# Comparative Analysis of Hedge Funds in Financial Markets using Machine Learning Models

Saurabh Aggarwal
Computer Science Graduate,
Software Developer,
Financial Markets Researcher,
New Delhi 110026, India

## ABSTRACT

The aim of this paper is to layout machine learning models for comparative analysis of hedge funds in financial markets for investments. It has been difficult to compare the hedge fund's performance due to myriad of classification techniques adopted by hedge fund managers such as strategy model, asset class, liquidity score, hence producing disparity in numbers. The machine learning and deep learning neural network models have been effective in exploiting the exogenous and complex data interactions in financial domain datasets and producing useful insights. The author in this paper discusses models such as – semi supervised learning, decision tree learning and hybrid time series classification along with experimental results to juxtapose the hedge funds and hence producing useful results for investments. The analysis shows that the discussed models in the paper can be used for comparative analysis of funds and the hybrid time series classification model is more effective rather than using the semi-supervised and decision tree models individually.

## General Terms

Hedge Funds, Liquidity, Strategy, Finance, Risk Profile

## Keywords

Deep Learning, Machine Learning, Artificial Intelligence, Hedge funds, Financial Markets, Decision Trees, Time Series Classification

## 1. INTRODUCTION

The term hedge fund in financial markets is defined as an investment fund that aggregates the capital from independent investors, firms and institutions and thereby further invests in variety of assets using risk profile analysis and portfolio selection techniques [1]. The comparative analysis of hedge funds is important from investment point of view. There has been complete reliance on databases and primitive schemes for bringing out hedge fund's performance while constructing a strategy model in financial domain, which makes it difficult to compare results as there are different ways to classify funds. The hedge fund databases provide description of investment styles, periodic return, fee structure, investment size which are not enough for prediction and comparison analysis. There has been need for unified models for classification that include most of the financial derivatives. The machine learning and deep learning models have made a remarkable impact in artificial intelligence field, for providing useful results from huge data sets with different input parameters. Deep Investment techniques such as neural networks, smart indexing, and auto-encoding have been helpful in credit risk analysis and portfolio selection. Models such as semi-supervised, decision tree and time series classification are useful for classification and can be effective in financial domain.

## 2. RELATED WORK

**[2] Hedge Fund Classification using K-means Clustering, Nandita Das, 2003:** It has been shown that the hedge funds can be classified using k-means clustering method of machine learning. The parameters for classification are taken as size of hedge fund, fees, and investment strategy and asset class. The classification results are compared with US and Non-US of ZCM or with hedge fund databases computing k-means to be an effective technique.

**[3] Pattern extraction for Time Series Classification:** The paper discusses that the useful results can be brought about by using pattern extraction from time series classification technique in machine learning. Artificial and real time experiments are also done to highlight the interest of approach for accuracy for the classifiers.

**[4] Deep Learning in Finance, 2016:** The paper analyzes the ability of the neural networks to solve the complex problems in financial markets for investments. It further explores the use of hierarchical deep learning models such as dropout, deep feature policy for solving financial domain problems.

**[5] A Comparison of Machine Learning Classifiers Applied to the Financial Datasets, 2010:** The paper analyzes some machine learning techniques and compares their classification accuracy. The algorithms compared are Naive Bayes learning model, feed forward neural networks model, and decision Trees learning. The analysis is done over two datasets-Japanese and European companies and about 59 financial attributes are used for them. It has been concluded that the decision trees give the best classification accuracy.

[**6] Machine Learning for Financial Market Prediction:**, the paper discusses about the use of machine learning techniques for the prediction of financial time series problems. Standard vector machines (SVMs), Relevance Vector machines, and neural networks are found to be performing well as compared to change point models and hidden Markov models.

## 3. COMPARATIVE ANALYSIS

### 3.1 Semi-Supervised Learning Model

A self-labeling or a semi-supervised learning model falls between supervised and unsupervised learning models that make use of unlabeled data set for training [7]. It is helpful in scenario where unlabeled data sets are far more than the labeled data set. A semi-supervised learning model uses at least one of the assumptions - a) Smoothness assumption: here the points close to each other are more likely to share a same label. b) Cluster assumptions: points lying in the same cluster as formed by the data are likely to share same label [8]. The hedge fund databases vary on the classification scheme and the type of hedge fund they include. The data-set returned from these databases is unlabeled in majority along with some labeled set. Hence, the self-labeling or self-training of hedge funds will be required for their classification and hence for comparative analysis. Consider the data set {F1, F2, F3… $F_M$ } for M hedge funds, where N can be time steps. For instance, one time step is equivalent to one month. Let F[x] is defined as return vector, to which each hedge fund is mapped. Hence Equation (1) represents the hedge funds as:

$$F_1[x_1, x_2, x_3, x_4, x_5 \ldots\ldots x_N] \quad F_2[x_1, x_2, x_3, x_4, x_5 \ldots\ldots x_N]$$

$$F_M[x_1, x_2, x_3, x_4, x_5 \ldots\ldots x_N] \qquad \text{- Equation (1)}$$

Let Q, be the quarter range hence, the return vectors can be distributed to collection of four months each representing quarter. Then,

$$\{F_1[x_1, x_2, x_3, x_4]\}_1 \ldots\ldots \{F_1[x_{N-3}, x_{N-2}, x_{N-1}, x_N]\}_k$$

$$\{F_2[x_1, x_2, x_3, x_4]\}_2 \ldots\ldots \{F_2\,[x_{N-3}, x_{N-2}, x_{N-1}, x_N]\}_k$$

Here, (k) is defined as the total range for (N) time steps, each equal to 4 monthly returns or can be defined as Quarterly returns [Q1, Q2, Q3, Q4 ]. Then, (M) Hedge funds will be defined as following:

$$\{F_M[x_1, x_2, x_3, x_4]\}_1, \{F_M[x_{N-3}, x_{N-2}, x_{N-1}, x_N]\}_2$$
$$\{F_M\,[x_{N-3}, x_{N-2}, x_{N-1}, x_N]\}_k \text{ - Equation (2)}$$

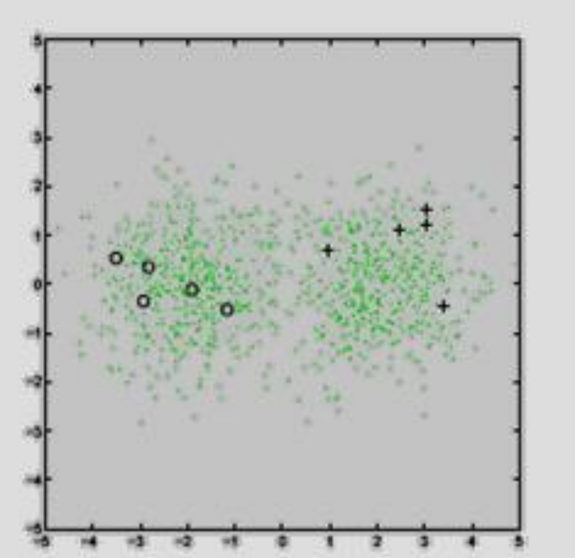


**Figure 1(a): Labeled (big dots) and Unlabeled (small dots) data set representation of hedge funds**

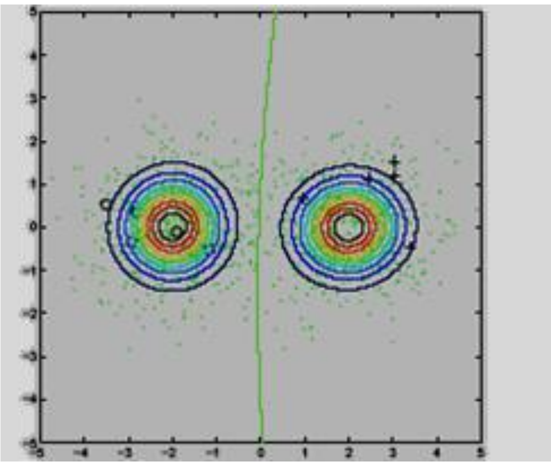


**Figure 1(b): Model Learned from labeled and unlabeleddata-set of hedge funds.**

For experimental analysis the hedge funds labeled groups are plotted along with the unlabeled data from the database. Figure 1 (a), shows the distribution of labeled as well as unlabeled representation of funds. Now assuming each fund class set has Gaussian distribution, it is plotted as in Figure 1(b). 50 hedge funds were taken in consideration from database for time range (2014 -2016), of which classified funds namely-SeCap(x), Point(x), Pela(x), Sha(x) and 40 others were labeled and hence classified appropriately (actual names have been masked accordingly). Hence, the raw hedge fund data as fetched from the database can be labeled appropriately and can be used for comparison based on their quarterly returns and other factors. The accuracy level for the classification this calculated here is 60.33% when compared with actual values from the database.

