Crypto-Currency Price Trend Prediction using k-Means Clustering

Evan Madhi Hamzh Al Rubaie
College of Engineering, University of Babylon, Hillah, Iraq

Abstract: Crypto-currency is new emerging asset. Many people are trading in crypto currency for money making. To predict the correct trend in crypto coin fluctuation, technical analysis is used. Many technical factors are present in analysis which help in predicting price trend. Few of this indicator are RSI, Ichimoku clouds, moving average convergence and divergence, different moving averages and Bollinger Band. This indicators are based on mathematical models. With advancement in data mining algorithm it is possible to use this indicators data as input and find out the price trend for different crypto currency. In this paper data pre-processing technique is used to remove noisy data, clustering is used to find cluster based on specified outputs. It is observed that accurate of this model is very good in predicting short term prices and fluctuation.

Key words: Clustering, data mining k-Means, neural network, MACD, RSI, KDJ, moving averages, Ichimoku

INTRODUCTION

Crypto currency is evolving with rapid growth. When Bitcoin was first introduces and it was making new high daily, people used to call bitcoin and other crypto currency as a bubble. After a long recent consolidation crypto market has proven that it was not a bubble but this a fluctuation which is present in all trading market (Velankar et al., 2018). With increased in mathematical function and extension in machine learning data mining technique it is now becomes simple to work simultaneously on complex math function and data mining simultaneously. Using this technique, it is possible to predict the trend of crypto currency prices. There are many exchanges which provides this information about daily, it is possible to collect this information from this source for data mining purpose.

With proper handling of this data and appropriate machine learning algorithm it is possible to predict good result in crypto price trend prediction. There are many technical analysis and indicators, oscillators are present in the analysis. Trades and investor use this parameter for determining the trend in crypto price fluctuation (Kamrat et al., 2018). Different indicators and oscillators have different characteristics, few of them are used for long term prediction like 1 years on the other hand few indicators are used to predict trend for very short time frame like 1 month. As there are many indicators and oscillators are present in analysis, it becomes difficult for traders to select proper indicators (Sakiz and Kutlugun, 2018). In this study clustering is used to find buy, sell or hold decision for traders. Technical indicators which are used for this purpose are RSI, Ichimoku clouds, moving average convergence and divergence, different moving averages and Bollinger Band. Characteristic and mathematical model for each indicator is as.

Price data, indicators, trading concept

Price data: There are many brokerage houses which allows to trade crypto currency. This brokerage houses provides open, high, low, close prices for all crypto pair. Using their API, it is possible to fetch the real time data from this source. In this study, Binance a major crypto currency exchange data is used as raw data. For training purpose around 6 months to 1 year data is collected and used to training purpose while testing and training performance analysis of machine learning algorithm only closing prices of each raw data is considered.

Indicators: Crypto currency traders use technical indicators and oscillators for understanding market movement. This indicator is capable of deciding the trend in market. Every indicator is computed by open, low, high, close and volume data (Wu and Diao, 2015). These indicators are RSI, Ichimoku clouds, moving average convergence and divergence, different moving averages and Bollinger Band. Characteristic and mathematical model for each indicator is as.

Moving averages: Simple moving average is wildly used moving average method in technical analysis. There is different type of moving averages like simple moving average, exponential moving average. In this study, simple moving average is used to predict the price trend. There are 4 time-series moving averages are used those are 20, 50, 100 and 200 day (Wu and Diao, 2015):

Simple moving average = \( \frac{A_1 + A_2 + \ldots + A_n}{N} \)

Where:

- \( A_1 \ldots A_n \) : Closing price
- \( n \) : Timeframe
When closing price is above 20 day moving average then it is more bullish and when closing price is below 200 days moving average, it is in bearish mode.

**RSI:** It is a momentum indicator, RSI uses time and price relation. It is known as relative strength index. RSI is depending on recent changes in crypto currency price change. RSI is a numeric value range from 0-100. It is used along with other indicators to determine trend in market. Average gain and average loss are calculated over specified timeframe. It is a general perception that if RSI is below 30 then it is oversold while if RSI is above 80 then it is overbought.

