Competitive Benchmarking in Large Data Bases Using Self-Organizing Maps

Lee, Young Chan
Institute for Business Research, Sogang University

이 영 찬
서강대학교 경영학연구원 경영연구소

ABSTRACT

The amount of financial information in today's sophisticated large data bases is huge and makes comparisons between company performance difficult or at least very time consuming. The purpose of this paper is to investigate whether neural networks in the form of self-organizing maps can be used to manage the complexity in large data bases. This paper structures and analyzes accounting numbers in a large data base over several time periods. By using self-organizing maps, we overcome the problems associated with finding the appropriate underlying distribution and the functional form of the underlying data in the structuring task that is often encountered, for example, when using cluster analysis. The method chosen also offers a way of visualizing the results. The data base in this study consists of annual reports of more than 80 Korean companies with data from the year 1998.

Keywords: Neural networks, self-organizing maps, data bases, benchmarking

1. Introduction

Competitive benchmarking is an important company-internal process, in which the functions and performance of one company are compared with those of other companies. Financial competitive benchmarking uses financial information most often in the form of ratios to perform these comparisons. Financial competitive benchmarking is utilized, among other things, as a communication tool in strategic management, for example in situations where company management must gain approval, from internal and external interest groups alike, for new functional objectives for the company.

Multivariate statistical methods have been used as a tool of analysis for company performance, bankruptcy predictions, stock market predictions etc., although mostly in research contexts. However, many problems have been reported concerning these methods. The two most important problems are the assumption on normality in the underlying distributions and difficulties in finding an appropriate functional form for the distributions. Moreover, results of analyses are difficult to visualize when there are several explanatory variables [Vermeulen et al., 1994].

Many researchers have addressed these problems: Trigeueros [1995] reports on several studies that have shown the existence of positive or negative skewedness in the ratios and on different remedies to overcome these difficulties. He also explains the existence of symmetrical and negatively skewed ratios and offers guidelines for achieving higher precision when using ratios in statistical context.

Fernandez-Castro & Smith [1994] used a non-parametric model of corporate performance to overcome the need for specification of statistical distribution or functional form. Vermeulen et al. [1994] presented a way to visualize the results with inter-firm comparison when the explanatory variable was explained by more than one firm characteristic.

Vanharanta [1995] has used modern computer technology and built a hyperknowledge-based system for financial benchmarking. The system contains a data base with financial data on more

than 160 pulp and paper companies worldwide. The amount of financial information in this system is, however, so large that it makes comparisons between companies difficult or at least very time consuming.

Artificial neural networks are a promising new paradigm in information processing. Originally, they were developed as computer analogues for the human brain [Hecht-Nielsen, 1991]. Artificial neural networks are able to learn the pattern of a system from a given set of examples, which makes them very attractive. They are applicable to such processes as classification, prediction, control, and inference [Rumelhart et al., 1986].

Back et al. [1996] investigated the potential of self-organizing maps for pre-processing 120 world wide forest companies' financial data bases and presented an approximated position of one company's financial performance compared to that of other companies. The results were promising. By using self-organizing maps they have overcome the problems associated with finding the appropriate underlying distribution and the functional form of the financial indicators. Furthermore, the visualization capabilities of self-organizing maps provide a good way of presenting and analyzing the results.

Neural networks have previously been suggested by Trigueiros [1995] for use with computerized accounting reports data bases, and by Chen et al. [1995] to define cluster structures in large data bases. Martin-del-Brio & Serrano-Cinca [1995] used self-organizing maps for analyzing the financial state of Spanish companies.

This paper use the self-organizing maps to structure the financial information on more than 80 Korean companies in Yahoo Korea financial information data base into clusters based on the underlying weight maps. Each cluster is then named according to the financial characteristics of the cluster. We analyze the financial performance of the Korean companies in these maps in the year 1998. Even though we take a closer look only at these companies, any individual company or group of companies can be the focus of interest.

We anticipate that neural networks can be used in future for benchmarking purposes to help executives find company characteristics that will lead to sustainable excellence of a company, in other words to help answer the question: Which are the characteristics that lead a company towards long-lasting good performance? Some company characteristics seem to produce and maintain good overall company performance, sustainable profitability, increasing productivity and continuous growth.