## 3.2 Decision Tree Learning Models (DTLMs)

### *3.2.1 Architecture*

A decision tree is defined as a tree based learning model that is used mostly in classification problems of machine learning.[9] The categorical and continuous input as well as output variables can both use this model. Here, the data set can be split into at least two or more homogenous data-sets based on the deciding factor from the input variables.

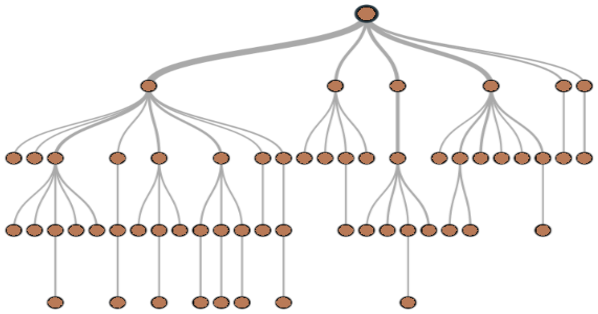
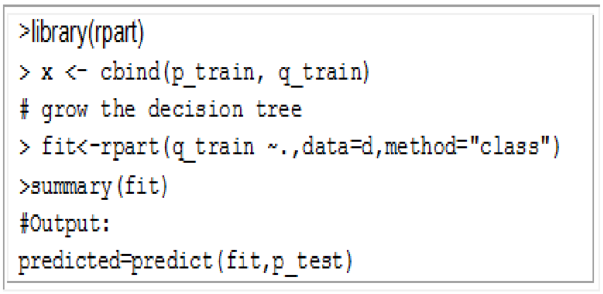


**Figure 2: A decision tree model performed on sample data-set**

In R framework, multiple packages are available to implement a decision tree such as- (Rpart). Following code generates a tree where p_train is the independent variable, q_train represents the dependent variable, and d is the training data while p_test is the output data.



```
>library(rpart)
> x <- cbind(p_train, q_train)
# grow the decision tree
> fit<-rpart(q_train ~.,data=d,method="class")
>summary(fit)
#Output:
predicted=predict(fit,p_test)
```

**Figure 3: Code to generate a tree in R framework using r-part library and training data set**

The decision trees can be regression trees or the classification trees depending upon the variable which can be categorical or continuous. Both these trees follow the top-down greedy approach which is called recursive binary splitting. The predictor space is split in two new branches down the tree as the tree is traversed from top to bottom. They are popular as being greedy because the algorithm cares about the current node split without any consideration to the future, thereby making the better tree.

### *3.2.2 Decision Trees for Hedge Funds*

The decision trees can be very impactful in financial domain problems. They can exploit the insights and complex dependent structures of financial data set to give a meaningful output. The nodes of the tree represent classification while the branches are the features that lead to those classifications. For hedge funds, each time slice represents a decision tree[10]. The nodes of the tree can have one or more records while each record represents a hedge fund. The nodes that have less records show better classification for the decision trees. The duration of time slices along with the number of hedge funds considered controls the number of nodes generated for the tree [11]. The input parameters play an important role in deciding the classification accuracy. The parameters such as-complexity penalty and minimal support affects the growth of the tree and it's split at lower levels. The computation time and growth is inversely proportional to the complexity penalty. Controlling the growth of tree favors the performance while it affects the classification accuracy. For the financial trees, the predictor value of the hedge funds is compared with other predictor values for the tree-nodes.

For experimental analysis for data set {F1, F2, F3… $F_M$ } for M hedge funds , equation (1) of section 3.1 can be the feature values for the financial tree while equation (2) values can be the predictor values. 50 Hedge funds were considered for analysis for the time range (2014 to 2016) equal to 8 quarters or 32 months. For each quarter a decision tree was made, hence a total of 8 decision classifying trees were made based on the investment return of the hedge funds (considered in $).

Some business rules were considered in generating the decision tree for figure (4). The rules defined on the basis of quarterly returns of the hedge funds.

The group of hedge funds gets further divided to sub-group based on the rules. For instance, for the main node, 50 hedge funds were sub-divided into 26 and 24 on the basis of rules a) 2016_(Q2) < - 5.18 || > = 1.9 b) 2016_(Q2) < -5.1 && < 1.9. Here, || represents or operator and && represents, and operator.

