USING NEURAL NETWORKS TO GROUP COMPANIES INTO CLUSTERS

Donald Joyner, Norfolk State University Norfolk, USA dtjoyner@nsu.edu

ANNs can be used for 3 tasks: classification of data, clustering of data, and pattern recognition in data. An ANN can be adjusted to improve results, but they have a "black box" characteristic to them. This refers to the hidden layer of ANNs and the fact that there is no way to explain how or why an ANN reaches a particular conclusion regardless of its accuracy. The proliferation of financial data in the form of XBRL perfectly coincides with the development of artificial neural networks. The goal of this study is to apply a neural network to quarterly XBRL data to derive clusters. A neural network can analyze a large amount of data and recognize patterns and based on those patterns classify companies into clusters.

The time covered in this study is quarterly financial information from 2012 through 2017 (22 quarters). This data will be input into an ANN in order to derive clusters based on similarities between companies. These clusters will be based on information other than SIC classification. If there are recognizable patterns the neural network should be able to categorize companies based on shared independent variables. If certain companies have similar values related to their independent variables an ANN should recognize it, refine it, and ultimately produce increasingly more accurate categorizations.

While clustering companies into categories isn't as powerful as actual stock price prediction, it is a good start. Nemes and Butoi (2013) concluded that neural networks can be used to develop strategies based on better pattern recognition. This creates a situation where a competitive advantage can be gained through the development of an ANN which ultimately produces better data for purposes of developing a better trading strategy than can be obtained using existing trading models. If an ANN groups companies into different categories, one could anticipate how a particular stock's price might change by observing the movement of stock prices related to companies in the same category. Often, investment analysts rely too much on past data, which doesn't necessarily incorporate non-numerical values. One way to increase prediction accuracy is the usage of ANNs. The clusters developed in this study are solely dependent on numerical values.

According to Krishnawamy, et al. (2000) "the lure of huge profits from their use makes neural network applications immensely popular in corporate finance and security markets." XBRL was made mandatory by the SEC in 2009. Ince and Trafalis (2007) provided an excellent overview of Artificial Neural Networks (ANN). ANNs work like the human brain. Like the human brain, ANNs have 'neurons' that analyze data. In regard to large amounts of data, ANNs are able to develop many 'neurons' which combined, form the basis for output (i.e. decision). When a human is exposed to a manageable level of information, it can detect patterns and filter out irrelevant variables (i.e. noise). However, the human brain can only process a finite amount of information. A human would not be able to handle the amount of data associated with XBRL and other financial data. In contrast, a computerized Neural Network *can* analyze large amounts of data and successfully detect patterns.

As has been noted, forecasting has traditionally been based on time series data analysis. Ince and Trafalis (2007) observed that the focus of time series analysis is based on charting stock prices and trading volumes along with other technical indicators that have been deemed relevant. The major obstacle encountered when following this approach is identifying all of the proper technical indicators. Ince and Trafalis (2007) note that over 100 indicators have been identified. It would be very hard to apply all of these indicators using traditional statistics-based models. This study uses over 60 different variables to sort companies into categories. A potential benefit of having clusters is that instead of applying time series analysis to the entire market or companies within the same SIC grouping, such an analysis could be applied to companies within the same cluster.

It also must be noted that a particular stock exchange does not exist in a vacuum. Stock prices can be influenced by world events and economic activity in one country, can affect multiple other countries. According to a study by Valadkhani and Chen (2014) events that occur in the US stock market can have significant effects on the output and volatilities of the UK, Canadian, and Australian stock markets. As a result, negative trends on the US stock market lead to negative trends in other markets. The best historical example of the effect the US stock market can have on the rest of the world is the Great Depression of the 1930's. It started with the US stock market crash in 1929 and ultimately enveloped most of the world for the next decade. This particular study doesn't use macroeconomic variables, only company specific information.

Stock exchanges are most influenced by company specific information such as quarterly and annual financial statements. Annual financial statements in the United States are audited and are viewed as the most comprehensive and reliable financial information provided by companies. Financial information should have a direct impact on stock price. Menaje (2012) reminds us that "the importance of the share price cannot be overemphasized" because it is of prime concern to analysts, investors and lenders. If two companies are placed into the same cluster, and these companies do indeed share similar traits, it could mean if one company in a cluster is undervalued, another company within the same cluster may also be undervalued.

