MIKADO (MARKET INVESTIGATION AND KNOWLEDGE ACQUISITION THROUGH DATA OBSERVATION): FIRST RESULTS AND PERSPECTIVES[♦]

Fabien Boniver and Patrick Meyer [♣]

December 15, 2002

Abstract

We show how financial databases of companies and their financial ratios, as well as analysts' recommendations, can be built and dealt with by means of statistical analyses and machine learning procedures. In a first approach, we aim at validating daily life assertions about stocks, such as: "For stock XX, the price to earnings ratio is low: it's a bargain." In particular, we focus on measuring levels of association between fundamentals and recommendations. We pay attention to the variations of these associations with respect to the industrial sectors. We also build rules, represented by decision trees, to deduce the recommendations from the available fundamentals.

Keywords: Financial ratios, ratings of stocks, knowledge extraction, human expertise centred decision aid, Internet, data mining.

1. Introduction

It is common knowledge that the complexity of decision procedures increases in the presence of disruptive elements among the information available to decision makers. Those disruptive yet often intrinsic parts of the information include randomness, unreliable sources, interacting criteria, extreme values and missing data. The management of a stock portfolio certainly involves such complex decisions. Mathematical methods provide various tools to investigate financials topics such as time series of stock prices or probability distributions associated to risks. They also constitute as usual an acknowledged framework for rigorous assertions.

Unfortunately, everyday assertions about stock markets are often far from rigorous. Consider for instance the following assertion: "Analyst XX raised its recommendation on stock YY from *market outperform* to *buy*." The semantic interpretation of such a recommendation can be misleading. Indeed, while it is clear that such a recommendation corresponds to a valuation on an ordinal scale, it is not obvious that it contains further information. In other words, *buy* indicates something better than *market outperform* but its meaning might be unrelated to the act of buying the stock under consideration. Besides, while financial *fundamentals* (also known simply as *financials*) associated to a particular stock are better defined, their semantic interpretation may also lead to some ambiguities. Consider the sentence: "According to its Price/Earnings (PER) ratio, stock YY appears to be cheap," which is sometimes followed by a *buy* recommendation. What does cheap mean when it comes to stock YY, except that its PER is low? And what should we do with a cheap stock? Buy it, because it is a bargain, or sell it, because it is out of fashion? Furthermore, is there any measurable association between the recommendations on YY and its financial *fundamentals* such as the PER?

♦ This short note was submitted to the 14th Mini-Euro Conference on Human Centered Processes to be held in Luxembourg in May 2003. An expanded version is under construction.

♣ The first listed author is FNRS Postdoctoral Researcher. The second listed author is Researcher for the Région Wallonne. Both are affiliated to the Department of Mathematics, University of Liège, B37, 4000 Liège, Belgium. Equal authorship is implied. The second listed author is the corresponding author.

The MIKADO project aims at two successive goals:

1. Collect enough data in a form that allows to validate statistically assertions about fundamentals or ratings pertaining to a stock;
2. Formulate such assertions that are supported by the data and study the evolution of their validity with respect to time.

In this note, we first present its methodology in Sections 2, 3, and 4. This includes a description of the gathered data and of the statistical and data mining techniques implemented to extract knowledge from the databases. In Section 5, we present results about the association between fundamentals and recommendations and the explanation of the latter from the former. They seem to open promising perspectives, which we finally briefly comment in Section 6.

2. Data description

The database consists in a bimonthly updated set of pairs of tables of financial information freely available on the Internet. The data we chose to analyse belong to two distinct categories. On the one hand, they are composed of various fundamental attributes that indicate characteristics such as the health or the size of the companies. This type of data is objective and each attribute is clearly defined. On the other hand, we focus on what we call recommendation profiles. For a given stock, it is a summary of the recommendations given by financial analysts regarding the management of the stock inside an investor's portfolio. Each recommendation is a value on a 5-levelled ordinal scale which goes from the worst *strong sell* to the best *strong buy*. Often, this scale is built by associating integers from 5 (strong sell) to 1 (strong buy) with these levels. We therefore adopt the sign convention that a recommendation is greater than another one if it is *worse*.