The rest of the paper is organized as follows: Section 2 describes the methodology used, the network structure, the data base, the list of companies in the study and the criteria for and the choice of financial ratios. Section 3 presents the results of applying neural networks to the problem and section 4 presents the empirical results. The conclusions of our study are presented in Section 5.

2. Methodology

2.1 Benchmarking

Competitive benchmarking is a company-internal process in which the activities of a given company are measured against the best practices of other, best-in-class companies. In the process of competitive benchmarking, internal functions are analyzed and measured using financial (i.e. quantitative) and/or non-financial (i.e. qualitative) yardsticks. Functions measured from one company are compared with similar functions measured from leading competitors, or they are compared with the best practices in other industries. The differences between compared functions are measured. The overall management goal of competitive benchmarking within a given company is to close the measured "gap" by changing the company's characteristics in ways that will improve company performance.

The generic benchmarking process consists of a planning phase, an analysis phase and an integration and action phase. The specific activity of financial competitive benchmarking is an integral part of the generic benchmarking process. In financial benchmarking, the aim is to compare the company with its competitors using available financial information, financial yardsticks. At the beginning of a benchmarking process, in its planning phase, financial benchmarking plays an important role in the identification and selection of the right competitors and/or good performers, those that will act as the benchmarks in the non-financial benchmarking to be done later in the generic process. Financial benchmarking is also important in the analysis phase when performance gaps are being measured and future performance levels projected. In the integration and action phase, financial benchmarking is useful for monitoring and tracking progress and for re-calibrating the benchmarks. Financial benchmarking achieves its greatest potential, however, as a communications tool at times when company management must gain approval, from internal and external interest groups alike, for new functional objectives for the company, i.e. in

strategic management.

The financial information needed for financial benchmarking work is, however, invariably available only from large commercial data bases or from specialized reports and publications, from where it must be gleaned with difficulty. Such information is thus far removed from its active users. If the needed financial information is to be brought closer to the active users, it must first be pre-processed, i.e. refined and classified. The overall objective of the present study is to pre-process, with the help of neural networks, the data and information needed for financial benchmarking purposes. Thus pre-processed, the information can be used in computerized benchmarking systems and executive support systems, making the task of competitive financial benchmarking easier and more effective.

2.2 Neural networks

A neural network is a computing device that is able to learn from examples. It consists of a set of simple processing units, neurons, that are connected to each other to form a network topology. A neural network compares input data with output data, and tries to approximate some complicated, unknown functionality between the two. When developing a neural network, the first step is to find a suitable topology for the network and thereafter train it so that it gradually learns the desired input/output functionality. There are two ways to train a network, *supervised* and *unsupervised*. In supervised learning the network is presented with examples of known input-output data pairs, after which it starts to mimic the presented input-output behavior. The network is then tested to see whether it is able to produce correct output, when only input is presented to it. In unsupervised learning, the output data is not available and usually not even known beforehand. Instead, the network tries to find similarities between input data samples. Similar samples form clusters that constitute the output of the network. The user is responsible for giving an interpretation to each cluster.

Since companies [in the data base] do not have predefined labels describing their financial status, a network intended for pre-processing their data can have no pre-desired outputs. For this reason, we utilize an unsupervised learning method. The Kohonen network [Kohonen, 1998], being the most common network model based on unsupervised learning, is used in this study.

The Kohonen network usually consists of two layers of neurons: an *input layer* and an *output layer*. The input layer neurons present an input pattern to each of the output neurons. The neurons in the output layer are usually arranged in a grid, and are influenced by their neighbors in this grid. The goal is to cluster the input patterns in such a way that similar patterns are represented by the same output neuron, or by one of its neighbors. Every output neuron has an associated *weight vector*. The neighborhood structure of the output layer will cause neighboring neurons in it [the output layer] to have similar weight vectors. These vectors should represent some subclass of the input patterns, thus forming a map of the input space, a *self-organizing map* (SOM).

The network topology can be described by the number of output neurons present in the network and by the way in which the output neurons are interconnected, i.e. by describing which neurons in the output array are mutual neighbors. Usually, neurons on the output layer are arranged in either a rectangular or a hexagonal grid, see Figure 1. In a rectangular grid each neuron is connected to four neighbors, except for the ones at the edge of the grid. In all the networks we use, the output neurons are arranged in a hexagonal lattice structure. This means that every neuron is connected to exactly six neighbors, except for the ones at the edge of the grid. This choice was made following the guidelines of Kohonen [Kohonen, 1995].

