Improving Stock Market Prediction by Integrating Both Market News and Stock Prices

Xiaodong Li[1], Chao Wang[2], Jiawei Dong[2], Feng Wang[3], Xiaotie Deng[1,4], and Shanfeng Zhu[2]*

[1] Department of Computer Science, City University of Hong Kong, Hong Kong
xiaodonli2@student.cityu.edu.hk

[2] Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University, Shanghai 200433, China
zhusf@fudan.edu.cn

[3] School of Computer and State Key Lab of Software Engineering, Wuhan University, Wuhan 430072, China

[4] Department of Computer Science, University of Liverpool, Liverpool, UK

Abstract. Stock market is an important and active part of nowadays financial markets. Addressing the question as to how to model financial information from two sources, we focus on improving the accuracy of a computer aided prediction by combining information hidden in market news and stock prices in this study. Using the multi-kernel learning technique, a system is presented that makes predictions for the Hong Kong stock market by incorporating those two information sources. Experiments were conducted and the results have shown that in both cross validation and independent testing, our system has achieved better directional accuracy than those by the baseline system that is based on single one information source, as well as by the system that integrates information sources in a simple way.

Keywords: Stock market prediction; Information integration; Multi-kernel learning.

1 Introduction

Stock market is an important and active part of nowadays financial market. Both investors and speculators in the market would like to make better profit by analyzing market information. In Efficient Market Hypothesis (EMH, proposed by Fama [1]), it is thought that stock prices have already included and revealed all the information in the market, and that random walk is the most natural and possible way the stock market should behave. However, researchers in behavioral finance argue that EMH may not be right because of irrational behavior of players who are influenced by various kinds of market information as well as their psychological interpretation of the information [2]. Although there

* Corresponding author.

A. Hameurlain et al. (Eds.): DEXA 2011, Part II, LNCS 6861, pp. 279–293, 2011.

are differences between those two theories, neither of them ignores the effect of market information.

News articles, known as one of the most important part of market information, are widely used and analyzed by investors. With the development of Internet, both the speed of news broadcasting and the amount of news articles have been changing tremendously: 1) Bloomberg[1] and Thomson Reuters[2] could provide investors around the world with real-time market news by network; and 2) the number of online news articles could be thousands of times than that in the past newspaper-only age. With such a big volume of information, more and more institutions rely on the high processing power of modern computers for information analysis. Predictions given by support systems could assist investors to filter noises and make wiser decisions. How to model and analyze market information so as to make more accurate predictions thus becomes an interesting problem.

Researchers with computer science background have studied this problem, and some works modeled it as a classification problem [3,4,5,6]. Their algorithm could give a directional prediction (up/hold/down) based on the newly released news articles. Text classification, however, only considers news articles' impact, but ignores information hidden in the prices shortly before news is released. Taking into consideration both news articles and short-time history price, we believe that positive news may not always lead to going up the price immediately, it might just stop the price trend from falling down. Negative news does not necessarily drive the trend from up to down. Instead, it might just make price curve appear flat. Figure 1 illustrates several possible scenarios that could happen. This is different from the traditional view, that is, "good news means up, bad news means down".

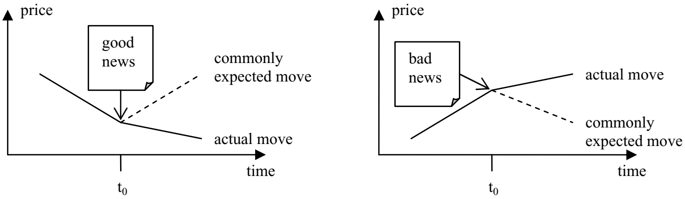

Fig. 1. Possible scenarios of price movements based on news articles and short time history price

In order to aggregate more-than-one information sources into one system, Multi-Kernel Learning (MKL) is employed in our system. The MKL has two sub-kernels: one uses news articles and the other accepts the short-time history prices. After learning the weights for sub-kernels, the derived model gives prediction that is supposed to be more accurate than traditional methods.

[1] `http://www.bloomberg.com/`

[2] `http://thomsonreuters.com/`

The rest of this paper is organized as follows. Section 2 reviews major existing approaches related to stock market directional prediction. Section 3 gives an overview of our proposed system and brief information about experimental design. Experimental results are reported in Section 4. The conclusion and future work are given in Section 5.

