A Map-Reduce based Approach for Mining Group Stock Portfolio

Chun-Hao Chen
Department of Computer Science and Information Engineering
Tamkang University, Taipei, Taiwan
chchen@mail.tku.edu.tw

Chao-Chun Chen
Institute of Manufacturing Information & Systems, Department of Computer Science & Information Engineering
National Cheng Kung University Tainan, Taiwan
chaochun@mail.ncku.edu.tw

ABSTRACT
In this paper, the map-reduce technique is utilized for speeding up the mining process and derived as similar results as our previous approach. The chromosome representation consists of four parts that are a mapper number, grouping part, stock part and portfolio part. According to mapper number, chromosomes in population are divided into subsets and sent to respective mappers. Fitness evaluation and genetic operations are the same with our previous approach, and executed on reducers. The evolution process is repeated until reaching the terminal conditions. Experiments are conducted on a real dataset to show the performance of proposed approach.

CCS Concepts
•Machine learning→Machine learning approaches→Bio-inspired approaches→Genetic algorithms •Theory of computation→Design and analysis of algorithms→Parallel algorithms→MapReduce algorithms.

Keywords
grouping genetic algorithms, grouping problems, group stock portfolio, map-reduce, stock portfolio optimization.

1. INTRODUCTION
Portfolio optimization is an attractive research topic since financial market has many financial instruments and the profit of the portfolio may be affected by various factors [20, 21]. The aim of portfolio optimization is to minimize the value at risk (VaR) and maximize the return on investment (ROI). Hence, a sophisticated approach for deriving a portfolio that takes these factors into consideration is needed. The mean-variance (M-V) model is the well-known approach for acquiring a stock portfolio [18], and many approaches have also been proposed it [1, 4, 17].

Investors may have various requests. An approach which can satisfy investor's requests for mining the stock portfolio was proposed by genetic algorithms in [4]. The fitness function that consists of objective criteria and subjective criteria was designed for chromosome evaluation. The suitability criterion, composing of a portfolio penalty (PP) and an investment capital penalty (ICP), was used to reflect the satisfactions of the users' requests.

Meanwhile, investors may not buy the suggested stocks for some reasons. For instance, the stock price of stock *A* may be too high to buy in recent trading date. When such situations are happened, other substituted stocks could be suggested are needed. Chen et al. thus proposed an approach for deriving group stock portfolio by grouping genetic algorithm [5]. Stocks in the same group mean they have similar properties. By utilizing the group stock portfolio, a set of stock portfolio could be generated. In other words, a set of substituted stocks could be provided when investors don't want to buy certain stocks. It encoded group stock portfolio into a chromosome by using grouping part, stock part, and stock portfolio part. Grouping and stock parts indicated how to split *n* stocks to *K* groups. Groups were then utilized to generate different stock portfolios. Stock portfolio part were exhibited whether the stocks were purchased and how many units they should be purchased. Then, each chromosome was evaluated by group balance and portfolio satisfaction. However, along with the increasing of stock number and group number, fitness evaluation of chromosomes is time-consuming.

In this paper, to deal with this problem, the map-reduce architecture which has good ability to speed up mining process is employed to improve our previous approach. The map-reduce based approach for mining group stock portfolio is presented. In the proposed approach, a group stock portfolio is represented by not only grouping part, stock part, and stock portfolio part but also a gene, namely *RID*. The gene indicates which mapper the chromosome should be assigned. It first generates initial population which are stored in HBase. In mapper phase, each mapper gets chromosomes from HBase according to *RID* in each chromosome. In reducer phase, each reducer receives chromosomes from each mapper. Genetic operations, fitness evaluation and reproduction are then executed on reducers. These two phases are repeated until the terminal conditions are reached. Experiments on the real dataset were also made to show the performance of the proposed approach.

2. PRELIMINARIES
In this section, related background knowledge is stated. The grouping genetic algorithm is given in Section 2.1. The map-reduce technique is then described in Section 2.2.