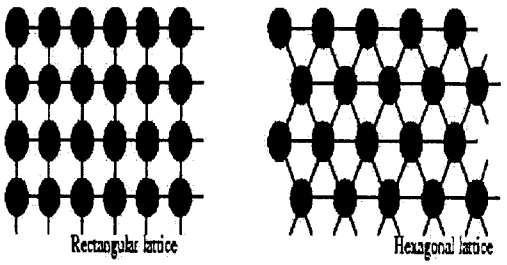


Figure 1. Network topologies

A Kohonen network is trained using unsupervised learning. During the training process the network has no knowledge of the desired outputs. The training process is characterized by a competition between the output neurons. The input patterns are presented to the network one by one, in random order. The output neurons compete for each and every pattern. The output neuron with a weight vector that is closest to the input vector is called the winner. For expressing the similarity between two vectors, we use the *Euclidean distance* between the two vectors. The weight vector of the winner is adjusted in the direction of the input vector, and so are the weight vectors of the surrounding neurons in the output array. The size of adjustment in the weight vectors of the

neighboring neurons is dependent on the distance of that neuron from the winner in the output array.

We use two learning parameters: the *learning rate* and the *neighborhood width* parameter. The learning rate influences the size of the weight vector adjustments after each training step, whereas the neighborhood width parameter determines to what extent the surrounding neurons, the neighbors, are affected by the winner. An additional parameter is the *training length*, which measures the processing time, i.e. the number of iterations through the training data.

Our criterion for the quality of a good map was the average quantization error, which is an average of the Euclidean distances of each input vector and its best matching reference vector in the SOM.

2.3 Data base and selection of companies

The Yahoo Korea financial information data base is used as the experimental financial knowledge base for the neural network tests. It provides standardized income statement, balance sheets, and cash flow statements of Korean companies. The data base also consists of more than 110 financial ratios, calculated using information from the standardized reports as well as general company information concerning products and production volumes. In this paper, 80 companies are randomly extracted from the data base. The companies are listed in Table 1.

Table 1. Companies

Company	Label	Company	Label
경남기업	F1	경남에너지	F2
극동전선공업	F5	금강피혁	F6
대덕전자	F9	대륭정밀	F10
대양금속	F13	대영포장	F14
동서산업	F17	동원수산	F18
디아이	F21	레이디가구	F22
모나미	F25	삼보컴퓨터	F26
삼애실업	F29	삼양제넥스	F30
삼환기업	F33	상림	F34
성원건설	F37	세계물산	F38
세원	F41	신광산업	F42
신영와코루	F45	신원	F46
엔케이전선	F49	엔케이텔레콤	F50
원림	F53	유성	F54
이구산업	F57	일동제약	F58
제일냉동	F61	제일모직	F62
코리아데이타시스템즈	F65	코오롱	F66
태창기업	F69	풀무원	F70
한미약품공업	F73	한섬	F74
현대엘리베이터	F77	현대자동차	F78
효성 T&C	F79	LG 전자	F80

2.4 Choice of ratios

The population consists of 52 financial ratios in the benchmarking system organized in the benchmarking system into six groups under the headings:

1) Profitability
2) Growth
3) Capital Structure
4) Activity
5) Productivity
6) Liquidity

The choice of ratios in this study was based on an empirical study conducted using MDA (multiple discriminant analysis), logistic regression, neural network, and decision tree from a large corporate bankruptcy data base [Kim, 1998]. The following nine ratios were selected. The numbers in parentheses indicate the appropriate ratio group number shown above.

- Net income to sales (1)
- Ordinary income to sales (1)
- Ordinary income to total sales (1)
- Ordinary income to stockholder's equity (1)
- Growth rate of sales (2)
- Growth rate of net income (2)
- Stockholder's equity to total assets (3)
- Stockholder's equity turnover (4)
- Cash flow/debt (6)

We note that there are four profitability measures, two growth measure, one capital structure measure, one activity measure, one liquidity measure, no productivity measures.

3. Training and testing the network

In this section we give a description of the construction process followed in developing the self-organizing maps. The actual construction work was performed using *The Self-Organizing Map Program Package version 3.1* prepared by the SOM Programming Team of the Helsinki University of Technology.