**Ichimoku Cloud:** This indicator is also known as Ichimoku Kinko Hyo. This indicators provides support and resistance and using pivot calculation it is possible to find out trend present in chart. Its mathematical function is computed by average of 9 days low and 9 days high. In Ichimoku analysis is based on conversion line and base line:

- Conversion line: (9-timeframe high+9-timeframe low)/2)
- Base line: (26-period high+26-period low)/2)
- Green span: (Conversion line+base line)/2)
- Red span: (52-period high+52-period low)/2)

Whenever there is a green cloud in Ichimoku analysis, trend suggest to be a positive while in case of red cloud, negative trend is considered.

**Moving average convergence and divergence:** MACD is moving average convergence and divergence is calculated based on moving averages of closing price. This value is calculated by 26 days exponential moving average is subtracted from 13 days exponential moving average. When MACD line crossover Signal Line, bullish trend is detected. When signal line crossover MACD line then bearish trend starts:

\[
MACD = 12 \text{ timeframe EMA} - 26 \text{ timeframe EMA}
\]

**Bollinger Band:** According to bollinger band equation, 4 different price lines are created and current closing price is compared with this line, Bollinger Band is mostly used with RSI and MACD indictors. Bollinger Band is also known as volatility band as it is based on price volatility:

- BB middle = 20-timeframe (SMA)
- BB upper = 20-timeframe (SMA)+(20-TF standard deviation of close x2)
- BB lower = 20-timeframe (SMA)-(20-TF standard deviation of close x2)

There are 3 types of Bollinger band which are upper band, lower band and middle band. Based on this three bands, 4 different assumptions are made by traders.

**MATERIALS AND METHODS**

**Model design**

**Price fluctuation analysis:** Crypto-currency price trend prediction system is based on several price fluctuation model. This involve all indictors mentioned above, these indicators are used as raw data. Pre-processing technique is used to reduce noisy and redundant data. Finally, appropriate machine learning algorithm is used to achieve better performance.

**Input and pre-processing:** Binance API provides data in open, high, low, close format which is known as OHLC data. It consists of coin pair open price, close price, high made in that interval and low made in that interval also it continues the quantity of trades volume happen in that interval of time. These values are used to find out indictors and oscillators which we explained in previous section. Using R mathematical library, it is possible to find out all these values is very easily.

This indicator and oscillators change their characteristic with time, hence, it is not possible to use direct value as input to machine learning algorithm that is the reason some noisy data and redundant data is removed as API call may content repeated data or data with null prices in case if trades don’t happen in the respective interval. Sampling is also used to sample this data into different format. Whenever data is not in form normalization is a required step.

**Sampling of technical indictors:** Bollinger Band, MACD, RSI and KDJ (stochastics) values are prepared by applying R functions to raw data. while analysing MACD, cross over line is more important than MACD values, hence cross over points are considered while creating a training model. After sampling these values are converted into MACD or signal as two different values for attribute. Just like MACD, Ichimoku also has importance’s to green and red cloud. When green cloud Is computed price likely to move forward and in case of red cloud, price is likely to move down. That’s why after sampling only two attribute values are present which are green or red. RSI is a numeric attribute and it is passed as a numeric value, machine learning algorithm is capable handling numeric value. Also it is observed that RSI with intermediate values between 30-80 also interpret different analysis which should not be missed out. Hence, RSI is used without sampling. Bollinger Band analysis creates 3 bands which are upper band, lower band and middle band.
Using this band, raw data is sampled to band1-4. If closing price is present above all bands then attribute value is assigned as band 1 December 18, 2019 if closing price lies in between then is labelled as band 2 if it is present below pivot point and above lower line then it is labelled as band 3 while if it is present in last band then band 4 is assigned in this way sampling is done in bollinger band analysis.

Generally, traders take their decision based on this analysis, hence using this inputs system as decide the output among 3 different attribute values like buy, sell and hold. TTR library is used to find MACD, RSI and Ichimoku values. These are required pre-processing steps for data analysis.

Clustering: Clustering is a process of forming group of similar objects. Each such group is called as cluster. Clustering is wildly used data mining technique. Clustering is used in data compression, computer graphics, image analysis and pattern recognition. Clustering is a technique, there are different algorithms present which specify how to form a cluster. This depend on the clustering algorithm. Popular way to form clusters are measuring distance by various distance measure, density-based clustering, distribution-based clustering, centroid-based clustering (Mishra et al., 2015).

