Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2010 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2010

An Effective Clustering Approach to Stock Market Prediction

Anthony J.T. Lee
National Taiwan University, jtlee@ntu.edu.tw

Ming-Chih Lin
National Taiwan University, d97725004@ntu.edu.tw

Rung-Tai Kao
National Taiwan University, r96725023@ntu.edu.tw

Kuo-Tay Chen
National Taiwan University, ktchen@management.ntu.edu.tw

AN EFFECTIVE CLUSTERING APPROACH TO STOCK MARKET PREDICTION

Anthony J.T. Lee, Department of Information Management, National Taiwan University, Taipei, Taiwan, ROC, jtlee@ntu.edu.tw

Ming-Chih Lin, Department of Information Management, National Taiwan University, Taipei, Taiwan, ROC, d97725004@ntu.edu.tw

Rung-Tai Kao, Department of Information Management, National Taiwan University, Taipei, Taiwan, ROC, r96725023@ntu.edu.tw

Kuo-Tay Chen, Department of Accounting, National Taiwan University, Taipei, Taiwan, ROC, ktchen@management.ntu.edu.tw

Abstract

In this paper, we propose an effective clustering method, HRK (Hierarchical agglomerative and Recursive K-means clustering), to predict the short-term stock price movements after the release of financial reports. The proposed method consists of three phases. First, we convert each financial report into a feature vector and use the hierarchical agglomerative clustering method to divide the converted feature vectors into clusters. Second, for each cluster, we recursively apply the K-means clustering method to partition each cluster into sub-clusters so that most feature vectors in each sub-cluster belong to the same class. Then, for each sub-cluster, we choose its centroid as the representative feature vector. Finally, we employ the representative feature vectors to predict the stock price movements. The experimental results show the proposed method outperforms SVM in terms of accuracy and average profits.

Keywords: stock price prediction, financial report, document clustering.

1 INTRODUCTION

Nowadays, a large amount of information is available for investment and research analysis. Researchers and investors can easily get access to such valuable information through various channels on the Internet. For example, a company's financial report, which provides accounting items and financial ratios, is an important indicator of financial performance. More importantly, within stock market research, it is believed that the information from quarterly reports and annual reports can influence the price of a stock, especially for unexpected earnings or unexpected loss surprises (Magnusson et al. 2005). The information is likely to be in different formats, which are numeric and textual data. In particular, a company's quarterly and annual reports are good examples of documents that contain both formats (Kloptchenko et al. 2004).

Stock market prediction is an appealing topic not only for research but also for commercial applications. In stock market research, the random walk theory (Malkiel 1973) suggested that short-term stock price movements were governed by the random walk hypothesis and thus were unpredictable. On the other hand, the efficient market hypothesis (Fama 1964) stated that the stock price was a reflection of complete market information and the market behaved efficiently so that instantaneous price corrections to equilibrium would make stock prediction useless. However, prior researches (Brown & Jennings 1989; Abarbanell & Bushee 1998) made use of a variety of methods to gain future price information. They proposed two types of stock market analysis. First, the fundamental analysis derives stock price movements from financial ratios, earnings, and management effectiveness. Second, the technical analysis identifies the trends of stock prices and trading volumes based on historical prices and volumes.

Stock market prediction based on structured data such as price, trading volume and accounting items has been widely employed on numerous researches (Chan et al. 2002; Lin et al. 2009). However, it is much more difficult to predict stock price movements based on unstructured textual data. One kind of unstructured textual data for stock market prediction is collected from financial news published on the newspapers or Internet. The methods used news articles to predict stock prices in a short period after the release of news articles (Schumaker & Chen 2009). Another kind of unstructured textual data is gathered from financial reports, which contain not only textual data but also numerical data. The numerical data provides quantitative information and the textual data contains a large amount of qualitative information related to the company performance and future financial movements. Moreover, incorporating the quantitative and qualitative information into stock market analysis can improve the prediction ability (Chen et al. 2009; Kogan et al. 2009). Thus, we propose a method and use both quantitative and qualitative information in financial reports to predict stock price movements.