2 Related Work

Some helpful observations and discussions about news and market prices are presented in finance domain. Ederington and Lee [7] observed that there is always a big increment of standard deviation of five-minutes returns on the day that a government announcement released at 8:30am. Engle and Ng [8] claim that positive and negative news present asymmetry impact curves, based on the empirical analysis of ARCH model family.

Analyzing news articles and market prices has also been reported in many works in computer science domain. Seo, Giampapa and Sycara [9] built a TextMiner system (a multi-agent system for intelligent portfolio management), which could assess the risk associated with companies by analyzing news articles. Fung and Yu [4] classified news articles into categories and predict newly released news articles' directional impact based on the trained model. AZFinText system, built by Schumaker and Chen [10], is also able to give directional forecast of prices.

As illustrated in Figure 2, the common steps of those works in general could be summarized as follows:

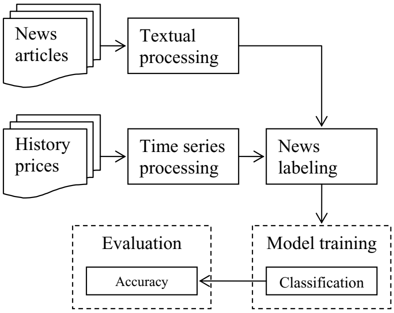

Fig. 2. Pipeline of traditional approach

1. **Representation of news articles.** News articles are basically textual documents. The vector space model from information retrieval treats textual documents as "bag of words", where each word is in a long vector $\langle word_1,\ word_2,\ \ldots\rangle$ without any duplicate. Weights w are assigned to words (or terms, features, etc.) by measuring their term frequencies and inverse document frequencies i.e. $w = tf \cdot idf$. Stop words, e.g. *the*, *of* and *is*, which frequently occur but are less informative, are removed in order to reduce

the noises in corpora. However, remaining features still form a very sparse space which requires a lot of memory and computation power. To reduce the feature space, dimension reduction and feature selection methods are applied [11,12,13,14].

2. **Representation of price data.** Price data is a series of trading statistics. At each time t_i, there is a corresponding trading record r_i. History price data is of different quality: 1) Inter-day data. Generally contains open, close, high, low and volume for each trading day, and daily collected; and 2) Intraday data. a.k.a. tick-by-tick data, the trading statistics collected in a smaller time unit, e.g. minute or second. Since price data is not smooth, time series segmentation techniques, such as the parametric spectral model [15] and fourier coefficient [16], are applied to smooth the price curve in order to emphasize the trend of prices.

3. **Alignment and news labeling.** Fung *et al.* [4] formulate the alignment of news articles and price data. They classify possible scenarios into three categories: 1) *Observable Time Lag* - a time-lag between news and price moves; 2) *Efficient Market* - No observable time-lag, price moves at almost the same time with news; 3) *Reporting* - News released after price moves, a summary or report of previous price moves. Alignment of news article with price data mainly relies on the time stamps attached with the news articles. News articles are sorted by their time stamps in ascending order and then aligned with price data in the corresponding time slots. Before training the classifier, news articles should be labeled with a tag indicating the directions of their impact. Besides simply labeling news articles by the trend of aligned price movement, linguistic methods and sentiment mining are also implemented for this purpose [17,18,19,20].

4. **Model learning.** Machine learning models, such as support vector machine, are used as classifier in this area because of their relatively short training time and comparatively high classification accuracy. Take SVM for example, SVM is fed with the labeled news articles. By selecting some points as support vectors, SVM finds a hyperplane that maximizes marginal space from hyperplane to the support vectors. After the training phase, prediction model is built up to make predictions for new coming data.

5. **Evaluation.** Besides evaluating the model by some standard benchmarks, such as precision, recall and accuracy, some researchers [3,21] conduct a preliminary simulation, which makes trades (buy/hold/sell) in a virtual market with real market data. Trading strategies are made on basis of the signals generated by prediction model. *Return rate* is calculated to measure the performance of different models. However, the performance in the simulation actually depends not only on the prediction model, but also on the trading strategies and risk management, which is beyond the discussion scope of this paper.

