Application of Data Mining Techniques on Dow Jones Index Stocks using WEKA

Rhea Rai

May, 2020

Final Project Report

Proposed Guide: Dr. Paul Benjamin, Professor, Computer Science, Pace University, NY.

Introduction

Dow Jones is the largest stock market index of the United States which calculates the stock performance of the top 30 companies. Predicting every stocks performance based on its previous weeks performance can bring about a significant amount of change in the margin therefore being profitable for investors for a safe investment. The Data Mining techniques that would be implemented are Linear Regression, K-Nearest Neighbour and Support Vector Machine and ZeroR.

The goal of this project is to be able to predict the net gain in the following week on the basis of the stocks performance in the previous week.

Description of the dataset

The given data set includes 750 instances of Dow Jones stock weekly metrics including sixteen features. The data represents the performance of every DJIA stock in the first and second financial quarters of 2011. The data was retrieved from the UCI Machine Learning Repository [1]. The features in the dataset include the financial quarter, the stock symbol, the date of the last

business day of that week, the weekly opening, high, low, and close prices for the stock, the volume of the stock traded over the week, the percent change in price of the stock, the percent change in volume over the last week, the previous weeks volume, the next week's open and closing prices, the percent change in the next week's price, the days until the next dividend for the stock, and the percent return on the next dividend. The data will be partitioned into two sets based on financial quarter. There are 360 examples in the first quarter and 390 in the second quarter. The stocks that made up the index were:

3M	MMM
American Express	AXP
Alcoa	AA
ATandT	T
Bank of America	BAC
Boeing	BA
Caterpillar	CAT
Chevron	CVX
Cisco Systems	CSCO
Coca-Cola	KO
DuPont	DD
ExxonMobil	XOM
General Electric	GE
Hewlett-Packard	HPQ
The Home Depot	HD
Intel	INTC
IBM	IBM
Johnson Johnson	JNJ
JPMorgan Chase	JPM
Kraft	KRFT
McDonald's	MCD
Merck	MRK
Microsoft	MSFT
Pfizer	PFE
Procter Gamble	PG
Travelers	TRV
United Technologies	UTX
Verizon	VZ
Wal-Mart	WMT
Walt Disney	DIS

Attribute Information:

- quarter: the yearly quarter (1 = Jan-Mar; 2 = Apr=Jun).
- stock: the stock symbol (see above)
- date: the last business day of the work (this is typically a Friday)

- open: the price of the stock at the beginning of the week
- high: the highest price of the stock during the week
- low: the lowest price of the stock during the week
- close: the price of the stock at the end of the week
- volume: the number of shares of stock that traded hands in the week
- percent_change_price: the percentage change in price throughout the week
- percent_change_volume_over_last_week: the percentage change in the number of shares of stock that traded hands for this week compared to the previous week
- previous_weeks_volume: the number of shares of stock that traded hands in the previous week
- next_weeks_open: the opening price of the stock in the following week
- next_weeks_close: the closing price of the stock in the following week
- percent_change_next_weeks_price: the percentage change in price of the stock in the following week
- days_to_next_dividend: the number of days until the next dividend
- percent_return_next_dividend: the percentage of return on the next dividend