The K-means clustering method (K-means for short) is a widely-used clustering method. However, its major disadvantages can be described in two aspects. First, the number of clusters is often unknown in different datasets but it is required to be specified in advance. Second, randomly choosing initial centroids of the clusters makes it impossible to obtain reliable results. On the other hand, HAC (Hierarchical Agglomerative Clustering method) produces better resultant clusters and provides a more interpretative hierarchical understanding of the document collection (Steinbach et al. 2000). However, as the size of a cluster grows, the centroid of a cluster might no longer be adequate to represent any feature vectors in the cluster. This drawback makes further investigation into the characteristics of the clusters difficult. Numerous hybrid methods have been made to mitigate the disadvantages in both approaches. Cheu et al. (2004) combined the K-means, HAC or SOM (Self-Organizing Maps) for the two-level clustering. In the first level of clustering, the prototypes of vectors are generated to reduce the number of samples for the second level of clustering. Chen et al. (2005) and Hu et al. (2007) presented a hybrid clustering method by using HAC to divide the data into clusters and then using K-means to group the clusters generated by HAC. Han et al. (2009) proposed the parameter-free hybrid clustering algorithm, which uses HAC to generate initial clustering and then iteratively uses K-means to choose the best number of centroids.

Therefore, in this paper, we propose an effective clustering method, which combines the advantages of K-means and HAC, to perform stock market prediction. Unlike the previous hybrid clustering methods, we first utilize HAC to do the initial clustering and then *recursively* perform K-means to do the second clustering. The proposed method consists of three phases. First, we convert each financial report into a feature vector and use HAC to divide them into clusters. Second, for each cluster, we recursively apply K-means to partition each cluster into sub-clusters so that most feature vectors in each sub-cluster belong to the same class. Then, for each sub-cluster, we choose its centroid as the representative feature vector. Finally, we employ the representative feature vectors to predict the stock price movements.

The contributions of this paper are listed as follows. First, we use a weight to consolidate both qualitative and quantitative features to analyze financial reports. Second, we combine the advantages of the K-means and HAC methods to develop an effective clustering method to cluster financial reports and select the representative feature vectors. Third, we employ the proposed method to investigate the relationships between financial reports and short-term stock price movements. Finally, the experimental results show the proposed method outperforms SVM (Support Vector Machine) in terms of accuracy and average profits.

2 LITERATURE REVIEW

The methods used unstructured textual data to predict stock prices or market indices have to extract relevant information from a large number of text documents. LeBaron et al. (1999) suggested that the relationships between news articles and stock prices do exist. They developed a stock trading system with simulated traders and discovered a lag between the release of information and the price movements. Lavrenko et al. (2000) employed naïve Bayes and language model to predict forthcoming trends in stock price. Schumaker & Chen (2009) employed SVM to predict stock prices at the time of news release and showed that their model containing both article terms and stock price had the best performance on predicting the stock prices of twenty minutes later.

Public companies are required to file periodic financial reports through the EDGAR database pursuant to section 13 or 15(d) of the Securities Exchange Act of 1934. Thus, the financial reports are important data sources for stock market prediction. Many methods used the numerical information of the financial reports to predict stock price movements (Carnes 2006; Chen & Zhang 2007). Besides, Kloptchenko et al. (2004) suggested that the textual information in the financial reports contains not only the description of events, but also explains why they have happened and how long the effect of such events will continue. Chen et al. (2009) built an earning prediction model by incorporating the textual information about the risk sentiment contained in financial reports, which significantly improved the accuracy of earning prediction. Moreover, the textual information holds some forward-looking statements about the future performance of the company. Exploiting the related textual information in addition to the numeric information should increase the quality of prediction.

Back et al. (2001) used SOMs to cluster the companies based on the quantitative and qualitative information in the annual reports. They compared the resultant clusters and suggested that the performance of considering both quantitative and qualitative information is better than that of using just quantitative or qualitative information. Kloptchenko et al. (2004) combined SOMs and prototype-matching methods to analyze the quantitative and qualitative information of quarterly reports. The experimental results suggested that the quantitative part reflects the past financial performance, but the qualitative part holds some messages about the future performance of the companies. Magnusson et al. (2005) analyzed the effects of seven financial ratios by SOMs and the effects of the qualitative data by collocational networks (Williams 1998). They concluded that: (1) a change in the textual data usually indicates a change in the financial data of the following quarter; and (2) the relationship is a consequence of the fact that the texts reflect the plans and future expectations, whereas the ratios reflect the current financial situation of the company.

Many stock prediction methods based on SVM have been proposed (Qiu et al. 2006; Schumaker & Chen 2009). Qiu et al. (2006) built SVM-based predictive models with different feature selection methods from ten years of annual reports. The results showed that document frequency threshold is efficient in reducing feature space while maintaining the same classification accuracy compared with other feature selection methods. Furthermore, the results showed the feasibility of using text classification on current year's annual reports to predict next year's company financial performance, namely the return on equity ratio.