2.1 Grouping Genetic Algorithm
Genetic algorithms (GA) has proposed by Holland in 1975 and has been applied to various fields [11]. The advantage of GA is that it provides feasible solutions in a limited amount of time when the problem is difficult to be solved [10, 11]. An improved algorithm, namely grouping genetic algorithm (GGA), has been proposed for solving grouping problems based on GA [7]. Brown et al. compared performance of GA and GGA in different domains on some empirical tests, and pointed out that GGA was superior to the GA for large grouping problems [2]. The components of GGA are described in the following. The chromosome in GGA consists of two parts that are grouping part and object part. For example, a possible chromosome is shown as follows:

ACBBC: ABC.

In the chromosome, before the semicolon, the string "ACBBC" is the object part, which means that there are five objects. After the semicolon, the string "ABC" is the grouping part. It means that objects in object part should be divided into three groups. Therefore, the chromosome means that five objects

ASE BD&SI 2015, October 07-09, 2015, Kaohsiung, Taiwan
© 2015 ACM. ISBN 978-1-4503-3735-9/15/10□$15.00
DOI: http://dx.doi.org/10.1145/2818869.2818901

are partitioned into three groups. In this example, object o1 belongs to group 'A'. Objects o3 and o4 belong to group 'B'. Objects o2 and o5 belong to group 'C'. There are three genetic operations of GGA that are crossover, mutation and inversion. The crossover operation in GGA is different from GA, which exchanges groups instead of genes by using four steps. Continue previous example, the groups in chromosome C1 are A: {o1}, B: {o3, o4} and C: {o2, o5}. Assume that chromosome C1 is base chromosome and chromosome C2 is insertion chromosome, they are shown as follows:

$$C_1: A: \{o_1\}, B: \{o_3, o_4\}, C: \{o_2, o_5\},$$

$$C_2: a: \{o_1, o_2\}, c: \{o_3\}, b: \{o_4, o_5\}.$$

In first step, the position of the base chromosome in the group part is randomly selected. Assume that the position of the insertion in base chromosome is between groups B and C. And, the insertion group selected from C2 is group b. After crossover, the result shows as follows:

$$C_1': A: \{o_1\}, B: \{o_3, o_4\}, b: \{o_4, o_5\}, C: \{o_2, o_5\}.$$

Since the new formed chromosome C_1' has four groups and objects o_4 and o_5 are duplicated, it should be adjusted. In third step, the objects o_4 and o_5 are removed from groups *B* and *C*, and the result is shown as follows:

$$C_1'': A: \{o_1\}, B: \{o_3\}, b: \{o_4, o_5\}, C: \{o_2\}.$$

From chromosome C_1'', since the desired number of groups is three and the number of groups in C_1'' is four, in the fourth step, one group will be selected for removal. Assume that the group *B* is picked and the object o_3 in group *B* is moved to group *C*, the final result shows as follows:

$$C_1''': A: \{o_1\}, b: \{o_4, o_5\}, C: \{o_2, o_3\}.$$

The mutation operation in GGA is moved an element in a group to another group. For example, if object o_2 is selected and added to group *A*, then the chromosome becomes C_1'''': *A*: $\{o_1, o_2\}$, *b*: $\{o_4, o_5\}$, *C*: $\{o_3\}$. The last genetic operation is inversion. The goal of this operation is to change the order of the groups in chromosome such that the crossover operation can improve the diversity of chromosomes. Take chromosome C_2 as an example, if groups *A* and *B* are exchanged, after inversion, the result is C_2': *B*: $\{o_4, o_5\}$, *c*: $\{o_3\}$, *A*: $\{o_1, o_2\}$.

