ISSN 2277-2685

IJESR/Feb 2021/ Vol-11/Issue-1/1-8
Chilakala Hari Krishna et. al., / International Journal of Engineering & Science Research

RETRIEVING TOP WEIGHTED TRIANGLES IN GRAPHS

[1]Chilakala Hari Krishna, [2]Dr.Gudivada Vijay Kumar, [3]Bathina Bharathi

[1] Associate Professor Department of CSE, RISE Krishna Sai Gandhi Group of Institutions,[2] Professor Department of CSE,RISE Krishna Sai Gandhi Group of Institutions, [3] Assistant Department of CSE,RISE Krishna Sai Gandhi Group of Institutions

ABSTRACT

Many network analysis tasks use pattern counting in graphs as a fundamental primitive, and there are several ways to scale sub-graph counting to large graphs. Although existing scalable methods for pattern mining are built for unweighted graphs, many real-world networks have a concept of the strength of connection between nodes, which is frequently described by a weighted graph. Here, using the generalised mean of the triangle's edge weights, Authors create deterministic and random sampling methods that make it quick to identify the 3-cliques (triangles) with the largest weight. For instance, one of our suggested algorithms may, in a reasonable amount of time on a commodity server, discover the top 1000 weighted triangles of a weighted graph with billions of edges, which is orders of magnitude faster than existing "fast" enumeration techniques.

Key Words:

I.Introduction

The understanding of the complex network structure depends on small sub-graph patterns, often known as graphlets or network motifs. The triangle (3-clique) is one of the most basic non-trivial sub graph patterns, and the fundamental issue of counting and enumerating triangles has received extensive theoretical and practical research.

Triangles are widely used in graph mining applications such as community detection, network comparison, representation learning, and generative modelling, which has contributed to the focus on them.The social sciences also make considerable use of triangle-based network statistics like the clustering coefficient.

The majority of algorithmic research on scalable triangle counting or enumeration concentrates on unweighted graphs. However, the edges of many real-world network datasets naturally have a sense of weight linked to them. Edge weights, for instance, can be used to reflect co-occurrence counts in projections of bipartite networks, traffic flows in transportation networks, or tie strength in social networks. These edge weights provide more information about the network structure. Edge weights can also improve the kinds of small subgraph patterns that are used in analysis.

A triangle is given, a weight that is calculated from the weights of its constituent edges in these instances, for example, when the network clustering coefficient has been generalised to account for edge weights. The weight of a triangle, roughly speaking, is often some combination of the minimum, maximum, arithmetic mean and geometric mean of the triangle's edge weights.

We Need effective techniques for retrieving triangles in large weighted graphs because existing triangle enumeration algorithms do not scale to massive graphs for these problems. We address the issue of counting the top-weighted triangles in a weighted graph in this study.

Let G = (V, E, w) be an undirected, simple graph with positive edge weights. The p-mean of the edge weights is what we use to define the weight of a triangle (i,j, k) with the edge weights W_{ij},W_{jk},W_{ik}.

II. Literature Review

A lot of research was done on this survey in fields like Deep Learning, stock markets, factors affecting stock prices, etc. Many studies have shown that stock prices can be predicted by taking a variety of factors into account.

The study by Ashish Sharma and his colleagues [6] provided a theoretical framework for predicting stock prices using regression models. The study explained what regression models are and how they can be applied to price prediction. The goal of this study was to provide some insight into how stock price influencing factors can be chosen as variables to offer some stock price predictions.

According to Amit Kumar Sirohi's study [7], the first tier provides information about feature selection, such as opening price and closing price, and the second tier builds kernels based on the extracted features. The most accurate value was predicted using appropriate features. As a result, we gained a fair understanding of how to apply suitable features to models.

Pushkar Khanal and Shree Raj Shakya [4] published a study that found Support vector machines gave the best results for prediction with a high level of accuracy. As a result of this study, we were able to understand how classification can be used to make predictions.