We started by standardizing the ratios in the data base using Tukey's Rank Normalization [SAS Institute Inc., 1994] in order to ease the SOM's learning process and to improve its performance.

$$y_i = \Phi^{-1}(r_i - 1/3)\ /(n + 1/3)$$

where, r_i : i^{th} rank of observations

n : # of observations after removing missing value
Φ^{-1} : inverse function of cumulative normal distribution

All the maps were trained in two phases. The purpose of the first training phase was to order the randomly initialized reference vectors of the maps to "approximately correct" values. During the second phase the maps are "fine-tuned," i.e. final ordering of the reference vectors takes place.

Training of maps like here is very fast due to the small amount of training data available (see next section). Furthermore such maps enable comparisons between the financial situations of companies to be made. However, this approach does include the presumption that the input space for each year contains an adequately comprehensive description of the whole possible input space, i.e. all the realistically possible combinations of financial ratios.

I constructed map of the year 1998. The network topology chosen was hexagonal with 20 * 15 neurons in the map. The parameters of the best maps with respect to the average quantization error are given in Table 2.

Table. 2 Network parameters

Year	Phase	Training length	Learning rate	Neighborhood width	Quantization error
1998	1	1000	0.05	10	0.178910
	2	100000	0.02	3	

4. Results

In the construction process hundreds of maps were initialized and trained. The best ones, in respect of average quantization error (shown above in Table 2), were more carefully inspected, i.e. the locations of the companies and the values of weights (corresponding to financial ratios) were visualized.

The groups, or clusters in Figure 1 (Appendix 1) were identified by analyzing the weight distributions of the maps for the year 1998 in the forms of standard 2D U-matrices and weight maps as produced by the *Self-Organizing Map Program Package ver. 3.1* we used. The figures in Appendix 2 are the "weight maps" for the resulting map for the year 1998. On these maps the value of each weight in each neuron is visualized by gray-level imaging light shades representing high values and dark shades representing low values.

One interpretation of the defined groups based on weight maps for year 1998 is as follows:

- **Group A** can be considered as a "high performance" group. The group is doing rather well regardless of which ratio is used as an indicator.
- **Group B** is can be considered as an "average performance" group. The group has slightly low profitability, growth, capital structure, and liquidity, but has higher values in activities than others.
- **Group C** is the worst case. Most of companies of group C have low values in all financial ratios.

Weight maps in Appendix 2 show an informant thing. All the companies have the lowest value in stockholder's equity to total assets except for F4. That result shows one of reasons why our companies trapped to IMF. Too much indebtedness result in crisis of overall management activity.

5. Conclusions and future research

The objective of this study was to investigate the potential of self-organizing maps, to support in managing the complexity in a large data base by pre-processing the vast amount of financial data available on companies. The data base contained financial data on more than 80 companies in Korea. Using nine different ratios as variables – four profitability measures, two growth measure, one capital structure measure, one activity measure, one liquidity measure, no productivity measures – this paper constructed different maps for the year 1998. However, competitive benchmarking is well conducted in world-wide comparison, and by analyzing the financial performance of individual companies over time in a world-wide scale. The future research should be focused to that point.

References

Back, B., Mikko Irjala, Kaisa Sere, and Hannu Vanharanta, "Managing Complexity in Large Data Bases Using Self-Organizing Maps," TUCS Technical Report No. 8, Turku Centre for Computer Science, Sep. 1996, pp. 1-17.

Chen, S. K., P. Mangiameli, and D. West, "The Comparative Ability of Self-organizing Neural Networks to Define Cluster Structure," *Omega, International Journal of Management Science*, Vol. 23, No. 3, 1995, pp. 271-279.

Klimasauskas, C.C., "Applying Neural Networks, Part IV: Improving Performance," *PC/AI Magazine*, Vol. 5, No. 4. 1991.

Kohonen, T., "The Self-Organizing Map," Neurocomputing, Vol. 21, 1998, pp. 1-6.

Kohonen, T., Jussi Hynninen, Jari Kangas, and Jorma Laaksonen, "SOM_PAK: The Self-Organizing Map Program Package," Report A31, Helsinki University of Technology, 1996, pp. 1-25.

Martin-del-Brio, B. and C. Serrano-Cinca, "Self Organizing Neural Networks: The Financial State of Spanish Companies," In *Neural Networks in the Capital Markets*, edited by Refenes, John Wiley & Sons, 1995.