It has been shown that the performance of considering both quantitative and qualitative information is better than that of using just quantitative or qualitative information. However, quantitative and qualitative information of financial reports are considered separately in the previous studies (Back et al. 2001; Kloptchenko et al. 2004; Magnusson et al. 2005). In this paper, we use a weight to combine both qualitative and quantitative information together and propose an effective clustering method to predict the stock price movements.

3 PROPOSED FRAMEWORK

We first extract a feature vector for each financial report. Each feature vector comprises two parts, namely qualitative and quantitative. The qualitative part is extracted from the textual contents of the financial reports. To obtain the qualitative part, we first transform financial reports into bag of words by the stemming algorithm (Porter 1980) and removing stop words. Then, we compute the TF-IDF weight of each term by multiplying the term frequency and the inverse document frequency. The term frequency $tf_{t,d}$ represents the number of occurrences of term t in the financial report d. The inverse document frequency idf_t is defined as $\log_2(n/df_t)$, where n is the total number of financial reports in the collection, and df_t is the number of financial reports containing term t in the collection. We select the terms with top k TF-IDF weights to form the qualitative part of a feature vector.

In addition, the quantitative part of a feature vector comprises some ratios about the performance of the company. Based on the prior research (Magnusson et al. 2005), we select five important financial ratios regarding company performance, namely operating margin, return on equity (ROE), return on total assets (ROTA), equity to capital, and receivables turnover. Incorporating the qualitative information with the quantitative information of the financial reports may generate more valuable information to explain the stock price dynamics.

Thus, each feature vector contains k qualitative features and five quantitative features. The similarity between two feature vector, f_1 and f_2, is defined by α times the Euclidean distance of qualitative features plus 1-α times the Euclidean distance of quantitative features of f_1 and f_2, where the combination weight α is used to measure the relative importance of qualitative and quantitative features.

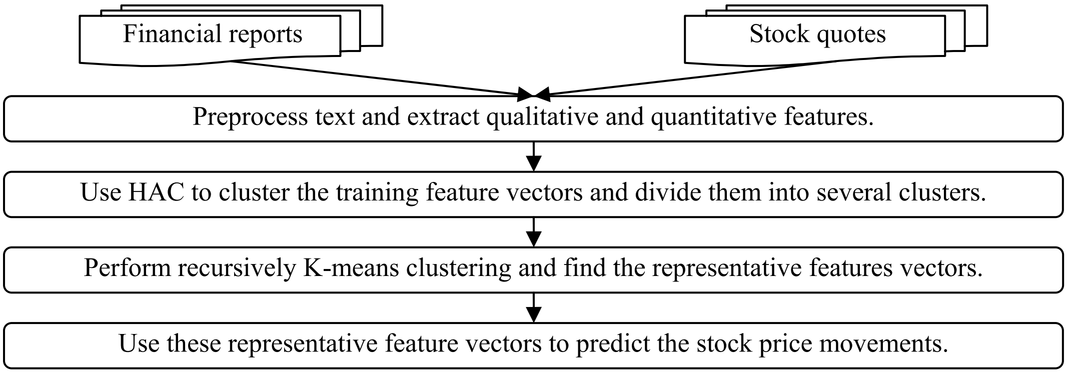

Financial reports
Stock quotes
Preprocess text and extract qualitative and quantitative features.
Use HAC to cluster the training feature vectors and divide them into several clusters.
Perform recursively K-means clustering and find the representative features vectors.
Use these representative feature vectors to predict the stock price movements.

Figure 1. The proposed framework for stock market prediction.

To distinguish the influence of financial reports on the direction of stock price movements, we classify the financial reports into three categories: "rise", "no movement", and "drop", which are represented by 1, 0, and -1, respectively. Specifically, we follow the categorization scheme used in Mittermayer (2004). We define the time window for a financial report from the release day to one trading day after the release. Then, we label a financial report as "rise" if it leads to a peak, with an increase of at least 3% and triggers a shift of average price at least 2% above the open price of the release day during the defined time window. Similarly, we label a financial report as "drop" if it leads to a drop, with a decrease of at least 3% and triggers a shift of average price at least 2% below the open price of the release day during the defined time window.

Next, we propose an effective clustering method, HRK (Hierarchical agglomerative and Recursive K-means clustering), for stock market prediction as shown in Figure 1. The proposed method consists of three phases. First, we apply HAC to cluster the training feature vectors and divide them into clusters. Second, from the clusters generated by HAC, we recursively perform K-means to accomplish further clustering until the purity of the cluster exceeds a predefined purity threshold *p*, where the purity is defined as the number of feature vectors of the dominant class divided by the total number of feature vectors in the cluster. Then, we compute the centroid for each cluster. The centroids are called the representative feature vectors of the clusters. Finally, we use these representative feature vectors to predict the stock price movements.