According to Yaojun Wang and Yaoqing Wang [5], social media mining helped explain the effect of stock comments on stock prices. Incorporating this factor with other important factors led to more accurate results. In this study, SVM with an emotion index was applied.

Different techniques have been used by researchers for feature selection from data sets with a large number of features. Nowadays, machine learning researchers are recognizing the importance of feature selection for analysing data because data with high dimension not only affect the learning models but also increase computational time and are considered as information poor [9]. Moreover, due to the large number of features, we face the curse of dimensionality, which states that in space of high dimension, data turn out to be sparse [10].

To solve the problems that arise from high dimensional data, researchers use two approaches: feature extraction and selection. In the first approach, new feature space with low dimensionality is created while in the second approach, redundant and irrelevant features are removed and a small subset of more relevant features is selected.

Feature selection has been done increasingly with swarm intelligence (SI) algorithms [11]. The reason is that the technique is popular for solving various optimization problems and finding optimal features is definitely this kind of problem. SI algorithms were very popular in recent years, and nowadays, its two popular algorithms are PSO and ant colony optimization (ACO).

Some researchers also used hybrid approaches for feature selection by combining individual techniques. For example, combined slap swarm algorithm with PSO and found an enhancement in performance and accuracy [12]. Similarly, also proposed a hybrid approach by combining genetic algorithms [13] and PSO and found that the proposed approach was capable to obtain accurate classification. [14] Compared three techniques for feature selection namely, kernel-based principal component analysis (PCA), fuzzy robust PCA, and PCA in order to reduce 60 economic and financial features in the data to forecast S&P 500 Index. They found that PCA gave slightly higher accuracy performance than the other two techniques.

III. Objective of Research Work:

Over the past decades, several researchers have discussed and refreshed machine learning algorithms related to stock price prediction. There have been a number of recent publications that have been surveyed. In addition, there have been some problem statements that have been found to help solve the issues left unresolved by researchers.

- It consists of adapting global prediction rules to implement multiple datasets for balancing scalability of stock price prediction, and performing various experiments to find out the impact of features in stock markets, like social media and financial news.
- In order to predict stock price, you must explore and select the features consists in each dataset. Features play a key role in making the prediction.
- Reducing the dimensionality of the dataset is key, because if the dataset has too many dimensions, it is difficult to predict. The selection of appropriate algorithms for reducing is also crucial.
- Separate the dataset into train and test data, it consists various split ratios like 90:10, 80:20, 70:30, 60:40 and 50:50. When creating any model, training data must consist more than test data. Choose appropriate split ratios.

- A great number of algorithms are presented in machine learning and deep learning for classification and regression problems. Choose or create an algorithm based on the dataset since all algorithms do not always perform well.
- Hyper-tune the parameters of the algorithm to improve model accuracy. Determine the best metrics to evaluate classification models or regression models.

IV. Methodology:

This section describes the individual steps performed in our framework for stock prediction. Our framework includes basic steps. The steps are as follows.

1. The raw data used are from various datasets.
2. The attributes used for feature extraction are the 'date' and ‘closing price’ of a stock.
3. Among the factors used to predict the momentum of the stock price of a particular company are 'stock momentum', 'index volatility', and 'sector momentum'.
4. The dataset is split into a training and test dataset.
5. The training dataset is used for model training, and the test dataset is used for prediction. The significance of a feature is determined based on R2 values
6. The values of the test data are predicted, and the results are evaluated. The result is given on the basis of accuracy, confusion matrix, and time required for the model used.