3 System and Experimental Design

Unlike the traditional way that considers only single information source, our system is designed to enable to integrate multiple information sources. The architecture of the system is shown in Figure 3.

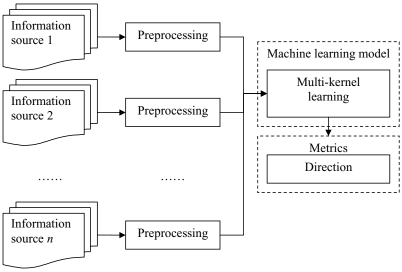

Fig. 3. Architecture of system integrating multiple sources

Two information sources are used in our system: market news and the history prices shortly before the news released, the later of which is referred to as *ex ante* prices in the rest of this paper. The processing pipeline is illustrated in Figure 4.

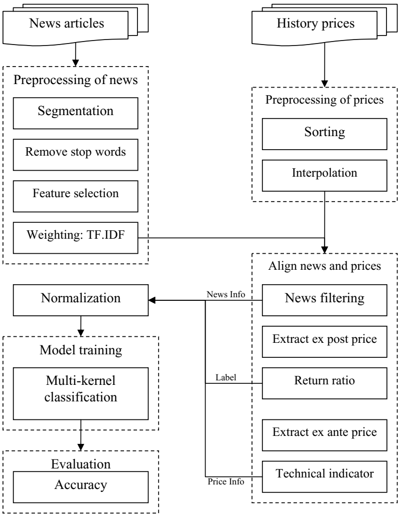

Fig. 4. Detail processing pipeline of our system

Section 3.1 describes brief information about the information sources fed into the system. Section 3.2 and Section 3.3 will talk about the processing of news articles and history prices respectively. Section 3.4 presents how to align news and prices. Data normalization and model training are discussed in Section 3.5 and Section 3.6.

3.1 Information Sources

The system is designed in a way that enables to integrate two information sources, which comes from news articles and the ex ante prices. The input data should have following characteristics:

- **Time stamped**. Each news article is associated with a time stamp with proper precision indicating when the news was released. With this time stamp, the system could identify the order of news and find corresponding price information.
- **Tick based**. Tick based data means trading data is often recorded in a short interval.
- **Parallel**. Since system needs to tag news using price movement, news articles and history prices should be about the stories of the same time period.

3.2 Preprocessing of News Articles

News articles are regarded as raw materials that need to be preprocessed. The main steps are listed below:

1. **Chinese segmentation**. We segment the news articles by using an existing Chinese segmentation software[3]. Although the segmentation software could produce outputs with high quality, many words that are specific to finance domain cannot be segmented correctly. A finance dictionary is thus employed to refine the segmentation results.
2. **Word filtering**. This step actually does two things: 1) remove stop word; and 2) filter out other unimportant words (only leave representative words, such as nouns, verbs and adjectives).
3. **Feature selection**. Not all the words would be included into the final feature list. Feldman [22] (Chapter IV.3.1) selects about 10% of the words as features. Similarly, top 1000 words (out of 7052 words after filtering) with high χ^2 score are selected as features in the system.
4. **Weighting**. With the selected 1000 features, we calculate the widely used $tf \cdot idf$ value for each word as its weight.

3.3 Preprocessing of History Prices

With the development of high frequency trading, tick data is popularly used so that results based on tick data are more convincing. With the following properties, tick data are distinguished from daily data:

[3] http://ictclas.org/

- **Big amount**. Tick data have much more records than daily data over the same time period.
- **Disorder**. Tick data is not recorded by the order of their time stamps, but by the time they arrive at the logging system.
- **Variant interval**. Time intervals between different records may not be the same. As transactions may happen in any seconds, a time interval between consecutive records is not always the same.

Raw tick price data is preprocessed through following steps:

1. **Sorting**. Since transactions do not arrive in the order of their time stamps, we must first sort the whole list of records by their time stamps.
2. **Interpolation**. Since time intervals between consecutive transactions are not the same. Over some time periods, there even does not exist any record, which leads to one problem: what price value should be filled in that time period. There are two ways to solve this problem: 1) linear time-weighted interpolation proposed by Dacorogna *et al.* [23]; and 2) nearest closing price. This method splits tick data in a minute basis and samples the closing price in each minute. If there is no record in a given minute, the closing price of last minute will be taken as the closing price of this minute. Although both methods make sense, we adopt the second method, which is simple and easy to implement.