In Fig. 1, different clusters are present. This clusters shows some common factors in that. This is useful in knowledge discovery. When a new input is given to system, it will find the correct cluster for input data and thereby a result of analysis is determined. In this study, k-Means clustering algorithm is used for training and building machine learning algorithm.

K-Means clustering: k-Means clustering is a vector quantization process. It is mostly used clustering algorithm in data mining. In k-Mean clustering all data is divided into k numbers of clusters. This clusters are created based on different calculation and distances between data points. It is an unsupervised learning algorithm. In k-Means algorithm all data points are distributed in k number of clusters based on the feature similarity. After applying k-Means algorithm to dataset, two observation are formed:

- Centroid on each cluster, this centroid can be used to label each cluster
- Training data can be assigned to a label

k determines the number of clusters to be formed after applying clustering algorithm. This is the major use of k-means algorithm. In this study, Euclidean distance is used to create cluster. Euclidean distance is a one of the distance measures equation for finding Euclidean distance is as:

\[ d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} \]

where, \( p = (p_1, p_2, \ldots, p_n) \) and \( q = (q_1, q_2, \ldots, q_n) \) are 2 points present in space. This distance measure is used in means clustering, using this distance, distance between 2 data points is calculated, based on this, cluster which is close from data point is selected for that data point.

**k-Means algorithm:**

- In the first step k is selected which is based on number of clusters require after clustering
- Using all Data points, randomly k numbers of points are selected which will acts as a centroid for each cluster
- Euclidean distance is used to place all data points to nearest centroid. Output of this step is k number of clusters but the centroids are not at centre
- These all k centroids are repositioned for optimizing cluster formation process
- These processes are carried out in iteration until centroid becomes static, meaning all clusters which are produced are with proper centroid

**Pseudo code:**

- Input: k and D (k is number of cluster and D is set of lift ratios)
- Output: set of k numbers of cluster
- Method: select k numbers of random cluster

**Repeat:**

- Assign or re-assign each data point to cluster which is more similar in this study it is based on Euclidean distance
Advantages of k-Means:

- k-Means is easy to implement and time complexity is better than another algorithm.
- When number of variables are large, k-Means is computationally faster.
- In k-Means it is possible to determine number of clustering before data mining.
- k-Means can produce more cluster than traditional hierarchical clustering algorithm.
- Because of iteration, appropriate centroid is computed in k-means algorithm.

**Trading model:** Trading model is used to take decision like buy, sell or hold. Data mining process consist of several steps which includes data collection, data pre-processing, training data, machine learning model creation, results analysis and testing. This process gives output as buy, sell or hold. Traders can use this decision for trading purpose (Fig. 2).

Data is collected from Binance API, this data is present in json format. OHLC values from each raw is extracted and R functions are used to find out values for MACD, Ichimoku and RSI. These values cannot be used directly, hence, prepossessing technique is used to sample this data. After sampling MACD data is sampled to signal or MACD. Bollinger Band data is sampled to 4 different attribute values, Ichimoku is sampled to green of red zone. If any data is present with null values, then it is discarded also if redundant is found that one entry is discarded in training phase.

After processing step, data forwarded to machine learning algorithm for model formation. In k-means clustering k is used as user input which is used to create k number of classes. User defined number of clusters is an advantage of k-means clustering over other clustering algorithm. In this example 3 clusters are formed which is responsible for buy, hold and sell decision:

- Input: \( k = 3 \), \( \text{seed} = 10 \).
- Distance measure = Euclidean distance

Figure 3 shows cluster formation, output of this model is 3 clusters. This clusters will be responsible for making decision for new inputs. Theses A, B and C cluster are formed of similar type of data, hence when new input arises it check the machine learning model and new data is added to cluster which is closer to the new data, this is measured with the help of Euclidean distance measure.

**Trading model:** Once buy decision is made, coin-pair is brought into portfolio and it will be remaining into portfolio until short term period or until system generates sell signal. The upper limit of 5% is also used to determine sell condition when a coin reaches 5% from entry it re-verifies the current status. In case of hold decision, no action is considered to be taken. If coin-pair is already present in long side, it is maintained while if coin-pair is present on short side, short call is maintained. Sell signal is very crucial in crypto trading as fluctuation is too high. Once sell signal is triggered then if portfolio already have coin-pair it gets sold out at current market price in case if no open position then new shorts are open to take short position. This is a simple discipline and training strategy which is followed in the research after buy, sell and hold call. It is observed that this trading strategy is very useful in booking timely profit and loss.
RESULTS AND DISCUSSION

Experiments: After training machine learning model, it is important to back test the accuracy of the algorithm for back testing clustering algorithm different techniques have been proposed which includes back testing by training set by supplied test set or percentage split can also be used to check the accuracy of formed cluster. In this experiment percentage split and supplied test are used to check the accuracy of clustering model.