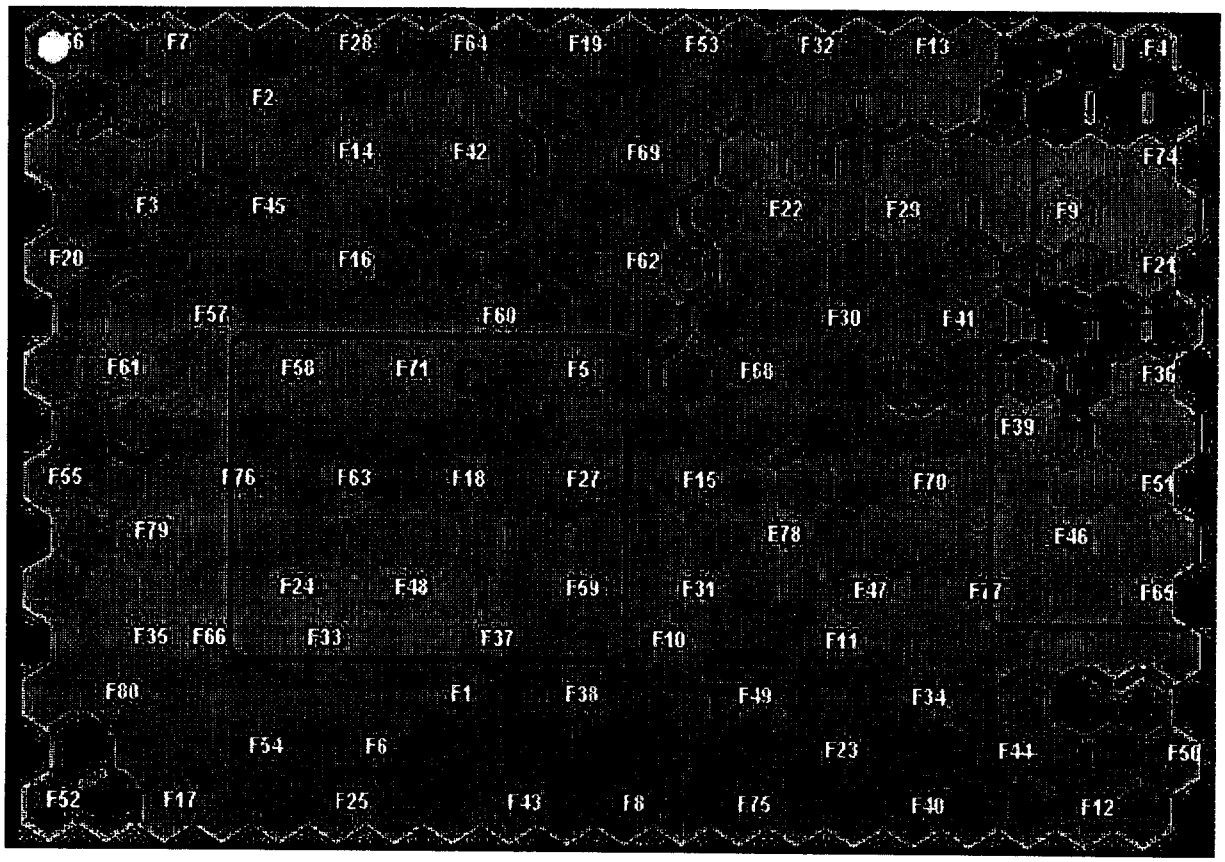
Rumelhart, D.E. and J. L. McClelland, *Parallel Distributed Processing*, The MIT Press, 1986.

SAS Institute Inc., *SAS Procedures Guide Version 6*, 3rd ed., 1994.

Trigueiros, D., "Accounting Identities and the Distribution of Ratios," *British Accounting Review*, Vol. 27, 1995, pp. 109-126.

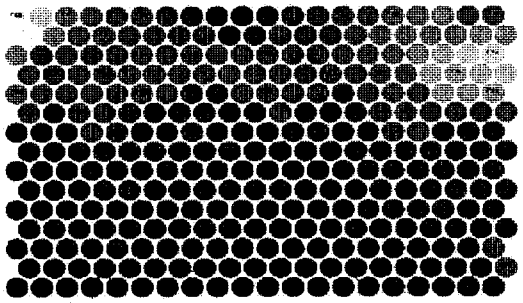
Vanharanta, H., "Hyperknowledge and Continuous Strategy in Executive Support Systems," Acta Academiae Aboensis, Ser. B, Vol. 55, No. 1. Turku, Finland, 1995.