The main purpose of grouping problems is attempted to divide *n* elements into *K* groups and each element can only belong to a group. Assume there is a set of instances $U = \{u_1, u_2, ..., u_n\}$, in accordance with [7], the definition of grouping problems is stated as follows:

$$\cup U_i = U, \text{ and } U_i \cap U_j = \varnothing, i \neq j.$$

In general, the grouping problems may have a cost function which is formed from hard constraints of each problem. Garey et al. indicated that grouping problems are NP-hard [8]

2.2 Map-Reduce Technique

The Hadoop is an open-source software framework and used to process large dataset on distributed system. In Hadoop architecture, it consists of three main components that are Hadoop MapReduce, Hadoop Distributed File System (HDFS) and HBase. The Hadoop MapReduce provides an easy way for users to process large dataset efficiently by utilizing its powerful computation ability. The HDFS is a distributed file system that provides a high efficient storage environment. The HBase is a distributed database by utilizing row and column as index to store data. In the following, an example is given to illustrate the concept of map-reduce technique and shown in Figure 1.

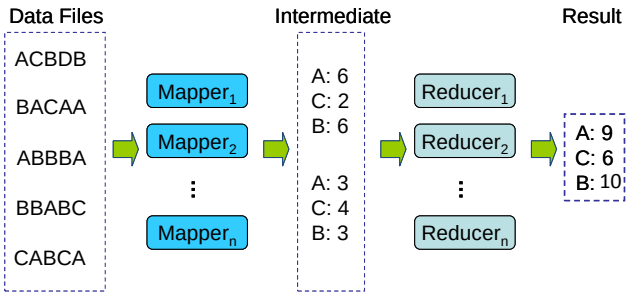

Figure 1. Example of map-reduce technique.

Figure 1 shows that there are five strings that are ACBCB, BACCA, ABBBA, BBABC and CABCA, and the appearance of each letter is counted after map-reduce process according to the predefined <key, value> pair. In this example, the key is the letter and the value is its appearance. By using the key value, frequent of each letter is counted in mapper procedure. The reducer then receives data from intermediate and appearance of each letter that are calculated by mappers is summed. As a result, A, B and C appear 9, 10 and 6 times are outputted.

Based on MapReduce framework, Nandi et al. proposed a MR-Cube framework that can derive useful Cube group by calculating historical data [19]. Chen et al. proposed closed frequent itemsets and association rule mining by using MapReduce framework [6]. By combined MapReduce and GA, some approaches were also proposed to solving job shop scheduling and time dependent vehicle routing problems [12].

3. STOCK PORTFOLIO OPTIMIZATION APPROACHES

Since M-V model is a well-known model for finding stock portfolio optimization, lots of portfolio optimization methods based on M-V model have been proposed for deriving portfolios [1, 4, 13, 15]. By evaluating profit and risk of protfolio, Hoklie et al. also proposed an optimization approach to dealing with portfolio optimization problem [13]. Since investors may have requests, to take them into consideration, Chen et al. proposed approaches for mining stock portfolio by GA [4]. Using ROI and VaR as objective functions, Bevilacqua et al. proposed an approach for finding a set of Pareto solutions [1]. Utilizing NSGA which is one of multi-objective genetic algorithms, the PONSGA that took different risk measures into consideration for mining portfolios was stated in [15].

In addition, many hybrid approaches that combine various mining techniques have been proposed [3, 9]. Gupta et al. presented an integrated approach for portfolio selection [9]. Assets were classified to three pre-defined classes by support vector machines. Then, GA was utilized to derive portfolios from the three classes in accordance with users' preferences. Bermúdez et al. proposed a MOGA-based approach that used a fuzzy ranking strategy for selecting efficient portfolios [3].

Furthermore, some approaches for deriving portfolios have been proposed [14, 16]. An approach for extracting portfolios with fixed costs and minimum transaction lots was proposed by a mixed-integer linear programming model [14]. Lin et al. presented three genetic-based models for the portfolio selection problem with minimum transaction lots based on M-V model [16].