There are two ways to make money in the stock market. Since a share of stock represents a person's relative proportion of ownership of a corporation, they are entitled to their share of the dividends paid by a corporation. The entire purpose of a corporation is to earn money. Therefore, firms return profits to its shareholders in the form of dividends. Dividends are not always the same nor are they always paid every year. There can be fluctuations. Many investors can be classified as "long-term." Their intent isn't to sell their stock investments. Many individuals live on the dividends paid through their stock portfolios. Further analysis would have to be performed but the usage of clustering could impact the way one develops their stock portfolio. Rather than diversifying a portfolio based on a company's industrial classification, diversification may be performed based on cluster category.

Menage (2012) indicates that for a "short-term" investor: "Any movement of the share price can mean income or loss to the investor." Since a "short-term" investor doesn't plan on retaining their shares for an extended period, every price fluctuation in the market prices of their stocks represents either a gain or a loss. Stocks that are perceived to be increasing in price are sought after while stocks perceived to be decreasing in price are avoided. The end result is the creation of price volatility. Andersson and Gärling (2009) elaborate on this concept. Since stock price prediction isn't an exact science, investors are strongly influenced by consistency. As has been noted, stock price prediction models rely on time series data which take the form of stock price observations over time. If the price predicted by the prediction model is proven accurate it is perceived as consistent. An area of future study would be analyzing whether price volatility can be explained by clustered categories developed by an ANN.

Closely tied to share price, is financial information. Financial information is the primary means by which investors not only evaluate individual companies, but also the primary means by which investors compare potential investments. Often, investment decisions are based on net income performance. However, cash flows are also very important when doing company analysis. Net income is based on the accrual-based method of accounting (which is the method required by United States Generally Accepted Accounting Principles (GAAP)). Therefore, while net income is a very important number, it is also important to understand whether a company has enough cash on hand to meet its debt obligations as well as its ability to pay cash dividends. An analysis of companies based on clusters derived from ANNs could identify variables which play a greater role on stock price performance than previously thought.

Indeed, stock pricing is not an objective process based on concrete and precise measures. No less an authority than Keynes believed that irrationality plays a significant role in stock pricing. The consequences of this environment, as noted by Baker, et al. (1995), is that it adds an extra layer of complexity to the way companies raise capital. Specifically, as it currently operates, the market makes it more difficult for less established companies to issue new stock because the price at which they issue the stock could be undervalued. On the other hand, however one could use stock performance history of companies within the same clusters to see how the cluster responds to events and based on such responses, predict whether investing in a company in a particular cluster is a good idea or a bad idea.

ANNs operate in a manner similar to a decision tree. For each variable entered into an ANN, a neuron is created to analyze it. The ANN then organizes and interconnected each neuron into a hierarchy. A static statistical prediction model doesn't necessarily 'learn' in the sense that it is capable of recognizing patterns the way a human would. ANNs can be trained to learn. ANNs operate in a nonlinear manner and are composed of three layers. The first layer is the input layer, the layer that addresses input variables. ANNs then organize these inputs into neurons which together, form a neural network's architecture or 'intuition.' This is referred to as the hidden layer. If a human is confronted with repetitive problems it must solve, over time they will develop intuition based on experience. Experience will teach the human brain how to analyze each factor relevant to the decision and assign weights to each factor. However, it would be very difficult, if not impossible for a human to explain and/or quantify how they arrive at decisions. The same applies to ANNs. That is why 'intuition' is referred to as the 'hidden layer.' Unlike statistical techniques, ANNs aren't capable of explaining the process used to arrive at decisions. Decisions are referred to as the output layer. As ANNs become more advanced, they will develop better 'hidden layers' and by extension better decisions. The clusters developed in this study cannot be designated by any specific name such as 'manufacturing' or 'service' etc. The only thing that can be said about each cluster is that it exists. It would be almost impossible to offer any further elaboration.