3.1 Hierarchical Agglomerative Clustering

First, we perform HAC to do initial clustering and construct a dendrogram, where the centrioid clustering is used and the similarity is computed by the Euclidean distance between feature vectors.

The clustering process of HAC is described as follows. Let us consider a document collection consist of nine financial reports $\{d_1, d_2, \ldots, d_9\}$, where the incidence matrix is shown in Table 1. The feature vector of the financial report d_i is illustrated in the *i*th column. The last five values are the quantitative features. After applying HAC, the resultant dendrogram is shown in Figure 2, where each financial report is represented by a node, and two merged clusters is linked by an edge.

Financial report		d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
Qualitative features	efficient	1	1	0	1	0	0	1	0	0
	growth	1	1	0	0	0	0	0	0	0
	advantage	1	1	1	0	0	0	0	0	0
	improvement	1	0	0	0	0	0	0	0	0
	deficient	0	0	0	1	1	0	0	0	0
	reorganize	0	0	0	1	1	0	0	1	1
	difficulty	0	0	0	1	1	1	1	0	0
	complaint	0	0	0	1	0	0	0	0	1
Quantitative features	operating margin	0.4	0.38	0.37	0.1	0.07	0.08	0.05	0.06	0.03
	ROE	0.3	0.28	0.27	0.01	0.04	0.04	0.07	0.02	0.05
	ROTA	0.25	0.23	0.22	0.02	0.05	0.07	0.04	0.01	0.04
	equity to capital	0.8	0.78	0.77	0.45	0.4	0.5	0.45	0.5	0.55
	receivables turnover	2.5	2.4	2.45	1.4	1.3	1.5	1.2	1.1	1.5
Class label		1	1	0	-1	-1	-1	0	-1	-1

Table 1. An example dataset.

Next, we divide the dendrogram constructed in the above step into *s* groups. If we want to split it into *s* groups, we remove the *s*-1 longest links, where the *s*-1 longest links refer to the links that merge two clusters in the last *s*-1 iterations in HAC. The reason why we could remove the longest links is that the longest links must merge clusters which are most dissimilar. Each group forms a cluster, which will be input to the K-means clustering method. In the example shown in Figure 2, if we want to obtain three clusters after the initial clustering, we just need to remove the two longest links. Consequently, we obtain three clusters: $\{d_1, d_2, d_3\}$, $\{d_4, d_5, d_6, d_7\}$, and $\{d_8, d_9\}$.

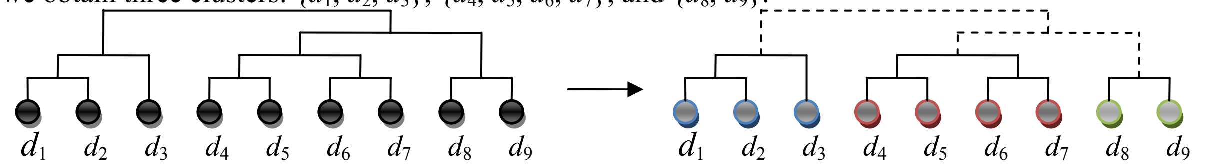


Figure 2. The dendrogram constructed by HAC and the clusters formed after removing the links.

3.2 K-means Clustering Method

We perform recursively the K-means clustering method to divide each cluster into sub-clusters until most feature vectors in each sub-cluster belong to the same class. However, to avoid the over-fitting problem, we use a purity threshold *p* in the recursive K-means clustering. When the purity of a cluster exceeds *p*, the recursion is finished. In addition, the class label of the resultant sub-cluster is set to the label of the majority class. In the proposed method, we modify the K-means clustering method in two aspects. First, the number of sub-clusters is determined by the number of different classes within a cluster. Second, the centroid of each sub-cluster is determined by averaging the features vectors belonging to the same class. We employ these two modifications to overcome the inherent weaknesses of the K-means clustering method.

For each cluster (or sub-cluster), we first examine how many different classes within the cluster (or sub-cluster), where the centroid of each class is determined by averaging the feature vectors which belong to the class. For example, there are two classes within the cluster $\{d_1, d_2, d_3\}$, namely class 0 and class 1. Thus, the number of sub-clusters in the K-means clustering method is set to 2. The centroid of class 0 is (0, 0, 1, 0, 0, 0, 0, 0, 0.37, 0.27, 0.22, 0.77, 2.45), which is the average of the feature vectors of d_1 and d_2, and the centroid of class 1 is (1, 1, 1, 0.5, 0, 0, 0, 0, 0.39, 0.29, 0.24, 0.79, 2.45). That is, the cluster $\{d_1, d_2, d_3\}$ is further divided into two sub-clusters: $\{d_1, d_2\}$, and $\{d_3\}$. The purity of each cluster obtained is 1.0. Thus, the recursion is finished.