ASE BD&SI 2015, October 07-09, 2015, Kaohsiung, Taiwan
© 2015 ACM. ISBN 978-1-4503-3735-9/15/10□$15.00
DOI: http://dx.doi.org/10.1145/2818869.2818901

4. PROPOSED ALGORITHM

In this section, according to Hadoop architecture, the flowchart of the proposed algorithm is stated in Section 4.1. The proposed group stock portfolio mining algorithm is then described in Section 4.2

4.1 Flowchart of Proposed Algorithm

The flowchart of the proposed algorithm that integrates grouping genetic algorithm, stock portfolio, and Hadoop is shown in Figure 2

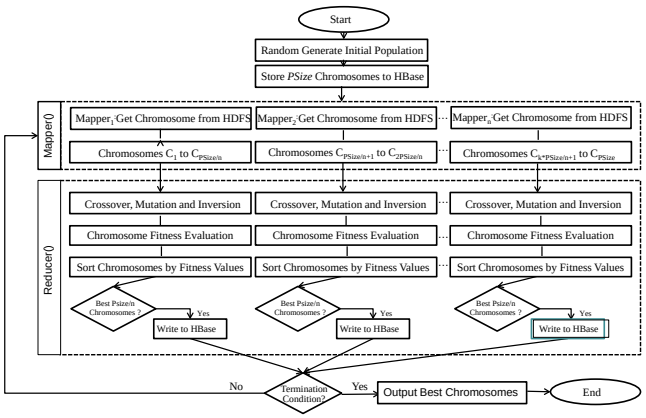

Figure 2.The flowchart of the proposed mining algorithm

Figure 2 shows that the proposed algorithm is divided into two parts that are mapper and reducer parts according to the Hadoop architecture. It first generates initial population using the given data randomly. The generated chromosomes are then stored to HBase. In mapper phase, each mapper gets chromosomes from HBase. Assume there are PSize chromosomes and n mappers, each mapper has PSize/n chromosomes. In reduce phase, firstly genetic operations are executed, including crossover, mutation and inversion. Secondly fitness of each chromosome is calculated. The top PSize/n chromosomes in each reducer are then kept and written to HBase. At last, if the termination conditions are reached, the best chromosome is outputted. Otherwise, it goes to mapper phase.

4.2 Components of Proposed Algorithm

In this subsection, chromosome representation, fitness evaluation and genetic operations are described.

4.2.1 Chromosome representation

In order to encode group stock portfolio into chromosome, three parts, namely grouping part, stock part and stock portfolio, are utilized in our previous approach [5]. Since map-reduce based approach is employed, a gene is added to the original chromosome representation to form new one. Let a set S consists of n stocks and they are divided into K stock groups. The new chromosome representation is shown in Figure 3

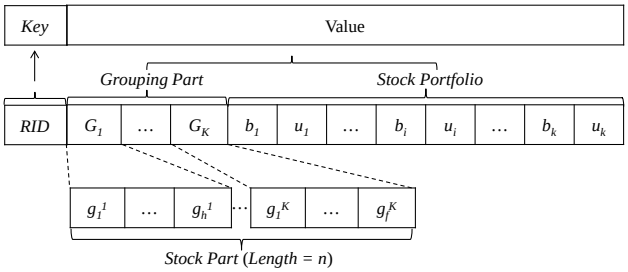

Figure 3. Chromosome representation

Figure 3 shows that the gene RID is the key in map-reduce architecture and will be used to assign a chromosome to a certain mapper. Other genes are the value in map-reduce architecture and will not only be represented a group stock portfolio but also indicate chromosome to be used for genetic operations in reducer part. Gi is the group number and stocks in the group means they are similar to each other. A stock si is represented by two genes. bi denotes a threshold and ui is the number of purchased units of si. When the value of bi is larger than or equal to 0.5, it means si is selected in the portfolio.