First percentage split technique is used in which 66% data is split into two part, one is used for training purpose while other set of data is used for testing purpose. Total 8590 instances were used, out of which 66% is used for training and rest is used for testing purpose. Conclusion of this test is as. Total instances: 8590, number of iterations: 52, clusters 1: 1280, cluster 2 630 and cluster 3 102.

After applying test data is observed that out of 2920 tested instances 2686 instances were correctly clustered which indicates the accuracy of this model is around 91%.

In another approach different testing set are provided which is completely unknown to training data. Total 1000 records were tested on trained model and it is observed that 923 instances were correctly classified which indicates the accuracy of the model is around 92%.

For further analysis Weka tool is used to create cluster on supplied data, so that, it will be easy to back test and visualize all cluster formation. Diagram shows the visualization of clusters formed by applying k-means clustering to dataset with k =3 and seed =2 (Fig. 4).

Weka is used to find accuracy of the given dataset with k-means clustering. Experimental result of Weka shows the incorrectly classified instances as 11%.

Accuracy on real time data: Algorithm accuracy may be dynamic from real time accuracy, so that, it is also important to check the accuracy of model with real time data. In this test method real time data is collected from Binance and it is used as testing data for machine learning model. Output of this data is recorded and back test with real time growth. Many such data with different time frame is captured and back test with real time growth. Around 1000 different real time instances were used for back-testing with target as 5% and stop loss of 2%, time duration for this trade is maximum 2 months. It is found that around 836 trades generate buy or sell signal. Out of which 785 symbols has proven that buy and sell decision which is computed by machine learning algorithm was correct. Which proves the real time accuracy of the following model is 93%.

CONCLUSION

Finding: In this study k-means clustering algorithm is used for short term price prediction of crypto-pair. Noise removal and sampling technique is used on raw data. 3 clusters are formed with accuracy of around 90-93% which can predict trend for new inputs. In it observed that risk and reward ratio of this machine learning model is very good.

RECOMMENDATION

Different technical analysis techniques can be implemented which can provide great insight to data and also can improve the accuracy of model. Along with this twitter sentiment analysis and news-based technique can be implemented with this model to handle news-based event by which sudden price drop can be avoided. Theses technique can further improve accuracy of this model.

REFERENCES

Kamrat, S., N. Suesangiamsakul and R. Marukatat, 2018. Technical analysis for cryptocurrency trading on mobile phones. Proceedings of the 2018 3rd International Conference on Technology Innovation Management and Engineering Science (TIMES-iCON’18), December 12-14, 2018, IEEE, Bangkok, Thailand, pp: 1-4.
Mishra, R.K., K. Saini and S. Bagri, 2015. Text document clustering on the basis of inter passage approach by using K-means. Proceedings of the International Conference on Computing, Communication & Automation, May 15-16, 2015, IEEE, Noida, India, pp: 110-113.
Sakiz, B. and E. Kutlugun, 2018. Bitcoin price forecast via blockchain technology and artificial intelligence algorithms. Proceedings of the 2018 26th International Conference on Signal Processing and Communications Applications (SIU’18), May 2-5, 2018, IEEE, Izmir, Turkey, pp: 1-4.
Velankar, S., S. Valecha and S. Maji, 2018. Bitcoin price prediction using machine learning. Proceddings of the 2018 20th International Conference on Advanced Communication Technology (ICACT’18), February 11-14, 2018, IEEE, Chuncheon-si, South Korea, pp: 144-147.

Wu, M. and X. Diao, 2015. Technical analysis of three stock oscillators testing MACD, RSI and KDJ rules in SH & SZ stock markets. Proceedings of the 2015 4th International Conference on Computer Science and Network Technology (ICCSNT’15) Vol. 1, December 19-20, 2015, IEEE, Harbin, China, pp: 320-323.