STEVENS INSTITUE OF TECHNOLOGY

FE 800

PROJECT IN FINANCIAL ENGINEERING

LEARNED SECTORS

Authors:
Choyon Anwar
Jiashi Li
Zhengkun Ye

Supervisors:
Dr. Thomas Lonon
Dr. Dragos Bozdog
Dr. Ionut Florescu
Dr. Papa Ndiaya

May 16, 2019

1. Abstract

Companies are often categorized into artificial groups referred to as industries and, more specifically, sectors based on qualitative assessments. This project focuses on demonstrating the efficacy of grouping companies with quantitative evaluations. This is done by employing unsupervised machine learning techniques to partition quantitative data, gathered for each company, to generate novel sectors. These newly formed sectors are then compared to benchmark sectors. Naturally, the comparison would be done at the sector-level, however, since only quantitative data was used, no linkage between any single novel sector to a benchmark sector can be determined. Thus, the comparison is done via groups of sectors. First, we assume each sector is a financial portfolio. The assumption then allows us to describe each sector with a portfolio performance measure. The sectors are now defined with their respective performance measures; we utilize a five-number summary to illustrate the workings of the group of sectors. We find that the novel sectors perform relatively identically as the benchmark sectors.

2. Introduction

Grouping companies into justifiable sectors is becoming increasingly important as investors are interested in monitoring the dynamics of company-to-company interactions in the financial realm. There are many taxonomies already established for this purpose including: Global Industry Classification Standard (GICS), Industry Classification Benchmark (ICB), Thomson Reuters Business Classification (TRBC), and more. Each of these taxonomies is created by a different organization that uses their own respective methods in creating the partitions and classifying the sectors within which the companies are sorted. However, this is all seemingly subjective to the organizations' methods – which are mostly likely dependent on qualitative assessments. It is this qualitative nature that we refer to as artificial. One issue with these artificial groupings is the possibility of one company being sorted into different sectors using different taxonomy schemes.

Instead of using qualitative assessments, we conduct an experiment that uses unsupervised machine learning (UML) techniques to process quantitative data to generate novel sectors. These new sectors are what we refer to as generated sectors (GS) which are then compared to a benchmark sector (BMS), where the benchmark is one of the multiple artificial taxonomies. In our case we use the GICS taxonomy. This experiment is primarily to investigate the efficacy of using such UML techniques to generate more robust sectors. One application of this new way to sectorize, featured in this paper, is to track risk and return of the new sectors. We hypothesize that the GS will perform better than the BMS since we are using quantitative data as opposed to qualitative data – the former being easily measurable and comparable between companies.

Note, since we only use quantitative data there is no way to link a GS to any one of the BMS. Thus, a comparison at the sector level is not accomplishable. We then move a level up in the scope of the project to achieve comparison. At this level, defined as a setting, we group the GS that were formed using specific parameters in the UML technique and compare this setting to the setting that which is the group of benchmark sectors. Now that the method of comparison is established, all that is left to do is assign quantitative descriptors for each sector, both generated and benchmark. This is done by assuming each sector to be its own financial portfolio and then computing portfolio performance measures, namely the Sharpe and Treynor Ratio, to demonstrate the featured application. Both measures analyze the return of a portfolio compared to risk, total risk of the standard deviation of portfolio and the systematic risk of the portfolio, respectively. Once the sectors are defined with the

performance measures, we utilize a five-number summary at the setting level to analyze the performances of GS to BMS.

For this project, we use R and RStudio exclusively for all data analysis and visualizations. We do this to take advantage of the prebuilt packages that compute many of our calculations such as: Sharpe and Treynor Ratio, UML techniques, and more. Also, we sourced our data from Bloomberg exclusively since there were two other groups working on this project who already established their data source. It was decided to have a wide variety of data from differing sources for novelty purposes.

3. Literature Review

The most significant piece of literature impacting our project was one about disovering user communities on the internet using unsuprivsed machine learning techniques. This literature, although does not concern itself with financial data, proved to be a fantastic blueprint for our project. We took the procedures the authors used in data collection, extraction, cleaning, as well as methods for analysis and selecting appropriate algorithms and applied them to our project. The two projects' ideas are similar. Both set out to group data points in a meaningful way such that valuable information is extract for further use. This allowed us to replicate much of their procedures but with quantitative, financial data set.

4. Methodology – Phase One

4.1. Data Collection

For this project, we were not limited to any specific source for data collection. It was proposed that any and all data, ranging from data companies to social media, was eligible for analysis. Due to the time limit, we forwent the vast amount of data sources available online and choose to specifically work with data from a well-established and trusted source such as Bloomberg. But even then, Bloomberg itself has many datasets that are eligible for our project. Still, we needed to create a standardized method of collecting data. We then found, with the help from our advisors, templates of data via a Bloomberg-Excel connection. These templates are organized by the type of data they collect for various ticker symbols in the market. One such template, and the one we used, is the Financial Analysis. This template contains multiple spreadsheets of data for a ticker symbol, namely: Income Statement, Cash Flows, Balance Sheet, Ratios, and a few more displaying graphics. We initially choose the first four spreadsheets mentioned and then only used the first three as an issue with data extraction and cleaning became apparent. Once we had our data source, standardized in format, all we simply did was save each spreadsheet for a corresponding ticker symbol as csv files. Csv format was chosen due to an issue with reading the data set into R with other file formats.