ANNs have been developed to predict bond ratings and identify companies facing financial distress. They can even be used to detect potential fraud. ANNs are commonly used by businesses to make decisions whether or not to grant credit. Krishnaswamy, et al. (2000) identify the most powerful characteristic of ANNs are their ability to analyze "massive unstructured and complex data sets." In a conventional statistical model, adding too many independent variables to a prediction model can create poor predictions due to co-variance amongst those variables. Far more independent variables can be added to an ANN. However, as has been noted, ANN's learn like a human brain and must be 'trained' to observe patterns and predict results. Whereas co-variance isn't a problem faced by an ANN, it is possible to 'over train' an ANN or to be more technical, run it through too many iterations of training. This can adversely impact predictive accuracy. Krishnaswamy, et al. (2000) identify 4 situations where the "heuristic" nature of ANNs are best suited:

1)one can specify particular influences on a phenomenon whose outcome is known with certainty
2) the relationship cannot be described
3) the relationship is not necessarily linear
4) there are no known models

Kiranyaza et al. (2009) calls ANNs "universal approximators" because they constantly work on a trial and error basis where interconnected groups of neurons work together in solving a specific task. Qing, et al. (2011) reminds us that ANNs have already been employed in multiple business fields and have "consistently outperformed other, more traditional quantitative forecasting methods." ANNs are only able to provide output. This means they don't explain how they developed said output (i.e. hidden layer). Qing, et al. (2011) indicates that within the hidden layer, ANNs are constantly adjusting the weights of coefficients in an ongoing effort to improve accuracy. ANNs are capable of evolving as they recognize and adjust various patterns. According to Krishnaswamy, et al. (2000) ANNs "possess learning ability" and when applied to detailed databases are "potent tools for forecasting and analysis." An ANN can train itself and improve its accuracy with minimal input from humans, meaning it is intuitive.

This study used data derived from 21 quarterly XBRL data releases. The neural network was asked to classify companies into one of eight groupings. It should be noted that each quarter was treated individually, meaning that the neural network wasn't 'trained' over 21 quarters. If it had been, the results would likely be different. For the sake of simplicity, a cumulative learning process was not used. The following chart is a summary of the results:

Category	Average Percent	Standard Deviation
1	14.18%	7.61%
2	18.56%	6.88%
3	10.58%	8.37%
4	6.88%	6.92%
5	8.58%	5.32%
6	13.14%	8.98%
7	12.63%	6.86%
8	15.46%	4.94%

This study is intended to be an introduction to one possible benefit that ANNs can provide, which is the clustering of companies into groups. There are multiple ways clustering can be used to analyze stock performance. This study does not incorporate the usage of historical stock prices, but it is highly probable that a time-series analysis based on stock price and cluster category would yield useful information in regard to understanding individual companies and predicting future stock performance.

References:

Andersson, M., Hedesström, T. M., & Gärling, T. (2009). Social influence on predictions of simulated stock prices. *Journal Of Behavioral Decision Making*, *22*(3), 271-279.

Baker, M., Stein, J. C., & Wurgler, J. (2003). When does the market matter? Stock prices and the investment of equity-dependent firms. Quarterly Journal of Economics, 118(3), 969-1005. doi:10.1162/00335530360698478

Ince, H., & Trafalis, T. B. (2007). Kernel principal component analysis and support vector machines for stock price prediction. *IIE Transactions*, *39*(6), 629-637.

Kiranyaza, S., Ince, T., Yildirim, A, and Gabbouj, M. (2009). Evolutionary artificial neural networks by multi-dimensional particle swarm optimization". Neural Networks, 22(10), pp. 1448-1462.

Krishnaswamy, C. R., Gilbert, E. W., & Pashley, M. M. (2000). Neural Network Applications in Finance: A Practical Introduction. Financial Practice & Education, 10(1), 75-84.

Menaje J. (2012). Impact of Selected Accounting and Economic Variables on Share Price of Publicly Listed Banks in the Philippines from 2002-2008. DLSU Business & Economics Review. 22(1):35-62.

Nemes, M.D., & Butoi, A. (2013). Data Mining on Romanian Stock Market Using Neural Networks for Price Prediction. *Informatica Economica*, *17*(3), 125-136.

Qing, C., Parry, M. E., & Leggio, K. B. (2011). The three-factor model and artificial neural networks: predicting stock price movement in China. Annals of Operations Research, 185(1), 25-44.

Valadkhani, A., & Chen, G. (2014). An empirical analysis of the US stock market and output growth volatility spillover effects on three Anglo-Saxon countries. International Review Of Applied Economics, 28(3), 323-335.