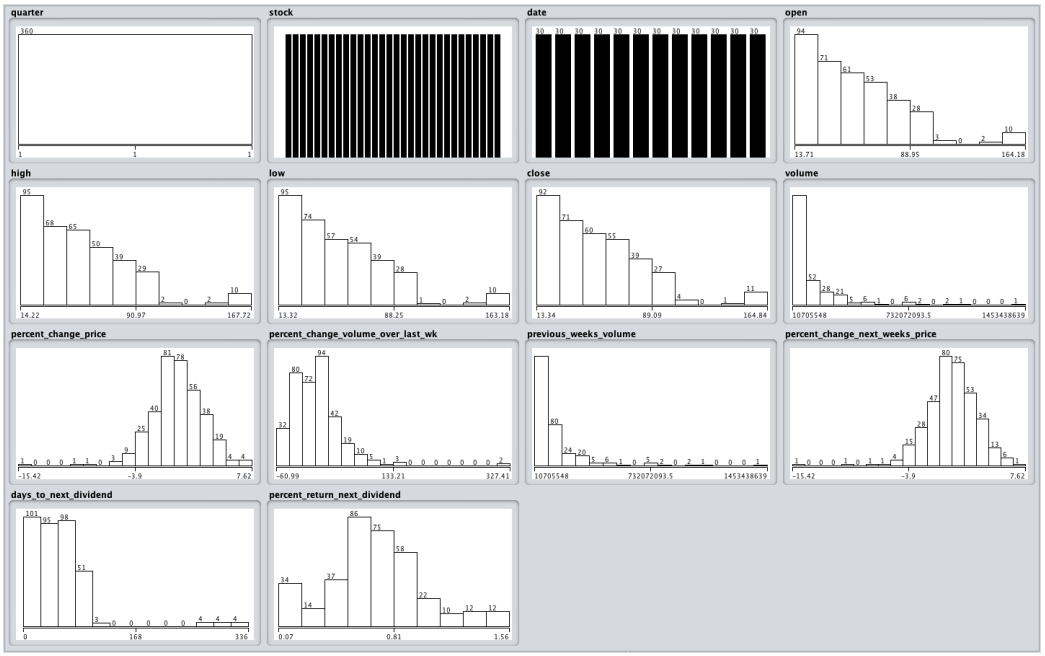
Link to the dataset: http://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index

Description of the problem solved

- Being able to predict the performance of more than 50 percent margin.
- Evaluate the effectiveness of different classifiers on the first quarter's data and then test the best classifiers on the second quarter's data, using the first quarter as a training set.
- Compare results with different classifiers along with accuracy.

Data Preparation

- Initially converted the dataset from UCI of .data format to .csv format.
- Partitioned the dataset into training and testing divisions as the training dataset would be used for both training and the development phase. After dividing the dataset into required forms, the final feature percent change in next weeks price, into binary classifier.
- If the stocks go up the following week, the binary classifier would be set to 1 and if the stocks go low the following week, it would be to 0.
- The unnecessary features such as quarter, stock symbol and date were removed. Features such as next weeks opening had no impact on the results, therefore those were eliminated too.
- The final features that were considered were as follows:
  - Opening price
  - Closing price
  - Weekly high price
  - Weekly low price
  - Volume
  - Percent change in price
  - Percent change in volume
  - Previous weeks volume
  - Days to next dividend
  - Percent return on next dividend
  - Result(price goes up or down)
- Further on, to impute missing values that were present in the dataset, ReplaceMissingValues filter in weka was applied. This filter imputed the missing values with mean of numerical distribution.



Data Analysis

The algorithms that have been implemented to analyze the DJIA stocks are: ZeroR, Logsitic Regression, K-Nearest Neighbour (Instance based Learner) and Support Vectore Machine. After due diligence considering various factors, these algorithms provided the best possible fit factors to provide optimal results.

ZeroR

ZeroR is the most naive of all the classifiers. The initial result provided by this algorithm can be later compared with other classifiers.This classifier emphasizes on the final target relies on the target and neglects all the predictors. ZeroR classifier simply predicts the majority category (class).

Logistic Regression

Logistic regression classifies examples with a multinomial logistic regression model. It is used for binary classification. It learns a coefficient for input

value that it is fed with, which is then combined into a regression function. The output of this function is then carry forwarded to a sigmoid function. A sigmoid function provides its output in the form of 0 and 1.

K-Nearest Neighbour (Instance based Learner)

Stocks with similar attributes show similar behaviour, in cases like so this algorithm performs the best. This algorithm takes in the dataset directly. It would consider the instances in its raw form. The predictions therefore generated are based on these raw instances. Also, since this algorithm takes in the dataset directly, its important to update the dataset as often as possible.

Support Vector Machine

Support Vector Machines work well with numerical values,but also convert nominal values to numeric values as input. Support vector machine is based on finding a line that equi distance from two groups and divides them equally.

Results

ZeroR, Logistic Regression, K-NN and SVM classifiers were tested on both the partitions of the dataset. The 1st quarters data was fed into the model as training set and the 2nd quarters data was computed to 10 fold cross validation. Before feeding into the network, both of the partitions were converted from numeric to nominal for efficient computation purposes.