3.4 Align News Articles and Market Prices

In order to train machine learning model with information from two sources, we need to prepare the raw data and change them as what the algorithm needs.

3.4.1 News Filtering

(Not all the news articles are used because of two reasons: 1) *trading hour limitation.* Take Hong Kong Exchanges and Cleaning Ltd (HKEx) for example, according to regulation of HKEx, 10:00am-12:30pm and 14:30pm-16:00pm are trading hours. Only news articles released during trading hours are considered to have impact on HKEx stock prices. Besides the trading hour limitations, Schumaker [21] suggests eliminate the opening 20 minutes in the morning and opening 20 minutes in the afternoon in order to absorb the impact of news that is released during the night and lunch break. 2) *news impact overlapping.* As illustrated in Figure 5, if the prediction length is Δ and there are two news articles (d_1 and d_2) released within Δ. In this scenario, it is hard to tell whether the change of price at time $t_{+\Delta}$ is caused by either d_1 or d_2 or both. In our system, d_1 will be eliminated.

3.4.2 Extract and Process Ex Post Prices

Before feeding news article into machine learning model, each news should have a category tag. In high frequency trading, people would like to know what is the short-time impact of news on prices, which means people are much more caring about the price change shortly after the news released (named as *ex post*

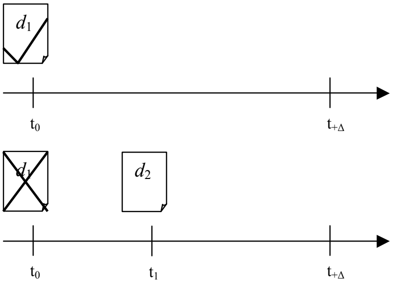

Fig. 5. Example of news filtering

prices). Gidofalvi [24] shows that news impact has the biggest power 20 minutes after it's released. Without the knowledge on how long the news impact lasts, we label the news by the change of price in future 5, 10, 15, 20, 25 and 30 minutes respectively, which can be regarded as an extension of work [21].

Corresponding to news time stamp t_0, current price value p_0 in the sorted tick price series as well as the prices of future 5, 10, 15, 20, 25 and 30 minutes intervals denoted as p_{+5}, p_{+10}, p_{+15}, p_{+20}, p_{+25} and p_{+30}, respectively, can be found. if $t_0 + \Delta$, for example, $t_0 + 20$, exceeds the requirement of *trading hour limitation*, the news article will be eliminated. We convert ex post prices into the return rates by

$$R = \frac{p_i - p_0}{p_0}$$

We set a threshold of 0.3% (average transaction cost in the market), which means if R is greater than 0.3%, news is tagged as positive. In contrast, news is tagged as negative if R is less than -0.3%.

3.4.3 Extract and Process Ex Ante Prices

Prices from 30 minutes to 1 minute before news is released, and sampled at 1 minute interval are extracted as ex ante prices in our experiment. However, if we naively take the 30 points as 30 features, the machine learning model will assume that the 30 features are independent from each other, which means the sequential dependency of the price serial p_{-30}, p_{-29}, ..., p_{-1} is not preserved and the machine learning model will not be able to use the sequence information. Cao and Tay [25,26] convert the price series into RDP indicators. Following their method, we use the same formulae of RDPs which are listed in Table 1.

In addition to RDPs, we employ some other market indicators from stock technical analysis. The formulae of market indicators are listed in Table 2, where p_i is the price at minute i, and q refers to the order counted in minute.

After all, 30 ex ante price points are converted to 6 RDPs and 5 market indicators, all of which will be simply referred to as indicators in the following sections.