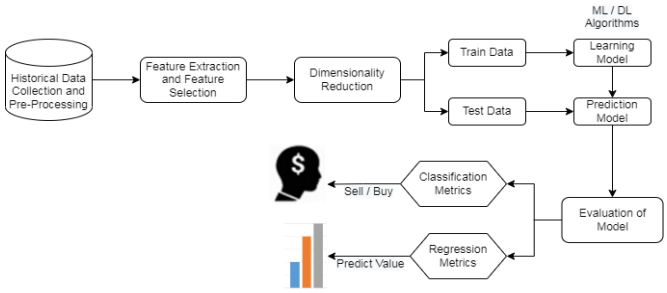


Fig. 1. Work flow of research proposal

1. Data Collection

This subsection describes the data collection process, sources of the collected data, and structure of the collected data. The stock markets that are selected as case study for this research and their tweets and news counts are given in Table 1. In this table, the stock exchanges show the overall stock markets while the other stocks are stock markets of individual companies. Note that the stock market terminology will be used interchangeably to refer to the stock market. Stock market, social media, and news data of the selected stock markets and S&P 500 index price data are gathered for 2 years of individual companies and overall markets.

Table 1. Stock markets symbols and tweets, news count summary

2. Data Preprocessing, Feature Selection and Dimensionality Reduction

Currently, the data is available in raw format, which makes analysis impossible. This data includes the highest value, lowest value, opening value, closing value, and volume of traded stocks on a particular date. The date and closing value of the stock are the most significant to us out of these. Based on a stock's closing value, we calculate momentum and volatility for each company, sector, and index. There are three types of momentum in the Yahoo dataset: stock momentum, sector momentum, and index momentum. A company's volatility, a sector's volatility, and an index's volatility are also taken into account. Every company undergoes this process.

3. Data Classification and Prediction

The first step in building a machine learning model is to obtain an optimal dataset. The open soured data which is available on the internet consists of many discrepancies like having missing data, having repeated rows of the same data, data being unstructured etc. Before feeding the data to the machine learning model, the data needs to be modified or preprocessed. This is so that the model is able to deliver results that are as accurate as possible.

The main attributes that are found in financial datasets (historical data about stock prices of a particular company) are as follows:

1. Date of that particular stock price
2. Opening stock price
3. High stock price (highest value of that stock during that day)
4. Low stock price (lowest value of that stock price during that day)
5. Closing stock price
6. Volume of stocks traded

Among all these parameters, the closing price is used primarily to feed the model. Various regression models available in machine learning can be used to predict the future stock price of a company based on this single value. In regression, curves are plotted on graphs based on input. The curve shows how stock prices have fluctuated over time. On the X-axis, we display the date of the stock, and on the Y-axis, we display the closing price.

The Classification algorithms used for data classification are Bayesian Classifiers, Decision Trees (DT), Logistic Regression, Support Vector Machine (SVM), Neural Network (NN), Multilayer Perception (MLP) Neural Network, Fuzzy Neural Network (FNN), k-Nearest Neighbourhood (KNN), and Genetic Algorithm

(GA) etc. Depending on the system requirements, we select an algorithm and then perform machine learning and optimization.

The Regression algorithms used for prediction of approximation values are 1. Simple Linear Regression, Polynomial Regression, Support Vector Regression (SVR), Decision Tree Regression, and Random Forest Regression etc. Depending on the system requirements, we select an algorithm and then perform machine learning and optimization.

4. Expected Outcome:

The surveyed systems focus on the prediction of stock price and its discrimination from machine learning and deep learning techniques, which can be made possible with the help of best parameters and experimental set-up design to execute the algorithms. The performance of the classification systems are calculated with the help of metrics namely, Accuracy, precision, recall, and F1-measure. All these parameters can be calculated using confusion matrix.

$$Accuracy = \frac{Numberofcorrectpredition}{Tota\ln umberofpredictions} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = 2 \times \frac{\mathrm{Pr}\,ecision * \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \quad (4)$$

Where TP – True Positive rate, FP – False Positive rate, and FN – False Negative rate

Similarly we can find the performance of the regression systems are calculated with the help of metrics namely, MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and Correlation Coefficient (R^2- Score).