Next, let us consider the cluster $\{d_4, d_5, d_6, d_7\}$. After the first iteration of the K-means clustering method, the cluster is divided into two sub-clusters: $\{d_4, d_5\}$, and $\{d_6, d_7\}$. However, there are two classes within the sub-cluster $\{d_6, d_7\}$. Thus, the sub-cluster is further divided into two sub-clusters: $\{d_6\}$, and $\{d_7\}$. Since the purity of each cluster obtained is 1.0, the recursion is finished. Moreover, there is only one class in the cluster $\{d_8, d_9\}$, and thus we don't need to perform the K-means clustering method. Finally, we obtain six clusters: $\{d_1, d_2\}$, $\{d_3\}$, $\{d_4, d_5\}$, $\{d_6\}$, $\{d_7\}$, and $\{d_8, d_9\}$.

For each resultant sub-cluster, its centroid is computed by averaging the feature vectors within the sub-cluster. These centroids are regarded as the representative feature vectors of the resultant sub-clusters, which is used to predict the stock price movements.

3.3 Stock Price Movements Prediction

When a financial report is released, we will transform it into a feature vector *f* according to the steps described in Section 3. Next, we assign *f* to the nearest representative feature vector. Then, we predict the direction of the stock price movement according to the class label of the nearest representative feature vector. For example, if the transformed feature vector *f* is assigned to the representative feature vector of cluster $\{d_1, d_2\}$, we predict the direction of the stock price movement to be "rise". Hence, we make a buy stock decision based on the prediction. On the other hand, if the prediction is "drop", we make a short stock decision. We don't make any trading decision if the prediction is "no movement".

4 ANALYSIS

We conducted the experiments to compare HRK with SVM. HRK was implemented by Microsoft Visual C++ 2008 and SVM was implemented by LIBSVM (Chang & Lin 2001). We chose the polynomial kernel and set all its other parameters to their default values since the polynomial kernel outperformed the others for the dataset. All the experiments were performed on an IBM Compatible PC with Intel Pentium 4 @ 3.40GHz, 2.0GB main memory, running on Windows XP Professional.

4.1 Dataset and Evaluation Metrics

We gathered financial reports and financial ratios from the EDGAR database. We focused on the companies listed in the S&P 500 index as of Sep. 30, 2008, and collected all available quarterly and annual reports released from Jan. 1, 1995 to Dec. 31, 2008. Besides, the daily open and close stock quotes were gathered. We also conducted the GICS (Global Industrial Classification System) experiments to investigate the performance of company groups based on their industry sectors, where the GICS was developed by Morgan Stanley in 1999. Therefore, we classified the companies into ten industry sectors according to the definition of their principal business activity. The codes and corresponding industry sectors are described in Table 2. In the experiments, we used the financial reports before Jan. 1, 2006 as the training reports. The remaining financial reports were testing reports. There are 20,884 training reports and 5,371 testing reports. In the GICS experiments, the numbers of training and testing reports are shown in Table 2.

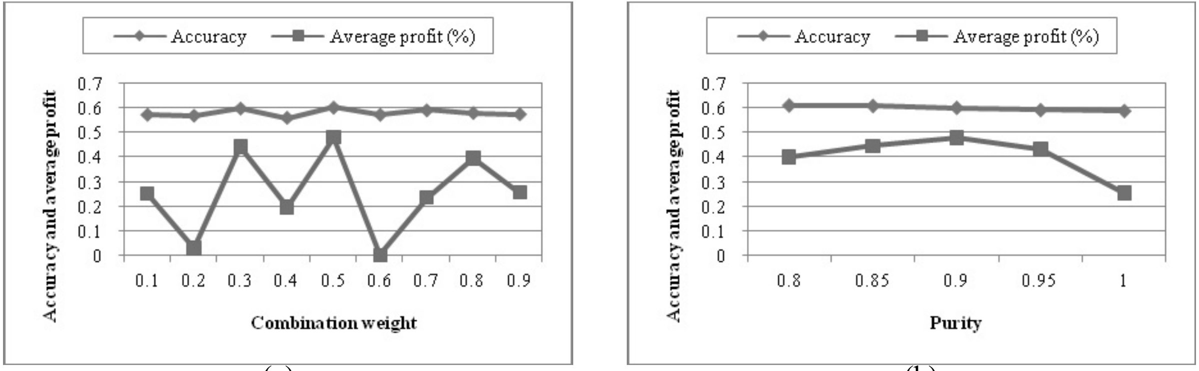
Code	Industry sector	Number of training reports	Number of testing reports
10	Energy	1,710	442
15	Materials	1,226	329
20	Industrials	2,688	651
25	Consumer discretionary	3,519	887
30	Consumer staples	1,890	442
35	Health care	2,361	585
40	Financials	2,684	743
45	Information technology	3,118	831
50	Telecommunication services	355	100
55	Utilities	1,333	361