The universe of our data was limited to the S&P 500 companies due to time constraint of this project and since the data collection process was a time consuming one.

4.2. Data Extraction

It is at this part where most of our troubles in the entirety of the project occurred. We began using data files saved in excel format but when reading these files into R, regardless of the package used in R to read the files, issues with converting the numbers to numeric values were complex to resolve. So complex, that we then resorted to using csv files as mentioned above. Even then, issues with data structure consistency, from the initial csv read into the R data structure, surfaced. After much trail and

error, and research on reading complex file structures into R, we were able to create data frames in R with similar initial data structure.

This was important because the initial data structure from the spreadsheet had rows of dimensions being reported by the companies and columns of varying time frames. For this project, we needed a way to guarantee that we could correctly identify and extract the data cell corresponding to the dimension and time frame we wanted. We could do this even if the columns and rows did not correspond in positions from company to company all because we had created data frames in R of consistent structure. Simply, all we did is identify the name of the dimensions and the time frames for the cells we wanted to collect for data analysis.

4.3. Data Cleaning

Once we had a method of extracting desired data from the raw data structure, we now began to clean the data. What we found is that many of the initial dimensions chosen to investigate had large variations in the number of missing values. Essentially, there were very few missing values from the dimensions in the Income Statement spreadsheet but many missing values in the Cash Flows and Balance Sheet. Due to the quantitative nature of our project, missing values cannot be processed by the UML algorithms and so we decided to handpick dimensions from all spreadsheets. Our handpicking criteria was that a dimension must not have any more than 3 missing value for a specific time frame for all the companies to be analyzed. This was to ensure that most of the S&P 500 companies were analyzed as originally planned.

5. Methodology – Phase Two

5.1. Generated Sectors

Now that we have our data collected and cleaned, we can being to use the quantitative data to create our generated sectors via UML techniques. The appropriate technique chosen for this project was cluster, hierarchical clustering to be exact. This is due to the innate ability for this technique to group objects in a fashion such that objects in the same group are more similar to each other than objects in other groups. This is exactly what we want because it is basically the quantitative analogy to the grouping method done by the qualitative taxonomies. Hierarchical clustering itself has two main algorithms: agglomerative and divisive. They both produce a clustering structure typically represented via dendrograms with the main difference in their starting condition. Agglomerative uses a bottom-up approach where each data point begins as its own cluster and then merging pairs as the algorithm iterates upward whereas divisive does the opposite and beings by placing all data points in a single cluster iterating downward. Both algorithms are already implemented in their appropriate packages.

Applying the R functions calling forth these algorithms, we can immediately produce primitive dendrograms. Note, practically all default settings for these functions were used. These primitive clusters and dendrograms are meaningless until we apply a criterion determined by our choice of benchmark taxonomy. We chose to use two GICS taxonomies as our benchmarks, GICS-Sectors and GICS-Sub-Industries. They have 11 and 158 partitions respectively. It is these numbers that we must apply to our algorithms such that we force that many number of clusters to be generated for our financial data set. Once the algorithms creates either 11 or 158 partitions, and subsequently the clusters, we have now established a balanced comparison. We would now have the same number of generated sectors as we have the number of groups in the benchmark.

5.2. Setting

Ideally, we would take these clusters and treat them as financial portfolios by characterizing them with the portfolio performance measures. Then one would assume each generated sector is compared to an individual benchmark sector but that would be wrong. Since our data set is only quantitative in nature, we have not nor can we bridge a single generated sector to a benchmark sector. Thus, we cannot do a comparison at the portfolio level. We can, however, go one level above the portfolio, which would be the group of portfolios, and use a five-number summary method to achieve comparison. This grouping of portfolios is what we define as setting since we generated that group of portfolios using specific experimental parameters.

The reason why the comparison at the setting level works for our project goal is because using a five-number summary we can still investigate the happenings of clusters formed using UML techniques. The experimental parameters that can be varied allows use to produce multiple settings and subsequently a large number of portfolios to be analyzed. The parameters we varied are: time frame, number of benchmark partitions, and the UML algorithm utilized. Of course, there are probably more parameters that can be varied and even still more options for the parameters that we did choose to vary. A more exhaustive project would include as many variations within the experimental procedure for best results. Nevertheless, due to time constraints, we choose to vary only three parameters and doing so allows us to create 552 portfolios and most notably 8 sectors to be analyzed.