Appendix 1. Standard 2D U-matrices

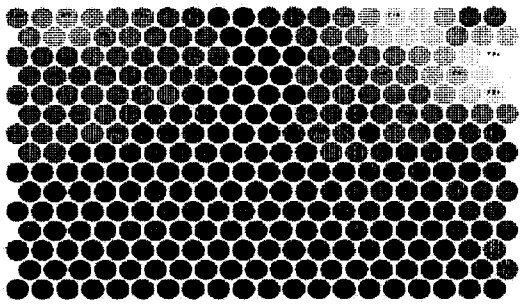


Appendix 2. Weight maps for year 1998

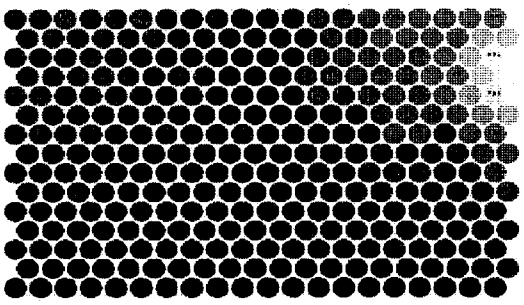
Net income to sales



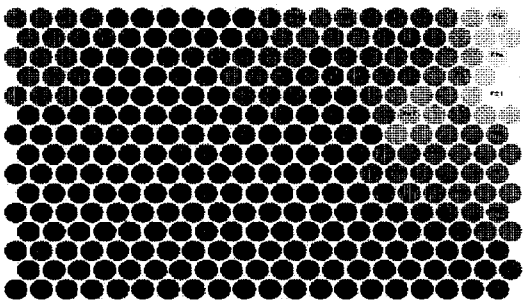
Ordinary income to sales



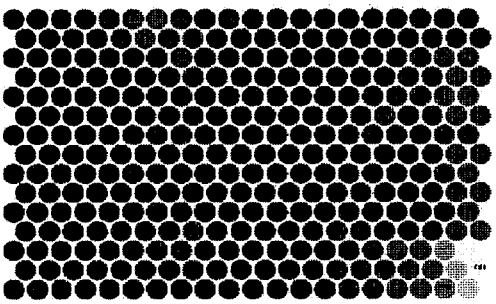
Ordinary income to total sales



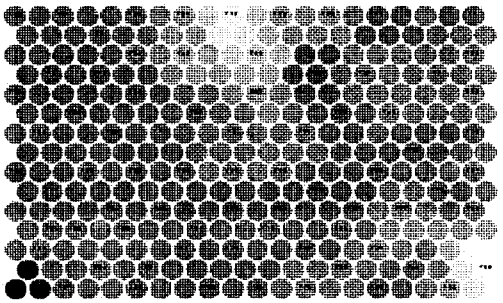
Ordinary income to stockholder's equity



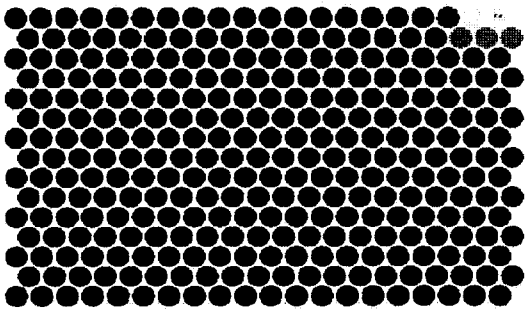
Growth rate of sales



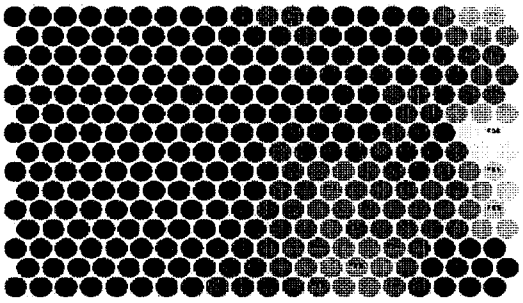
Growth rate of net income



Stockholder's equity to total assets



Stockholder's equity turnover



Cash flow/debt