Table 2. The GICS dataset.

We use two matrices to evaluate the performance in the experiments. One is the accuracy of the prediction. The other is the average profit per trade, which simulates the buy and short trading based on the predictions in the short-term stock market. If the prediction is “rise” (or “drop”), we make a buy (or short) decision at the open of the day of the financial report releases and even up at the close of the next trading day. Based on the prior research (Lavrenko et al. 2000; Schumaker & Chen 2009), we assume the transaction cost is zero since the trading costs are absorbed if the trading volume is large. The average profit per trade is calculated by averaging the profit rate of each trade.

4.2 Experimental Results

To decide the value of each parameter, we randomly sampled 10% of the data from each industry sector to conduct a series of experiments and found that HRK have the best performance when the number of qualitative features is 1,000 and the number of clusters generated by HAC is 10. Then, we used the rest data of each industry sector to evaluate the performance of HRK and SVM. Figure 3(a) shows the accuracy and average profit versus the combination weight, where the purity is 0.9. The experimental result shows that we have the highest average profits when the weight is set to 0.5. Moreover, Figure 3(b) illustrates the accuracy and average profit versus the purity, where the purity is from 0.8 to 1.0. The experimental result shows that HRK is most profitable when the purity is 0.9. Hence, we set the purity to 0.9 in the following experiments. Note that the accuracy decreases slightly and the average profit increases sharply when the purity varies from 0.8 to 0.9. When the purity threshold is low, the feature vectors of class 0 dominate some clusters. Hence, the feature vectors of class -1 and class 1 in these clusters would be merged into class 0. That makes the prediction bias toward class 0. Therefore, the average profit is low since fewer trades are executed. On the other hand, the accuracy decreases slightly and the average profit decreases sharply when the purity varies from 0.9 to 1. When the purity threshold is high, the resultant clustering becomes over-fitted. Therefore, the accuracy and average profit are lower.



(a) (b)

Figure 3. The accuracy and average profit: (a) combination weight and (b) purity.

Next, we compare HRK with SVM, HAC, and K-means methods, where the combination weight is set to 0.5 and the purity is set to 0.9 in HRK. The experimental results are shown in Figure 4. In this experiment, we adopt two settings of K-means clustering, namely K-means (avg_seed) and K-means (rand_seed). The difference between them is in the process of seed initialization. The seeds of K-means (avg_seed) are calculated as the average of the feature vectors of each class within a cluster, while the seeds of K-means (rand_seed) are randomly selected among the feature vectors within a cluster. Note that both of them are recursively performed until the purity of each cluster exceeds the purity threshold. Besides, we adopt three settings of HRK: HRK (with ratio) includes 1,000 qualitative features retrieved from financial reports and five financial ratios, HRK (w/o ratio) excludes the financial ratios, and HRK (ratio) only includes the financial ratios in the feature vectors.

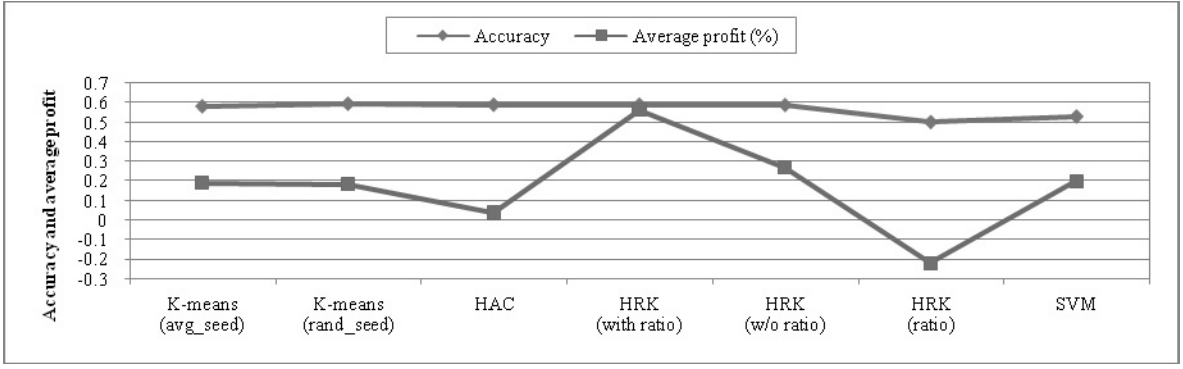


Figure 4. Comparing HRK with the K-means, HAC, and SVM methods.