$$(5)\ MAE = \frac{1}{N}\sum|Y - \hat{Y}|$$

$$(6)\ MSE = \frac{1}{N}\sum(Y - \hat{Y})^2$$

$$(7) RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum(Y - \hat{Y})^2}$$

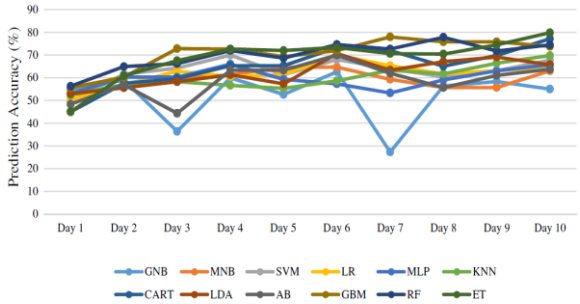
$$(8)\ R^2 = 1 - \frac{SSR}{SSM} = 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2}$$

Where SSR = Squared Sum error of Regression line

SSM = Squared sum error of mean line

Y – Actual value, $\hat{Y}$ – the predicted value of Y, and $\bar{Y}$ – Mean value of Y

The sample outputs for the stock price prediction using various classification and regression models



represented in the form of tables and graphs.

Table 3. Sample Result Analysis of Regression Models

REFERENCES

[1]. TrevirNath "How Big Data Has changed finance" https://www.investopedia.com/articles/active-trading/040915/howbig- data-has-changed-finance.asp.

[2]. Saheli Roy Choudhury. "Machines will soon will be able to learn without being programmed"https://www.cnbc.com/2018/04/17/machine-learning-investing-in-ainext- big- thing.html

[3]. Rohit Verma Astral institute Technical and Research, Indore (M.P.), PROF. Pkumar Choure (CSE) Astral institute Technical and Research, Indore (M.P.)"Neural Networks through Stock Market Data Prediction " International Conference on Electronics,Communication and AerospaceTechnology ICECA 2017

[4]. Pushkar Khanal ,Shree Raj Shakya "*Analysis and Prediction of Stock Prices of Nepal using different Machine Learning Algorithms*" Department of Mechanical Engineering, Pulchowk campus, Institute of Engineering, Tribhuvan University, Nepal

[5]. Yaojun Wang, Yaoqing Wang "Using Social Media Mining Technology to Assist in Price Prediction of Stock Market" .

[6]. Ashish sharma , dinesh bhuriya , upendra singh "Survey of Stock Market Prediction Using Machine Learning Approach "International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.

[7]. Amit Kumar Sirohi, Pradeep Kumar Mahato, Dr. Vahida Attar "Multiple Kernel Learning for Stock Price Direction Prediction " IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014),August 01-02, 2014, Dr. Virendra Swarup Group of Institutions, Unnao, India

[8]. Book-Elements of Artificial Neural Network,Kishan Mehrotra,Vhilukuri K K Mohan,Sanjay Ranka.

[9]. Cao J, Cui H, Shi H, Jiao L (2016) Big data: a parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce. PLoS ONE 11(6):e0157551

[10]. Cheng S, Shi Y, Qin Q, Bai R (2013) Swarm intelligence in big data analytics. In: International conference on intelligent data engineering and automated learning. Springer, Berlin, pp 417–426

[11]. Blum C, Li X (2008) Swarm intelligence in optimization. In: Dorigo M (ed) Swarm intelligence. Springer, Berlin, pp 43–85

[12]. Ibrahim RA, Ewees AA, Oliva D, Elaziz MA, Lu S (2019) Improved salp swarm algorithm based on particle swarm optimization for feature selection. J Ambient Intell Humaniz Comput. https ://doi.org/10.1007/s1265 2-018-1031-9

[13]. Moslehi F, Haeri A (2019) A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection. J Ambient Intell Humaniz Comput. https ://doi.org/10.1007/s1265 2-019-01364 -5

[14]. Zhong X, Enke D (2016) Forecasting daily stock market return using dimensionality reduction. Exp Syst Appl 67:126–139. https ://doi.org/10.1016/j.eswa.2016.09.027