We denote by *recommendation profile* the frequency table of the recommendations about a particular stock. The experts are not univoquely identifiable. But their position as institutional experts leads us to think that their recommendations are not completely out of sense. These data are more subjective.

The data we are focusing on have been chosen for multiple reasons. They are freely accessible on the Internet, represent a large amount of shares and can be systematically collected from stable sources. This last point is very important as it guarantees some homogeneity within the data. The fundamentals are collected from the website of *Multex Investor* whereas the recommendation profiles' source is the website of *Yahoo! Finance*.[i]

Let us proceed to a technical description of the available data. The fundamentals we are considering can be divided in a few general classes for an easier understanding. The following list considers only these general sets of attributes. The interested reader can refer to [2].

Category	Example
General information	Company name, symbol, sector, employees, …
Price and volume attributes	Recent price, beta, …
Equity size and value	Market capitalisation, shares outstanding, …
Dividend information	Yield, payout ratio, …
Financial strength/solvency	Quick ratio, Total debt over equity, …
Market value ratios	Price / Earnings, Price to Book ratio, …
Normalized (per share) financials	Earnings, Sales, Book value, …
Management effectiveness	Return on investment, …
Profitability ratios	Profit margin, …

Table 1: General categories of fundamentals

3. Investigation methodologies

3.1. Statistical estimation of association levels

We begin by some basic statistical analysis, namely, the estimation of association levels between fundamentals and recommendations. Further details are given in Section 5.

3.2. Machine learning

The second part of our research is focusing more particularly on machine learning procedures for data mining. These methods are used to extract implicit information from data. The data mining methods are based on algorithms that try to detect patterns and regular structures in the data. In order to allow experts to give a semantic interpretation of our results, we focus in a first approach to methods with a clear, easily understandable output. Obviously, the goal is to reproduce the original classification as accurately as possible.

4. Data pre-processing

Before going further into these investigations, we first pre-process the data. We only retain those data for which there is no missing value; furthermore, we require that the total number of recommendations be positive. That way, we obtain a dataset of 1621 stocks, each of which is recommended about 8 times on average.

To the variables, we add one that we name *Price Position*. It is defined by the formula

$$\frac{P-Pl}{Ph-Pl},$$

where, for a given stock, P denotes the last quote observed (at the time the database was built), Pl, the lowest price in the trailing twelve months, and Ph, the highest price in the trailing twelve months.

For each stock, we also compute an ordinal central tendency parameter for the recommendation profile. It is common practice to build a so-called *analyst consensus* based on a weighted mean of recommendations. This is obtained by associating weights from 1 to 5 to the ordinal labels “strong buy” to “strong sell”. Unfortunately, this procedure does not define an ordinal statistic if the weights are allowed to vary. We thus retain the median as a central tendency parameter. Then, we rearrange the observations into a *list of recommendations*. It is obtained by replicating the entry of a particular stock in the original dataset as many times as there are covering brokers for this stock. This replication takes into account the fact that not all stocks are reviewed by the same number of brokers. We plan to refine this duplication by matching exactly the recommendation profiles. Our study is thus based on a table made of 1,621 complete observations (stocks) of 39 variables, of which we derive the 12,646 entries in the list of recommendations.

5. First results

5.1. Statistical estimation of association levels

Our first results regard the critical examination and validation of daily life assertions about stocks. Assume that an investor is presented with the choice of investing in one of the two companies XX and YY. Assume furthermore that the available fundamentals for these companies are completely similar, except for one point: the profit margin of XX exceeds that of YY. Then we expect the potential shareholder to prefer investing in company XX. This kind of reasoning is often implied in many commentaries about investment opportunities. The sole problem is of course the fact that two companies are never *completely* similar. An analogous discussion can be found in [2, p. 65]. Section 3.3 in this reference is devoted to the analysis of financial ratios.