By comparing two settings of the K-means clustering, we find that K-means (avg_seed) has better average profit. That is, initializing the seeds as the average of the feature vectors of each class within a cluster contributes to the better quality of the clustering. By comparing three settings of HRK, we could confirm that the performance of considering both qualitative and quantitative features in financial reports is better than that of only considering the qualitative or quantitative features. Moreover, HRK (with ratio) outperforms K-means (avg_seed). Since HRK uses HAC to divide the feature vectors into several clusters and HAC localizes the resultant clusters, the average profit is better than K-means (avg_seed). Besides, HRK (with ratio) outperforms HAC method as well. The results show that HRK combines the advantages of two clustering methods and the performance is better than that of using K-means clustering or HAC method only. Furthermore, HRK (with ratio) performs better than SVM in terms of accuracy and average profits.

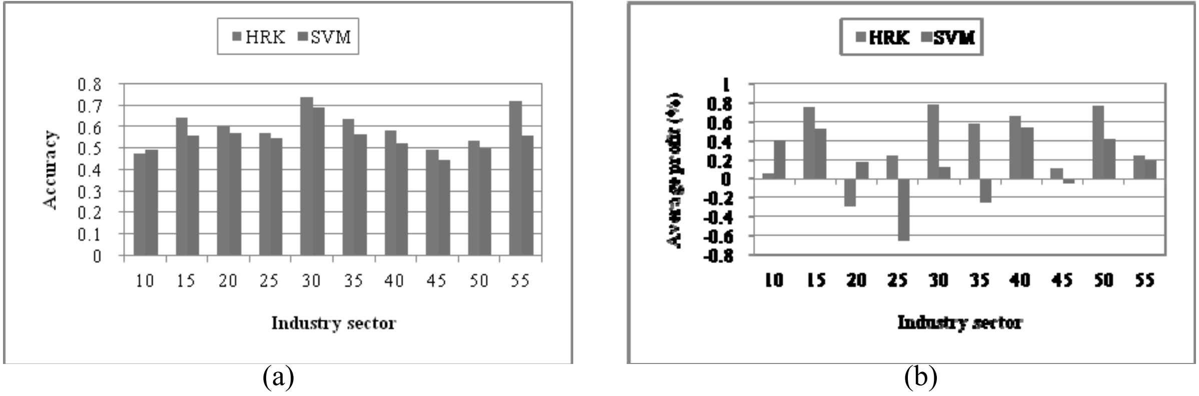


Figure 5. The (a) accuracy and (b) average profit of the GICS experiment.

Figure 5 shows the accuracy and average profit of the GICS experiment. For the accuracy, HRK outperforms SVM in nine industry sectors. By employing paired t-test over the results at 95% confidence level, the results show HRK performs significantly better than SVM with p-value 0.0027. For the average profit, HRK outperforms SVM in eight industry sectors. Furthermore, the total average profit of 10 industry sectors of HRK is 3.95%, while the total average profit of SVM is 1.46%. The results of the GICS experiment further validate that HRK is better than SVM.

In summary, HRK outperforms SVM in terms of accuracy and average profit. HRK can attribute its better performance to three aspects. First, we consider both qualitative and quantitative features in financial reports. Second, we combine the advantages of two clustering methods to propose an effective clustering method. Third, choosing an appropriate number of splits in HAC can localize the clusters generated and thus improve the quality of the clustering generated by the K-means clustering.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed an effective method, HRK, to predict the short-term stock price movements after the release of financial reports. We combine the advantages of HAC and K-means clustering methods to propose a hybrid clustering method. The experimental results show that HRK outperforms SVM. Besides, the performance of considering both qualitative and quantitative features in financial reports is better than that of only considering the qualitative or quantitative features.