4.2.2 Fitness evaluation

Here, each chromosome is evaluated by two factors as defined in our pervious approach [5]. They are portfolio satisfaction and group balance. The portfolio satisfaction is utilized to evaluate the average profit and satisfaction of user's requests of all possible stock portfolios in a chromosome. A stock portfolio with high portfolio satisfaction means that it can both maximize return on investment and satisfy user's requests. The group balance is used to make the groups represented by the chromosome have as similar number of stocks as possible. The fitness function is defined as formula (1):

$$f(Cq) = portfolioSatisfaction(Cq)*[balance(Cq)]^{\alpha}, \quad (1)$$

where parameter α which can be set according to the investor's favor is used to control the relative influence of the two factors..

4.2.3 Genetic operations

Genetic operations include crossover, mutation and inversion operations are executed in the proposed approach. Here, the crossover operations are executed on grouping part and stock portfolio, mutation operations are executed on stock part and stock portfolio, and inversion operation is executed only on grouping part.

4.3 Proposed Map-Reduce Based Mining Algorithm

The map-reduced based group stock portfolio mining algorithm:

INPUT: A set of stocks with stock prices $S = \{s_i \mid 1 \le i \le n\}$, a predefined maximum number of purchased stocks in portfolio *numCom*, a predefined maximum investment capital *maxInves*, a predefined maximum number of purchased units of a stock *maxUnit*, cash dividends of stocks $Y = \{y_i \mid 1 \le i \le n\}$, a number of groups K, a parameter α, a population size *PSize*, a crossover rate p_c, a mutation rate p_m, an inversion rate p_I, and generations G.

OUTPUT: The group stock portfolio $G = \{G_i \mid 1 \le i \le K\}$.

STEP 1: Generate initial population with *PSize* using the following sub-steps:

ASE BD&SI 2015, October 07-09, 2015, Kaohsiung, Taiwan
© 2015 ACM. ISBN 978-1-4503-3735-9/15/10□$15.00
DOI: http://dx.doi.org/10.1145/2818869.2818901

Sub-step 1.1: Generate grouping part with K randomly such that $G_1 \cup G_2 \cup \ldots \cup G_k = S$, $G_i \neq \phi$ and $\forall i \neq j$, $G_i \cap G_j = \phi$.

Sub-step 1.2: Calculate average cash dividend of each group G_i according to cash dividends of stocks Y using the following formula:

$$avgCD(G_i) = \sum_{h=1}^{|G_i|}(y_h / |G_i|),$$

where $|G_i|$ is number of stocks in group G_i, and y_h is the h-th cash dividend of group G_i.

Sub-step 1.3: Calculate proportion of average cash dividend of each group G_i to all groups using the following formula:

$$proportionAvgCD(G_i) = \frac{avgCD(G_i)}{\sum_{a=1}^{K} avgCD(G_a)},$$

where $avgCD(G_i)$ is the average cash dividend of each group G_i, and $\sum_{a=1}^{K} avgCD(G_a)$ is the summation of average cash dividend of all groups.

Sub-step 1.4: Randomly generate *numCom* values from the range [0, 1] and collect them in a set $R = \{r_i \mid 1 \leq i \leq numCom\}$.

Sub-step 1.5: For each element in R, if the random value r_i is between $proportionAvgCD(G_{i-1})$ and $proportionAvgCD(G_i)$, then group G_i is put into the candidate portfolio.

Sub-step 1.6: Generate stock portfolio according to the candidate portfolio. For each selected group G_i, set its b_i in the chromosome larger than 0.5. Otherwise, set it less than 0.5. Randomly generate the corresponding number of purchased units of each group from the range [0, *maxUnit*].

Sub-step 1.7: Assign a *RID* to the chromosome.

Sub-step 1.8: If *PSize* chromosomes are generated, go to the next step. Otherwise, go to Sub-step 1.1.

Sub-step 1.9: Store initial population to HBase.

Step 2: Divide each chromosome into corresponding mapper by its *RID*. Assume there is n mappers, each mapper thus contains *PSize*/n chromosomes.