We measure the level of association at a given time between a single fundamental variable and the recommendation median. This association is measured within each sector, e.g. financial,

healthcare. The association level is measured with an estimator that respects the ordinal nature of the recommendations. Specifically, we use Kendall's tau estimator. Recall that it is defined by the formula

$$\frac{C-D}{\sqrt{C+D+Tx}\sqrt{C+D+Ty}},$$

where C (resp. D) denotes the number of concordant (resp. discordant) pairs among all possible pairs of observations, and where T_x (resp. T_y) denotes the number of pairs for which only the first measures are equal (resp. for which only the second measures, i.e. the recommendation medians, are equal). The p-value associated to this statistic is also computed. All the computations involved here are performed using the statistical software R, which is documented in [1]. The previous procedure is repeated for all variables within each sector.

We retain only those associations that are very significant (p-value less than or equal to .01). Four variables are very significantly associated to the median of the recommendations in 10 sectors. We say that these variables *occur 10 times*. Among these, for instance, the total debt to equity ratio occurs five times with a positive level of association and five times with a negative one. With the same total occurrence, the payout ratio is almost always positively associated to the median. All other parameters being equal within a relevant sector, this can be summarized by: "The higher the payout ratio, the worse the recommendation" which might be considered as somewhat surprising. Table 2 lists a small sample of these occurrences.

Clearly, a summary examining the influence of groups of similar variables within each sector separately would be more readable. We thus group the variables according to their "intuitive" meaning and we count, for each sector and each group, the number of variables having a very significant positive (resp. negative) association level with the recommendation median. Table 3 lists these group influences sector by sector. If all variables in one group have a positive (resp. negative) association level for a given sector, we indicate the relevant sign in the corresponding cell. If different signs occur, we indicate the number of negative and positive associations, separated by a dash. We indicate "0" if no variable of the group under consideration is very significantly associated to the recommendation median. For an easier reading, it is worth remembering the sign convention adopted in Section 2.

Surprisingly enough, the definitions of the groups are quite coherent: when variables in a given group are associated to the recommendation median, the sign of this association is often common to the whole group. This seems to be supported by a factor analysis that is currently being performed.

	Negative occurrences	Positive occurrences	Total occurrences
Total debt to equity	5	5	10
Payout	1	9	10
Yield	0	9	9
Price position	8	0	8
Return on investment	6	2	8
Number of employees	1	6	7
Operational margin	7	0	7
Price to earnings	5	2	7
Profit margin	5	1	6
Book value per share	1	4	5

Table 2: Significant associations and their occurrences

		Basic Materials	Capital Goods	Conglo-merates	Consumer Non-Cyclical	Consumer Cyclical	Energy	Financial	Healthcare	Services	Technology	Transpor-tation	Utilities
	Price	0	1-3	-	-	1-1	1-2	4-1	4-1	2-3	3-1	1-1	-
	Size	0	+	-	+	+	+	-	-	+	+	+	1-1
	Dividends	+	+	0	+	+	+	+	+	+	+	+	-
Profitability	**Margins**	+	+	-	0	-	-	-	-	-	-	-	-
	Returns	+	0	-	-	-	-	1-1	-	-	-	0	0
	Mrkt value	-	-	-	1-1	2-1	-	-	-	-	-	1-2	-
	Norm. fin.	+	+	+	-	+	+	-	3-1	+	-	1-1	+
Solvency	**Short term**	+	-	-	-	-	-	-	+	+	0	0	0
	Long term	-	+	+	+	+	0	-	-	-	-	0	-

Table 3: Sectorwise group influences

5.2. Machine learning

For the purpose of this paper, we first limit ourselves to a decision tree classifier based on the c4.5 algorithm. It is implemented in the WEKA software package [4]. The main reason of this choice is the easy readability and the interpretability of the results. Further information on machine learning can be found in [3]. The data used for the building of the model is the list of recommendations.

The chosen output attribute is the median of the recommendation profiles on an ordinal scale. For the sake of simplicity, we say that a stock is *recommended i* and *assigned to j* if the median of the recommendation profile for this stock is *i* and the decision tree assigned it to class *j*. If *i* equals *j*, the assignment is said to be *correct*. We emphasize the fact that in addition to the variables observed in the analysis of association levels, the input attributes here include nominal ones, like the sector and the industry.