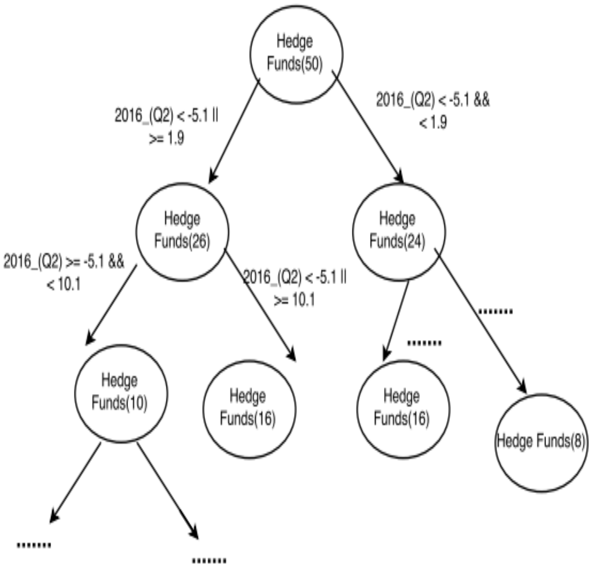


**Figure 4: Decision Tree for 50 Hedge Funds considered for 8 quarters between Years 2014 to 2016.**

When the computations for decision tree classifications get completed, all the funds falling in one category nodes can be considered for similarity. As per the results, node (a) got 3 hedge funds; node (b) resulted in 2 funds and so on thereby classifying the funds for similarity and their comparative analysis. The accuracy level for the model comes out to be 69.93% for the classification when compared with values from hedge fund database.

## 3.3 Hybrid Time Series Composition model

A signal composition or a time series model is used when a number of observations recorded for a given time are to be considered from data-set using time steps. The model can be used as - a) Univarate model: where only single variable is observed over the time range from data set [12] and b) Multivariate model: where more than one variable is considered from the data set which is mostly used for complex applications [13]. To track similarly behaving hedge funds for a given time frame, a time series composition model can be used. For experimental analysis consider the data set {F1, F2, F3… $F_M$} for 'M' hedge funds. The time series composition model for hedge funds uses both the self-training and decision tree learning models. Using self-labeling in section 3.1 equation (2), label L (t) is generated as following:

$$L\ (t)\ _1 = \sum_{l=1}^{Q} F_1(x_1)$$

$$L\ (t)\ _2 = \sum_{l=1}^{Q} F_2(x_2)$$

$$L\ (t)\ _M = \sum_{l=1}^{Q} F_M\ (x_N)$$

L (t) generated is assigned to the time series for each of time slice. Now, generate decision trees for this labeled data set of hedge funds using decision tree learning model in section 3.2. The resultant funds are then plotted with x-axis as time series and y-axis as percentage of total returns. Hedge funds SeCap(x), Point(x) and Sha(x) from the data-set have been plotted and compared, where actual names have been masked.

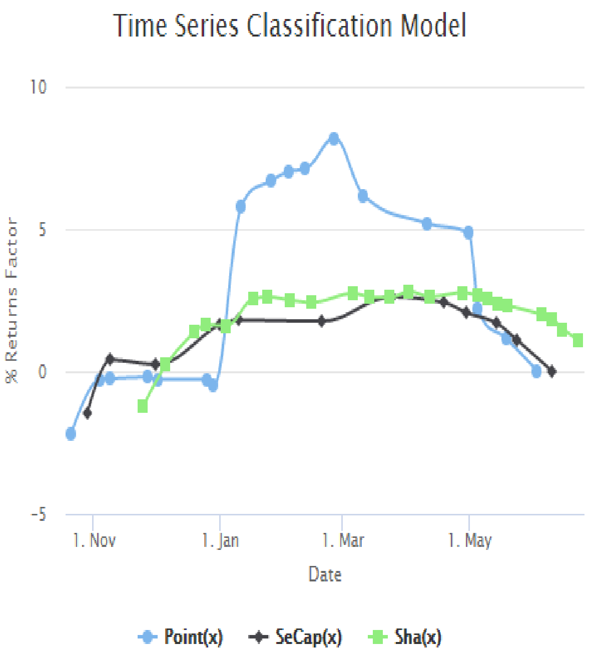


**Figure 5: Time series Classification model for Hedge Funds Point(x), SeCap(x), Sha(x)**

Pearson Product-Moment Correlation Coefficient (r) values [14] between the hedge funds are then evaluated. A higher correlation coefficient value (r) indicates that there is a high correlation between the time-series classified representing the hedge funds.