Step 3: For each reducer, do the following substeps:

Sub-step 3.1: Receive *PSize*/n chromosomes from mapper.

Sub-step 3.2: Execute crossover operation on the population.

Sub-step 3.3: Execute mutation operation on the population.

Sub-step 3.4: Execute inversion operation on the population.

Sub-step 3.5: Calculate fitness value of chromosome C_q using formula (1).

Sub-step 3.6: Execute selection operation on the population to form the next population.

Sub-step 3.7: If the stop criterion is satisfied, go to the next step. Otherwise, store new generated chromosomes to HBase and go to Step 2.

Step 4: Output the chromosome with the best fitness value.

5. EXPERIMENTAL RESULTS

In this section, experiments are conducted to show the performance of the proposed approach. The parameter setting of the proposed approach is shown in Table 1.

Table 1. Parameter setting of the proposed approach

PSize	60	*numCom*	3
p_c	0.8	*maxInves (million)*	2
p_m	0.3	*maxUnit*	40
p_i	0.6	α	2
Generation	100	K	4

The experimental dataset was collected from the TSE and 31 stocks are used for performance evaluation. The dataset contains the stock prices of stocks from 2013/01/01 to 2014/12/31, the cash dividends of stocks, and the risk values, where the risk value of each stock is calculated by history simulation. Experiments on the real dataset were made to show the perofrmance of the proposed approach. The comparison result between the proposed appproach and our previous approach in terms of execition time is shown in Figure 4.

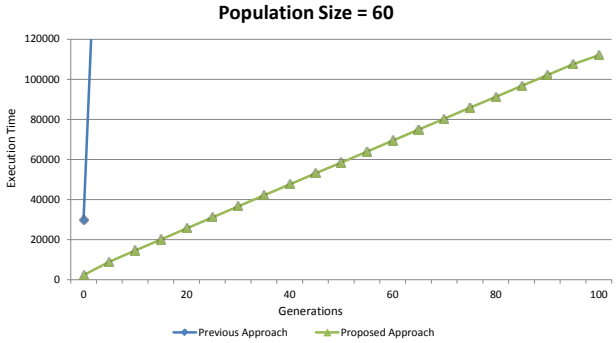

Figure 4. Comparison of propsoed and previous approaches in terms of execution time (population size = 60)

Figures 4 shows that the execution time of proposed approach is better than previous approach. From the results, by using map-reduc technique, the proposed approach can speed up the evolution time our previous.

6. CONCLUSIONS AND FUTURE WORKS

In this paper, the map-reduce technique has been utilized to improve our previous approach. The map-reduce based approach for mining group stock portfolio has been proposed. The proposed approach are divided into two phases, including mapper and reducer phases. In mapper phase, each chromosome is accessed from HBase and assigned to certain mapper according to the gene RID in chromosome. In reducer phase, fitness evaluation and genetic operations are executed. Experimental results on the real data also show that execution time of proposed approach is better than our previous approach. In the future, various map-reduce architectures will be investigated to improved the proposed approach.

7. ACKNOWLEDGMENTS

This research was supported by the Ministry of Science and Technology of the Republic of China under grant MOST 104-2221-E-032-040.

ASE BD&SI 2015, October 07-09, 2015, Kaohsiung, Taiwan
© 2015 ACM. ISBN 978-1-4503-3735-9/15/10□$15.00
DOI: http://dx.doi.org/10.1145/2818869.2818901

8. REFERENCES

[1] V. Bevilacqua, V. Pacelli and S. Saladino, "A novel multi objective genetic algorithm for the portfolio optimization," *Advanced Intelligent Computing*, pp. 186-193, 2012.

[2] A. L. Blum and R. L. Rivest, "Training a 3-node neural networks is NP-complete," *Neural Networks*, Vol. 5, pp. 117-127, 1992.