Table 1. The formulae of RDPs

RDP	Formula
RDP-5	$100 * (p_i - p_{i-5})/p_{i-5}$
RDP-10	$100 * (p_i - p_{i-10})/p_{i-10}$
RDP-15	$100 * (p_i - p_{i-15})/p_{i-15}$
RDP-20	$100 * (p_i - p_{i-20})/p_{i-20}$
RDP-25	$100 * (p_i - p_{i-25})/p_{i-25}$
RDP-30	$100 * (p_i - p_{i-30})/p_{i-30}$

Table 2. Indicator

Indicator	Formula	Description
RSI(q)	$100 * UpAvg/(UpAvg + DownAvg)$ $UpAvg = \sum_{p_i > (\sum_i p_i)/q}(p_i - (\sum_i p_i)/q)$ $DownAvg = \sum_{p_i < (\sum_i p_i)/q}(p_i - (\sum_i p_i)/q)$	Relative Strength Index
RSV(q)	$100 * (p_0 - \min_q(p_i))/(\max_q(p_i) - \min_q(p_i))$	Raw Stochastic Value
R(q)	$100 * (\max_q(p_i) - p_0)/(\max_q(p_i) - \min_q(p_i))$	Williams Index
BIAS(q)	$100 * (p_0 - (\sum_i p_i)/q)/((\sum_i p_i)/q)$	Bias
PSY(q)	$100 * (\sum 1\{p_i > p_{i-1}\})/q$	Psychological Line

3.5 Normalization

After performing the previous steps, we have obtained: 1) group of news instances (denoted as N); 2) group of indicator instances (denoted as I); and 3) vector L containing labels. Each instance of N corresponds to one piece of news and each feature of instance corresponds to one selected word. Each instance in I also corresponds to one piece of news and each feature in I corresponds to one of the indicators. For features in N and features in I that only takes non-negative value, denoted as f_k, we use

$$norm(w_{ki}) = \frac{w_{ki} - \min(w_{k*})}{\max(w_{k*}) - \min(w_{k*})}$$

to normalize. The range of values after the normalization is [0, 1]. For features in I that could take both positive and negative values, denoted as f_m, we use

$$norm(w_{mi}) = \frac{w_{mi}}{\max(w_{m*})}$$

to normalize. The range of values after the normalization is [-1, 1].

3.6 Model Training

We want to compare the ability of prediction between two information sources based model and single information source based model. SVM is selected to be the classifier. We implement four models that use information of news and ex ante prices in different ways. The details about the setup of those four models are described as follows:

1. **News article only**. This model takes labeled news instances as the input of SVM. It tests the prediction ability when there are only news articles. (Figure 6 (1))
2. **Ex ante prices only**. This model takes labeled price data as the input of SVM. It tests the prediction ability when there are only history prices. (Figure 6 (2))
3. **Naive combination of news article and ex ante prices**. This approach uses the simple combination of news articles and prices. Naive combination means combine the 1000 features from news and 11 features from indicators to form a 1011 feature vector. As instances of news and instances of indicators are one-one correspondence, the label for each instance is unchanged and the total number of instances remains the same. (Figure 6 (3))
4. **Multi-Kernel Learning (MKL)**. MKL is employed to aggregate the information within news articles and ex ante prices of each news (SHOGUN [27], an implementation of MKL, is used in our experiment.). Unlike naive combination which trains SVM using

$$f(\overrightarrow{x}) = \text{sign}(\sum_{i=1}^{m} \alpha_i l_i \mathbf{k}_{naive}(\overrightarrow{x}_i, \overrightarrow{x}) + b)$$

   where $\overrightarrow{x}_i, i = 1, 2, \ldots, m$ are labeled training samples of 1011 features, and $l_i \in \{\pm 1\}$, for the case of MKL, similarity is measured among the instances of news and instances of indicators respectively, and the two derived similarity matrices are taken as two sub-kernels of MKL (as shown in Figure 6 (4)) and weight β_{news} and $\beta_{indicator}$ are learnt for sub-kernels,

$$\mathbf{k}(\overrightarrow{x}_i, \overrightarrow{x}_j) = \beta_{news}\mathbf{k}_{news}(\overrightarrow{x}_i^{(1)}, \overrightarrow{x}_j^{(1)}) + \beta_{indicator}\mathbf{k}_{indiccator}(\overrightarrow{x}_i^{(2)}, \overrightarrow{x}_j^{(2)})$$

   with $\beta_{news}, \beta_{indicator} \geq 0$ and $\beta_{news} + \beta_{indicator} = 1$, where $\overrightarrow{x}^{(1)}$ are news instances of 1000 features and $\overrightarrow{x}^{(2)}$ are indicator instances of 11 features.

