. Haste. Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics*, vol.28, No.2:337–407, 2000.

[8] M. Hassan and B. Nath. Stock Market Forecasting Using Hidden Markov Model:A New Approach . *Fifth International Conference on Intelligent Systems Design and Applications (ISDA05)*, pages 192–196, 2005.

[9] W. Lin, M. Orgun, and G. Williams. An Overview of Temporal Data Mining. *Proc. of the Australasian Data Mining Workshop (ADM02)*, Sydney, Australia:83–90, 2002.

[10] S. Menard. *Applied logistic regression analysis*. Sage Pubns, 2002.

[11] G. Montana, K. Triantafyllopoulos, and T. Tsagaris. Flexible least squares for temporal data mining and statistical arbitrage. *Expert Systems with Applications*, vol.36:28192830, 2009.

[12] L. Ott and M. Longnecker. *An introduction to statistical methods and data analysis*. Duxbury Pr, 2008.

[13] S. Rao and J. Hong. Analysis of Hidden Markov Models and Support Vector Machines in Financial Applications. *EECS department, University of California, Berkeley, Technical Report No. UCB/EECS-2010-63*, http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-63.pdf, May 12, 2010.

[14] M. Wang and A. Yu. Proportionally Weighted Logistic Regression Model. *14th Pacific-Asia Knowledge Discovery and Data Mining (PAKDD) Competition 2010*, March, 2010.

[15] A. Weigend and N. Gershenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley, 1994.