ZeroR

For this classifier, the class attribute predicted whether the stocks went high or low in the following week in 0 or 1. Confusion matrix for ZeroR:

	True diagnosis		Classified as a and b
	a	b	
Positive	0	175	$a = 0$
Negative	0	215	$b = 1$

The model predicts the class to be 1, that is the stocks would go up. As observed, the mean absolute error computes up to be 0.4948 and root mean

square errors computes to be 0.4974. The ROC area is 0.487. THe accuracy on the test dataset was 55 %. The incorrectly classified instances are 175 and the correctly classified instances are 215

Logistic Regression

This algorithm would assume the input variables are of numeric format. A coefficient value for each input is learned in this algorithm. The input value of a linearly based function is learnt by this algorithm to transform it into a logistic based regression. This algorithm supports multi classification problem. Confusion matrix for Logistic regression:

	True diagnosis		
	a	b	Classified as a and b
Positive	1	174	$a = 0$
Negative	5	210	$b = 1$

As observed, the mean absolute error computes up to be 0.4959 and root mean square error computes to be 0.5015 The ROC area is 0.484. The accuracy on the test dataset was 54 %. The incorrectly classified instances are 179 and the correctly classified instances are 211.

K-Nearest Neighbour

This algorithm supports both classification and regression. This algorithm would take in the entire dataset and would query the instances around the given point with similar behaviour patterns and then make a prediction. For this prediction, Euclidean distance was selected. Confusion matrix for K-Nearest Neighbour:

	True diagnosis		
	a	b	Classified as a and b
Positive	98	77	$a = 0$
Negative	93	122	$b = 1$

The mean absolute error computes up to be 0.4713 and root mean square errors computes to be 0.5366. The ROC area is 0.569. The accuracy of the test dataset as per K-Nearest Neighbour is 56 %. The correctly classfied

instances using KNN are 220 and incorrectly classified instances are 170.

Support Vector Machine

In addition to being able to work with multi class classification and regression, Support Vector Machines also work with binary classification. The best functionality of Support Vector machine is it takes in both numerical and nominal data as input. If the data provided is nominal it will convert it into numerical. Before processing, it normalizes the data input. Support Vector Machine processes such that it creates a division between two groups differentiating them from each other into two groups. It would only consider the points that are close to the line drawn.
Confusion matrix for Support Vector Machine:

	True diagnosis	
	a	b
Positive	46	129
Negative	33	182

Classified as a and b

$a = 0$

$b = 1$

The mean absolute error computes up to be 0.4154 and root mean square errors computes to be 0.6445. The ROC area is 0.555 The accuracy of the dataset comes up to be 58 %. The correctly classified instances are 228 and incorrectly classified instances are 162.

Here, its evident now that not all the classifiers provided satisfactory results, rather majority of them performed moderately with not the best accuracy percent. But major similarity was found between ZeroR, Logistic regression and KNN, being clsoely ranged to each other whereas SVM performes the best among all with the highest accuracy percentage.

Conclusion

This intuitively makes a lot of sense since stocks of similar base perform closely similar within a weeks difference. Bur looking at this picture as a whole, a 55 % ranged performed is not the best of the lot but considering the stock market to be a fluctuating and volatile situation, a 55 % performance is decent enough for an uncertain dataset. This prediction could prove to be beneficial for an investor with monetary gain. To enhance the results gained

through these algorithms, consistent training with previous years datasets and a better range of stocks could be prove to be beneficial.

References

[1] UCI Machine Learning Repository: Dow Jones Index Data Set. [online] Available at: http://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index.

[2] Brown, M. S., Pelosi, M. Dirska, H. (2013). Dynamic-radius Species-conserving Genetic Algorithm for the Financial Forecasting of Dow Jones Index Stocks. Machine Learning and Data Mining in Pattern Recognition, 7988, 27-41.

[3] Gupta, A., Bhatia, P., Dave, K. and Jain, P., 2019. Stock Market Prediction Using Data Mining Techniques. SSRN Electronic Journal,.