**Table -1. Pearson Product-Moment Correlation Coefficient (r) values for Hedge Funds**

| Hedge funds | SeCap(x) | Point(x) | Sha(x) |
|---|---|---|---|
| SeCap(x) | 1.00 | 0.55 | 0.78 |
| Point(x) | 0.55 | 1.00 | 0.69 |
| Sha(x) | 0.78 | 0.69 | 1.00 |

As from the Table 1, high values of (r) were observed for SeCap(x) and Sha(x), thereby computing them to be similarly behaving hedge funds. Likewise, remaining funds from the set were evaluated accordingly. The accuracy level so computed by using all the three modules consecutively in hybrid time series was 81.44% as compared to values from hedge fund database.

## 4. CONCLUSION

The author demonstrated the experimental conclusions for the comparative analysis of hedge funds in financial markets for investments using machine learning models. The models along with their implementation are hence explained in the section 3 for classifying hedge funds from the raw data set. The classification accuracy for semi-supervised learning model and decision tree models individually were found to be less as compared to hybrid time series composition model which was 81.44% in contrast to 69.93% and 60.33%.

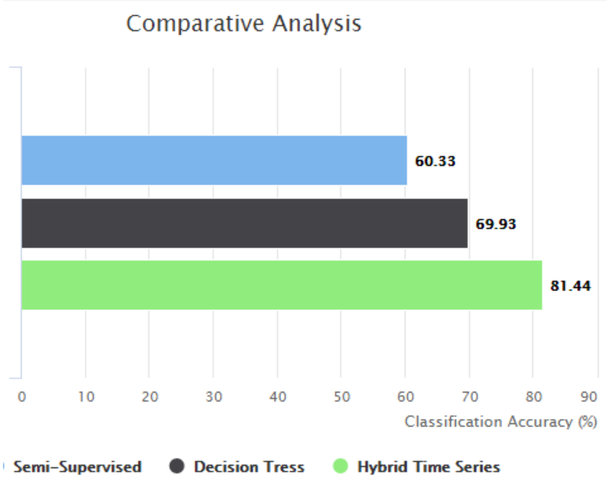


**Figure 6: Comparative Analysis of Hedge Funds using machine learning models**

The author suggests the use of semi-supervised learning model and decision trees combined with time series composition model to get better results for solving classification problems in financial domain. Hence, the discussed machine learning models can be useful in classifying hedge funds for investments and analysis. The machine learning model found to be most accurate for comparative analysis of hedge funds can be further used for classifying securities in future from the complex data set for portfolio management and recommendations. Risk analysis and investment analysis can be further done on the basis of the experimental conclusions.

## 5. FUTURE WORK

The next steps in the future will be to discuss and analyze further more sophisticated data dependencies in financial domain such as risk analysis for mortgagee using deep learning and machine learning techniques. The machine learning models presented in this paper along with other deep learning models will be applied to mortgage backed securities data set to uncover the risk involved in the investment.

## 6. REFERENCES

[1] Wikipedia, What is a hedge Fund, https://en.wikipedia.org/wiki/Hedge_fund

[2] Hedge Fund Classification using K-means Clustering method, Nandita Das, 2003, 9th International Conference on Computing in Economics and Finance.

[3] Pattern extraction for Time Series Classification, Pierre Geurts, University of Liege, Department of Electrical and Computer Science.

[4] Deep Learning in Finance, NG Polson, JH Witte, JB Heaton, Feb 2016, University of Chicago

[5] A Comparison of Machine Learning Classifiers Applied to Financial Datasets , Pablo D. Robles-Granda and Ivan V. Belik, Proceedings of the World congress on Engineering and Computer science, 2010

[6] Machine Learning for Financial Market Prediction, Tristan Fletcher, PhD Theses, University College London, Computer Science

[7] Wikipedia What is a Semi-supervised learning https://en.wikipedia.org/wiki/Semi-supervised_learning

[8] Introduction to Semi-supervised learning, MIT Press release, Semi supervised learning introduction.

[9] Wikipedia, What is decision tree learning, https://en.wikipedia.org/wiki/Decision_tree_learning

[10] White paper: A method for comparing Hedge Funds Uri Kartoun, Washington DC, USA

[11] Machine Learning: Decision Trees, CS540, Jerry Zhu University of Wisconsin-Madison

[12] Deep Learning Architecture for Univariate Time Series Forecasting, Dmitry Vengertsev CS229 Technical report 2014.

[13] Multivariate Time Series Classification with Temporal Abstractions Iyad Batal, Lucia Sacchi, Riccardo BellazziMilos Hauskrecht, Proceedings of the Twenty-Second International FLAIRS Conference (2009), Department of Computer Science, University Of Pittsburg

[14] Wikipedia, what is Pearson Product-Moment Correlation Coefficienthttps://en.wikipedia.org/wiki/Pearson_correlation_coefficient.