We have focused our research on financial reports dataset to predict the short-term stock price movements after the release of financial reports. In addition to financial reports, the proposed method may also be applied to predict the stock price movements on financial news articles immediately after the article release. Besides, we may also consider incorporating more financial ratios, accounting items, and technical indicators into quantitative features in the future. Prior researches (Mittermayer 2004) suggested that integrating with domain knowledge is effective in extracting textual information. It is worthy of consulting with domain experts to find the keywords which may influence the stock price movements. On the other hand, we will investigate the effect of using industry specific feature set for each industry sector instead of a global feature set. Finally, in the phase of predicting stock price movements, it will be worthwhile using the representative feature vectors to build a classification model such as the decision tree classification model in the future.

Acknowledgements

The authors are grateful to the anonymous referees for their helpful comments and suggestions. This research was supported in part by the National Science Council of Republic of China under Grant No. NSC 97-2410-H-002-126-MY3.

References

Abarbanell, J.S., Bushee, B.J. (1998). Abnormal returns to a fundamental analysis strategy. The Accounting Review, 73, 19-45.

Back, B., Toivonenb, J., Vanharanta, H., Visa, A. (2001). Comparing numerical data and text information from annual reports using self-organizing maps. International Journal of Accounting Information Systems, 2, 249-269.

Brown, D.P., Jennings, R.H. (1989). On technical analysis. The Review of Financial Studies, 2 (4), 527-551.

Carnes, T.A. (2006). Unexpected changes in quarterly financial-statement line items and their relationship to stock prices. Academy of Accounting and Financial Studies Journal, 10 (3).

Chan, M.C., Wong, C.C., Tse, W.F., Cheung, B., Tang, G. (2002). Artificial intelligence in portfolio management. Intelligent Data Engineering and Automated Learning, 403-409.

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

Chen, B., Tai, P.C., Harrison, R., Pan, Y. (2005). Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis. IEEE Computational Systems Bioinformatics Conference, 105-108.

Chen, K.T., Chen, T.J., Yen, J.C. (2009). Predicting future earnings change using numeric and textual information in financial reports. In Proceedings of Pacific Asia Conference on Knowledge Discovery and Data Mining, 54-63.

Chen, P., Zhang, G. (2007). How do accounting variables explain stock price movements? Theory and evidence. Journal of Accounting and Economics, 43, 219-244.

Cheu, E.Y., Kwoh, C.K., Zhou, Z. (2004). On the two-level hybrid clustering algorithm. International Conference on Artificial Intelligence in Science and Technology, 138-142.

Fama, E.F. (1964). The behavior of stock market prices. Journal of Business, 38 (1), 34-106.

Han, Z.X., Feng, S., Ye, Y., Jiang, Q. (2009). A parameter-free hybrid clustering algorithm used for malware categorization. In Proceedings of the 3rd International Conference on Anti-Counterfeiting, Security, and Identification in Communication, 480-483.

Hu, J., Ray, B.K., Singh, M. (2007). Statistical methods for automated generation of service engagement staffing plans. IBM Journal of Research and Development, 51 (3), 281-293.

Kloptchenko, A., Eklund, T., Back, B., Karlsson, J., Vanharanta, H., Visa, A. (2004). Combining data and text mining techniques for analyzing financial reports. Intelligent Systems in Accounting, Finance and Management, 12 (1), 29-41.

Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., Smith, N.A. (2009). Predicting risk from financial reports with regression. In Proceedings of NAACL Human Language Technologies Conference.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J. (2000). Mining of concurrent text and time series. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 37-44.

LeBaron, B., Arthur, W.B., Palmer, R. (1999). Time series properties of an artificial stock market. Journal of Economic Dynamics and Control, 23 (9-10), 1487-1516.

Lin, X., Yang, Z., Song, Y. (2009). Short-term stock price prediction based on echo state networks. Expert Systems with Applications, 36 (3), 7313-7317.

Magnusson, C., Arppe, A., Eklund, T., Back, B., Vanharanta, H., Visa, A. (2005). The language of quarterly reports as an indicator of change in the company’s financial status. Information and Management, 42, 561-574.

Malkiel, B.G. (1973). A Random Walk Down Wall Street. W. W. Norton & Company, New York.

Mittermayer M.A. (2004). Forecasting intraday stock price trends with text mining techniques. Proceedings of the 37th Hawaii International Conference on System Sciences.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14, 130-137.

Qiu, X.Y., Srinivasan, P., Street, N. (2006). Exploring the forecasting potential of company annual reports. In Proceedings of the American Society for Information Science and Technology.

Schumaker, R.P., Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: the AZFin text system. ACM Transactions on Information Systems, 27 (2), 1-19.

Steinbach, M., Karypis, G., Kumar, V. (2000). A comparison of document clustering techniques. In Proceedings of KDD Workshop on Text Mining.

Williams, G.C. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. International Journal of Corpus Linguistics, 3, 151-171.