5.3. Entanglement

But just being able to generate these portfolios and sectors is not enough to warrant further investigation. We need to make sure that each sector created and their set of portfolios are unique. If they are not unique then that means varying the experimental parameters leads to non-different results rendering the variations useless. We can do this analysis for uniqueness either visually or using a measure called entanglement. Entanglement rates the alignment between two dendrograms between 0 and 1 where 0 is no entanglement and 1 is full entanglement. Alignment essitnally being how the dendrogram structures mirror each other. For our case, an entanglement of 0 is unwanted as that would mean our variations to the experimental parameters are worthless. We are also limited to what variations we may measure with entanglement. We can only measure entangle for differing clustering algorthims, like hclust and diana, but not for other variations like number of partitions and time frame. The simple argument as to why we cannot compute entaglement for the later two variations is because those variations lead to either an uneven number of clusters or an uneven number of objects in the dendrogram. Regardless, it is obvious that the dendrograms created using a different number of partitions and number of ojbjects will result in unique dendrograms, accomplishing our goal. As for the entaglement mweasures between varying algorithms, we verify that these measures are non-zero and in summary, all of our sectors are unique and worth individual investigation.

6. Methodology – Phase Three

6.1. Portfolio Performance Measures

The way we investigate the mechanics of these sectors are through portfolios performance measures. These measues allow us to quantitative characterize the portoflios and in turn their sectors for comparison capabilities. By assuming each generated sector as well as the benchmark sectors to be

a financial portfolio, we simply have to compute a performance measure as one would for an ordinary portfolio. The measures chosen for this project are Sharpe and Treynor ratio, both for their ability to track return per risk. Their main difference is in what type of risk is used; Sharpe uses the standard deviation of the portoflio's excess return and Treynor uses the beta of the portfolio. Specifiallcy to the Treynor ratio, the portfolio used in the beta calculation is a chosen index where we have logically picked the SPY index to match our dataset's universe. Additionally, a 2.39% risk free rate for a three-month treasury yield, at the time of the project, was chosen for the risk free rate as we are looking at quarterly time frames.

6.2. Five-Number Summary

As mentioned before, though we can assign each portfolio with a performance measure we cannot necessarily compare a generated sector to a benchmark sector due to the lack of qualitative data connecting the two sector types. Thus we resort to the comparison at the setting level which is done via a basic five-number summary. The five-number summary provides a summary of the distribution of the observations, in this case our portfolios. This type of summary was chosen since our main goal is to investigate the efficacy of using UML techniques to create novel sectors which can be done by analyzing the generated sectors' constituents – the portfolios. With this summary we can identify if any setting performs better or worse than the benchmark setting and how often such events happen. Since the setting is defined as a group of portfolios, or sectors, we still are able to achieve our goal of assessing generated sector, albeit not individually but as a group. The five-number summary is easily graphically represented using a boxplot for effective visual comparison.

7. Results

Viewing the boxplots we can understand that for regardless of portfolio performance measure or number of partitions used, the generated sectors for 2018 quarter four out perform their benchmark counterparts and the opposite is true for 2018 quarter three. However, the argument of out or under performance is a rather extreme one as one can verify the scale of the vertical axis. The difference of the performance measures vary in the tenth place which is for the most part unrealistic. Usually these performance measures, especially the Sharpe and Treynor ratios, are valued for their non-negative integer vales. Another note is that much of our resulting number by applying the portfolio performance measures to our sectors are negative. For investors this is bad new since negative ratio for both Sharpe and Treynor indicate that investing in the risk-free rate is a better option. However, this project is more concerned with how the UML techniques influence the creation of sectors by comparing to a benchmark taxonomy which is still achieved. Ultimately, due to the scale of the numbers, we argue that UML generated sectors perform on par with regular benchmark taxonomies.

8. Conclusion

By creating an experimental procedure for testing UML techniques in their ability to sectorize financial companies we have uncovered underwhelming results. Our experimental procedure takes into account variations in the method used to create unique sectors which allows us to grasp the efficacy of using UML techniques with numbering evidence. However, a more robust experiment will entail the usage of countless more variations to fully understand how the UML techniques can be useful in sectorizing financial companies. Also, a much larger data set should be used to cover a large amount of the companies present in reality. Not only increasing the number of companies but also investigating

more time frames would better solidify the results. Tracking other performance measure will also be worth while for different type of investors.

In summary, our results were underwhelming. The generated sectors performed relatively on par with the benchmark sectors. This is due to the minute scale of changes in the performance measures of the portfolios. Too minute to realistic be valuable to financial experts. Yet we implore the continuation of this project on a much grander scale to fully grasp the efficacy of UML techniques to create financial sectors.

References

Paliouras, G., C. Papatheodorou, V. Karkaletsis, and C.d Spyropoulos. "Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques." Interacting with Computers 14, no. 6 (2002): 761-91. doi:10.1016/s0953-5438(02)00015-2.