For training models 1, 2 and 3, we use grid search and 5-fold cross validation to find the best combination of model parameters. As for MKL, parameter selection is a little bit different. As the best parameter combination for sub-kernels of news and indicators has been found during the training of model 1 and 2, we just need to adopt the derived parameters and search the best parameters which are specific to MKL.

4 Experimental Results and Discussion

4.1 Data Sets

Parallel news articles and market prices serve as the experiment data sets.

- News articles. The news articles of year 2001 used in our experiment are bought from Caihua[4]. All the news articles are written in Traditional Chinese. Each piece of news is attached with a time stamp indicating when the news is released.

[4] http://www.finet.hk/mainsite/index.htm

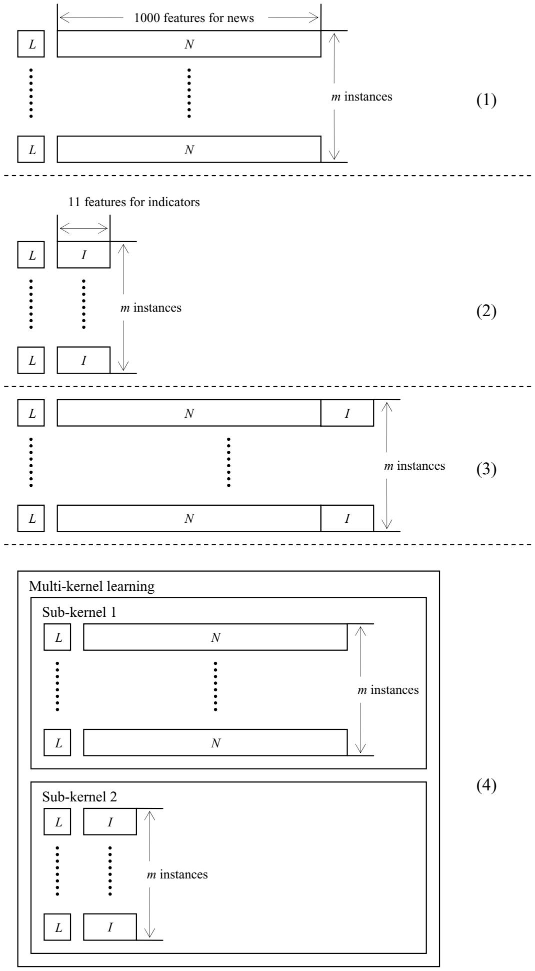

Fig. 6. Model setup: (1) News article only; (2) Ex ante prices only; (3) Naive combination; (4) MKL

- Market prices. The market prices contain all the stocks' prices of HKEx in year 2001.

Time stamps of news articles and prices are tick based.

HKEx has thousands of stocks and not all the stocks are playing actively in the market. We mainly focus on the constituents of Hang Seng Index[5] (HSI) which, according to the change log, includes 33 stocks in year 2001. However, the constituents of HSI changed twice in year 2001, which was on June 1st and July 31th. Due to the *tyranny of indexing* [28], price movement of newly added

[5] http://www.hsi.com.hk/HSI-Net/

constituent is not rational and usually will be mispriced during the first few months. We only select the constituents that had been constituents through the whole year. Thus, the number of stocks left becomes 23. The first 10-month data is used as the training/cross-validation set and the last 2-month data is used as the testing set.

4.2 Parameter Selection

During model training period, parameters are determined by two methods: grid search and 5-fold cross validation. Take model 1's training for example, SVM parameters to be tuned are γ and C. For γ, algorithm searches from 0 to 10 with step size 0.2; For C, the step size is 1 and C searches from 1 to 20. Thus, there are totally $50 \times 20 = 1000$ combinations of parameters (in other words, 1000 loops). In each loop, 5-fold cross validation is executed to validate the system's performance, which equally splits the first 10-month data into 5 parts and use 4 of them to train the model and the left 1 part to validate. Among the 1000 combinations, the one with the best performance is preserved and used to configure the final model for testing.

For models 1, 2 and 3, the method of parameter selection is the same. For model 4, instead of changing γ $50 \times 50 = 2500$ times (50 for sub-kernel of news and 50 for sub-kernel of indicator), we just adopt the γs which have already been selected in models 1 and 2's training. MKL's parameter C is selected by the same method as the other model.