[3] J. Bermúdez, J. Segura and E. Vercher, "A multi-objective genetic algorithm for cardinality constrained fuzzy portfolio selection," *Fuzzy Sets and Systems*, Vol. 188, pp. 16-26, 2012.

[4] C. H. Chen, Y. C. Hsieh and Y. C. Lee, "The YTM-based stock portfolio mining approach by genetic algorithm," *The IEEE International Conference on Granular Computing*, 2013.

[5] C. H. Chen, C. B. Lin and C. C. Chen, "Mining group stock portfolio by using grouping genetic algorithms," *IEEE Congress on Evolutionary Computation*, 2015.

[6] G. P. Chen, Y. B. Yang and Y. Zhang, "MapReduce-Based Balanced Mining for Closed Frequent Itemset," *IEEE International Conference on Web Services*, pp. 652 - 653, 2012.

[7] E. Falkenauer, "A New representation and operators for genetic algorithms applied to grouping problems," *Evolutionary Computation*, Vol. 2, pp. 123-144, 1994.

[8] M. R. Garey and D. S. Johnson, "Computers and Intractability: A Guide to the Theory of NP-Completeness", *Macmillan Higher Education*, 1979.

[9] P. Gupta, M. K. Mehlawat and G. Mittal, "Asset portfolio optimization using support vector machines and real-coded genetic algorithm," *Journal of Global Optimization*, Vol. 53, pp. 297-315, 2012.

[10] D. E. Goldberg, "Genetic algorithms in search, optimization, and machine learning," *Addison Wesley*, 1989.

[11] J. H. Holland, "Adaptation in natural and artificial systems," *University of Michigan Press*, 1975.

[12] D. W. Huang and J. Lin, "Scaling Populations of a Genetic Algorithm for Job Shop Scheduling Problems Using MapReduce," *IEEE Second International Conference on Cloud Computing Technology and Science*, pp. 780 - 785, 2010.

[13] L. R. Z. Hoklie, "Resolving multi objective stock portfolio optimization problem using genetic algorithm," *International Conference on Computer and Automation Engineering*, pp. 40-44, 2010.

[14] H. Kellerer, R. Mansini and S. M. Grazia, "Selecting portfolios with fixed costs and minimum transaction lots," *Annals of Operations Research*, Vol. 99, pp. 287-304, 2000.

[15] P. C. Lin, "Portfolio optimization and risk measurement based on non-dominated sorting genetic algorithm," *Journal of Industrial and Management Optimization*, Vol. 8, pp. 549-564, 2012.

[16] C. C. Lin and Y. T. Liu, "Genetic algorithms for portfolio selection problems with minimum transaction lots," *European Journal of Operational Research*, Vol. 185, pp. 393-404, 2008.

[17] K. Lwin, R. Qu and G. Kendall, "A learning-guided multi-objective evolutionary algorithm for constrained portfolio optimization," *Applied Soft Computing*, Vol. 24, pp. 757-772, 2014.

[18] H. M. Markowitz, "Harry Markowitz: Selected Works," *World Scientific Publishing Company*, 2009.

[19] A. Nandi, C. Yu, P. Bohannon, R. Ramakrishnan, "Data Cube Materialization and Mining over MapReduce," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24 , No. 10, pp. 1747 - 1759, 2012.

[20] T. P. Patalia and G. Kulkarni, "Design of genetic algorithm for knapsack problem to perform stock portfolio selection using financial indicators," *International Conference on Computational Intelligence and Communication Networks*, pp. 289-292, 2011.

[21] E. Wah, Y. Mei and B. W. Wah, "Portfolio optimization through data conditioning and aggregation," *IEEE International Conference on Tools with Artificial Intelligence*, pp. 253-260, 2011.

ASE BD&SI 2015, October 07-09, 2015, Kaohsiung, Taiwan

© 2015 ACM. ISBN 978-1-4503-3735-9/15/10□$15.00

DOI: http://dx.doi.org/10.1145/2818869.2818901