A first run on the data with the *median* of the recommendation profiles as the output attribute results in a decision tree with 482 leaves. A path from the top node to a leaf in such a tree can be seen as an *if-then* decision rule. A sample of the most salient attributes is given by the set {yield, price position, operational margin, employees, price/cash flow}. Roughly speaking, these are the attributes with the highest influence on the output. The following table summarises 3 types of tests performed on the data: the classification accuracy on the training set (12,646 instances), a 10-fold cross validation on the same training set, and the accuracy of the classification performed on the original (or test) set of 1,261 stocks. A second run on the data is performed with a slightly different output attribute: instead of considering intermediate classes due to the calculation of the median of the recommendation profiles, we compute the *pessimistic median* by merging an intermediate class with its worst neighbour class. Indeed, the usual definition of the median raises 4 intermediate classes that introduce an artificial supplemental precision for the assignments.

		Training set		10-fold cross validation		Test set	
	Output	Correct	Kappa	Correct	Kappa	Correct	Kappa
Median	{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5}	93.56%	0.90	89.89%	0.84	80.10%	0.72
Pessimistic median	{1, 2, 3, 4, 5}	94.65%	0.91	91.82%	0.86	84.15%	0.74

Table 4: Accuracy of the two runs

The confusion matrix and a detailed summary of the accuracy of the classification allow us to better understand the overall structure of the assignments. They are represented below for the second run, in the case of the test set. The confusion matrix should be read from the left to the right: the integer at the intersection of line *i* and column *j* is the number of times the decision tree assigned a stock recommended *i* to *j*. Therefore, the diagonal of this matrix represents the correctly

classified stocks. The *TP rate* of line *i* indicates the proportion of the stocks recommended *i* that have been correctly assigned. The *precision index* of column *j* indicates the proportion of the stocks assigned to *j* for which the assignment is correct.

	1	2	3	4	5
1	191	54	39	0	0
2	25	717	32	0	0
3	26	71	453	1	0
4	1	1	6	2	0
5	0	1	0	0	1

Class	TP Rate	Precision
1	0.673	0.786
2	0.926	0.850
3	0.822	0.855
4	0.200	0.200
5	0.500	0.500

Table 5: Confusion matrix and accuracy of classification per class for the pessimistic median

6. Conclusion

We have shown the possibility of gathering freely accessible financial information, and to structure it in order to allow its systematic and rigorous treatment. Restricting ourselves to a snapshot of such information at a given time, we have studied the relationship between fundamentals and analysts' recommendations of stocks. Not surprisingly, very significant associations exist between the fundamentals and the recommendations.

Yet, the "signs" of the associations are not always intuitive and can vary considerably among sectors. In view of the results of Subsection 5.2, machine learning techniques seem to deserve further consideration in order to structure the influence of fundamentals on the recommendations.

Our methodology helps making numerous further questions tractable. In particular, it would be of uttermost interest to examine the regularity of association patterns with respect to time; to obtain more readable rules regarding the rating of stocks, by conducting a decision tree analysis based upon fewer attributes. These attributes could be chosen according to the grouping of Subsection 5.1. Obviously, it seems also appealing to us to evaluate the predictive power of carefully crafted decision trees and submit the extracted knowledge to the critical examination of financial experts.

We hope that the very approach exposed in the present note can help in rigorously tackling these newly risen problems.

References

[1] Ross Ihaka and Robert Gentleman, *R: A language for data analysis and graphics*, Journal of Computational and Graphical Statistics, 1996, vol. 5 n. 3, 299--314.
[2] Stephen A. Ross, Randolph W. Westerfield, and Bradford D. Jordan, Fundamentals of corporate finance, 2000, Irwin McGraw-Hill, fifth edition.
[3] Louis Wehenkel, Applied Inductive Learning : Course Notes, 2000, University of Liège.
[4] Ian H. Witten and Eibe Frank, Data Mining: Practical machine learning tools with Java implementations, Morgan Kaufmann, San Francisco. (software downloadable at http://www.cs.waikato.ac.nz/ml/weka)

[i] Multex Investor is a service mark of Multex.com, Inc. Multex.com, Inc. is not involved in preparation of the data, does not review, edit or endorse them in any way. See http://www.multexinvestor.com . Yahoo! is a trademark of Yahoo! Inc. Yahoo! has not reviewed, and in no way endorses the validity of such data. See http://finance.yahoo.com .