4.3 Experimental Results

accuracy, which is adopted by many previous works [3,4,5,6], is used to evaluate the predication performance. The formula of accuracy is

$$accuracy = \frac{true\_positive + true\_negative}{true\_positive + false\_positive + true\_negative + false\_negative}$$

Cross validation results are listed in Table 3, and Table 4 lists the results of independent testing (numbers in bold font indicate the best results at that time point and the second best results are underlined). From the results, we can see that:

Table 3. Prediction accuracy of 5-fold cross validation (%)

Cross validation	5m	10m	15m	20m	25m	30m
Indicator	**60.24**	58.42	57.59	58.48	57.99	57.2
News	59.12	62.65	62.84	63.06	60.93	59.12
Naive combination	60.05	61.78	62.84	62.84	60.85	58.04
MKL	60.20	**62.94**	**63.68**	**64.23**	**61.14**	**60.44**

Table 4. Prediction accuracy of independent testing (%)

Independent testing	5m	10m	15m	20m	25m	30m
Indicator	53.48	50.23	47.84	48.92	45.38	50.80
News	52.94	51.13	44.40	52.38	**53.41**	48.80
Naive combination	56.15	**61.54**	49.14	52.38	45.78	49.60
MKL	**56.68**	57.84	**53.20**	**53.87**	50.34	**52.29**

1. MKL outperforms the other three models both in cross validation and independent testing, except for point 5m in cross validation and points 10m, 25m in testing. Although both Naive combination approach and MKL make use of market news and prices, naive combination does not outperform single information source based model as expected. The reason might be that simply combining the features of news and the features of indicator could lead to feature bias. As stated in Section 3.6, the number of news features has nearly 100 times than that of indicator. Features are thus greatly bias to the side of news. As can be observed in Figure 7 (I), the curve of *Naive* is quite close to *News*. On the other hand, due to a different learning approach, MKL balances the *predictability* of news and prices (Market news and stock prices have their own characteristics and the information hidden in either of them could be a complement to the other.). Comparing to cross validation, although MKL's performance in independent testing decreases, it still can achieve 4 best-results and 2 second-best-results.

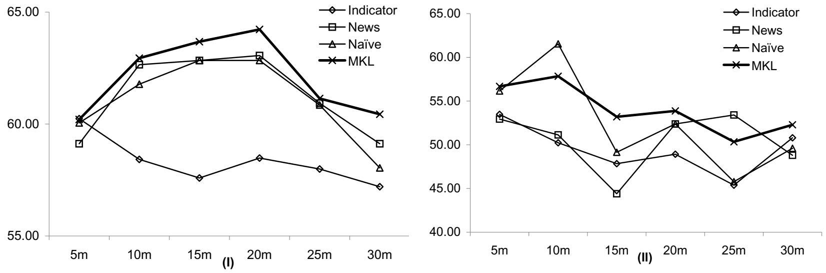

Fig. 7. Experimental results: (I) cross validation results, (II) independent testing results

2. From Figure 7 (I) and (II), it can be clearly observed that the slope of curve *Indicator* is almost negative, which means the predictability of prices decrease as time goes by. This observation is natural and consistent to the general knowledge that the impact of market information will be gradually absorbed by the market and the predictability will decrease as time goes by.
3. From Figure 7 (I), accuracy score at point 20m for curves *News*, *Naive* and *MKL*, all of which use the information of news articles, is better than

the other points, which means predictability of news articles reaches its peak at point 20m. This observation is consistent with the observation of Gidofalvi [24].

5 Conclusion and Future Work

In this paper, we build up a system which uses multi-kernel learning to integrate market news and stock prices to improve prediction accuracy for stock market. Experiment is conducted by using a whole year Hong Kong stock market tick data. Results have shown that multi-kernel based model could make better use of information in news articles and history prices than the model simply combines features of news articles and prices. It is also observed that multi-kernel model outperforms the models that just adopt one information source.

For future research on this topic, it is possible to further investigate this problem from two ways.

- Some news articles are usually not just talking about one specific stock but several stocks of the same industry. Instead of focusing on the constituents of HSI, industry section, which is at a higher level than individual stocks, could be a good research object.
- MKL in our system uses two information sources. More information sources could be found and added to this system. What kind of information sources could provide complementary information without redundancy is another question that is worth consideration.

Acknowledgements. The work in this paper is partially supported by the Research Grants Council of Hong Kong under Project No. RGC CityU 112909, the Natural Science Foundation of Hubei Province of China (No. 2010CDB08504), and Research Project of Shanghai Key Laboratory of Intelligent Information Processing under Grant No. IIPL-2010-007. Xiaotie Deng would like to acknowledge the support of a research grant from University of Liverpool.

References

1. Fama, E.F.: The behavior of stock market prices. Journal of business 38(1) (1964)
2. Barberis, N., Thaler, R.: A survey of behavioral finance. Handbook of the Economics of Finance 1, 1053–1128 (2003)
3. Fung, G., Yu, J., Lam, W.: News sensitive stock trend prediction. Advances in Knowledge Discovery and Data Mining, 481–493 (2002)
4. Fung, G.P.C., Yu, J.X., Lu, H.: The predicting power of textual information on financial markets. IEEE Intelligent Informatics Bulletin 5(1), 1–10 (2005)
5. Wu, D., Fung, G., Yu, J., Liu, Z.: Integrating Multiple Data Sources for Stock Prediction. In: Bailey, J., Maier, D., Schewe, K.-D., Thalheim, B., Wang, X.S. (eds.) WISE 2008. LNCS, vol. 5175, pp. 77–89. Springer, Heidelberg (2008)

6. Wu, D., Fung, G.P.C., Yu, J.X., Pan, Q.: Stock prediction: an event-driven approach based on bursty keywords. Frontiers of Computer Science in China 3(2), 145–157 (2009)
7. Ederington, L.H., Lee, J.H.: How markets process information: News releases and volatility. Journal of Finance 48(4), 1161–1191 (1993)
8. Engle, R.F., Ng, V.K.: Measuring and testing the impact of news on volatility. Journal of finance 48(5), 1749–1778 (1993)
9. Seo, Y.W., Giampapa, J., Sycara, K.: Financial news analysis for intelligent portfolio management. Robotics Institute, Carnegie Mellon University (2004)
10. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS) 27(2), 12 (2009)
11. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1), 59–69 (1982)
12. Kohonen, T., Somervuo, P.: Self-organizing maps of symbol strings. Neurocomputing 21(1-3), 19–30 (1998)
13. Ultsch, A.: Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. Kohonen Maps 46 (1999)
14. Fu, T., Chung, F.L., Ng, V., Luk, R.: Pattern discovery from stock time series using self-organizing maps. In: Workshop Notes of KDD2001 Workshop on Temporal Data Mining, pp. 26–29 (2001)
15. Smyth, P.J.: Hidden Markov models for fault detection in dynamic systems (November 7, 1995)
16. Pavlidis, T., Horowitz, S.L.: Segmentation of plane curves. IEEE Transactions on Computers 100(23), 860–870 (1974)
17. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
18. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of COLING, vol. 4, pp. 1367–1373 (2004)
19. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. In: ICWSM 2007 (2007)
20. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
21. Schumaker, R.P., Chen, H.: A quantitative stock prediction system based on financial news. Information Processing & Management 45(5), 571–583 (2009)
22. Feldman, R., Sanger, J.: The text mining handbook (2007)
23. Dacorogna, M.M.: An introduction to high-frequency finance (2001)
24. Gidófalvi, G., Elkan, C.: Using news articles to predict stock price movements. In: Department of Computer Science and Engineering. University of California, San Diego (2001)
25. Tay, F.E.H., Cao, L.: Application of support vector machines in financial time series forecasting. Omega 29(4), 309–317 (2001)
26. Cao, L.J., Tay, F.E.H.: Support vector machine with adaptive parameters in financial time series forecasting. IEEE Transactions on Neural Networks 14(6), 1506–1518 (2004)
27. Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., Bona, F., Binder, A., Gehl, C., Franc, V.: The SHOGUN machine learning toolbox. The Journal of Machine Learning Research 99, 1799–1802 (2010)
28. Ritter, J.R.: Behavioral finance. Pacific-Basin Finance Journal 11(4), 429–437 (2003)