Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 89 (2015) 36669-36672

ISSN: 2229-712X

Data Mining based Analysis of Stock Market Financial Data of Each Firms in Different Sectors

Layla Safwat Jamil and Ammar Riadh Kairaldeen
University of Baghdad, Iraq.

ARTICLE INFO

Article history:
Received: 20 October 2015;
Received in revised form:
22 November 2015;
Accepted: 27 November 2015;

Keywords
Data Mining in Finance,
Financial Data,
Stock Market,
Weka,
Analysis of Financial Data in Data Mining,
Data Mining Algorithms.

ABSTRACT
In this work we applied five data mining algorithms using Weka Tool (Data Mining Program) cases of monthly financial data of firms in the different sectors. These datasets have five years' experience between 2007–2012. After we applied these algorithms at these datasets, we displayed the results and compared the results that found by using these different algorithms.

Introduction

"Data mining" term now days refers to new methods on behalf of the intelligent analysis of large data sets. These methods have developed from several traditionally separate fields, for instance artificial intelligence, applied statistics, machine learning, information systems, data engineering, , and knowledge discovery. One of the attractive scope of this technology is finance application [1].

In general, data mining methods such as neural networks and decision trees can be a useful addition to the techniques available to the financial analyst. However, the data mining techniques tend to require more historical data than the standard models and, in the case of neural networks, can be difficult to interpret

Data mining

Data mining is a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating [2]

The Weka

Is (pronounced Way-Kuh) workbench covers a pool of algorithms and visualization tools used for predictive modelling and data analysis, organized with GUI support for easy to use. The original non-Java version was a TCL/TK front-end to modelling algorithms applied in other software design languages, In addition data pre-processing utilities in C language, and the Style of files used in the system used to running the experiments in machine learning. Weka previous version was mainly considered as a tool for analyzing data obtained from agricultural domains, but the version used to analyzing data selected in this paper are fully Java-based version (Weka 3), In this version, the developers started working on it since 1997, and now it used into different application, in research and different particular for educational purposes. Different advantages for Weka included but are not limited to:

- Free availability, used under the GNU (General Public License).
- Portability, since its implementation in Java and thus runs on different computer platform.
- Ease of use, because of its GUI with the comprehensive different modelling techniques and a big collection of data pre-processing.

Data sets

Each data for related company was taken from New York Stock Market. It was monthly value of each company and all of these taken found in Yahoo Finance Page. The data is related with 2007-2012, 5 year experience data.

	A
1	Date,Open,High,Low,Close,Volume,Adj Close
2	2011-12-01,25.56,26.19,25.16,25.96,49264800,24.50
3	2011-11-01,26.19,27.20,24.30,25.58,53693200,24.15
4	2011-10-03,24.72,27.50,24.26,26.63,60235300,24.95
5	2011-09-01,26.46,27.50,24.60,24.89,63522800,23.32
6	2011-08-01,27.51,27.69,23.79,26.60,77332100,24.92
7	2011-07-01,25.93,28.15,25.84,27.40,68186700,25.51
8	2011-06-01,24.99,26.00,23.65,26.00,61377000,24.21
9	2011-05-02,25.94,26.25,24.03,25.01,67821800,23.28
10	2011-04-01,25.53,26.87,24.72,25.92,81660300,23.97
11	2011-03-01,26.60,26.78,24.68,25.39,59744300,23.48
12	2011-02-01,27.80,28.34,26.43,26.58,61355100,24.58
13	2011-01-03,28.05,29.46,27.42,27.73,71312700,25.50
14	2010-12-01,25.57,28.40,25.56,27.91,48111900,25.66
15	2010-11-01,26.88,28.87,24.93,25.26,68402700,23.23
16	2010-10-01,24.77,27.20,23.78,26.67,66458300,24.37
17	2010-09-01,23.67,25.53,23.54,24.49,63542900,22.38
18	2010-08-02,25.99,26.38,23.32,23.47,61156600,21.45
19	2010-07-01,23.09,26.41,22.73,25.81,71053500,23.46
20	2010-06-01,25.53,26.93,22.95,23.01,79675500,20.92
21	2010-05-03,30.67,31.06,24.56,25.80,89381300,23.45
22	2010-04-01,29.35,31.58,28.62,30.54,65821100,27.64
23	2010-03-01,28.77,30.57,28.24,29.29,51043300,26.51
24	2010-02-01,28.39,29.03,27.57,28.67,58684900,25.95
25	2010-01-04,30.62,31.24,27.66,28.18,81765200,25.38

Tele:
E-mail addresses: Layla_70@yahoo.com, eng_ammar81@yahoo.cm

For enriching data we add some information about companies. Firm age, employee numbers, incomes, gain information are added to original data.

We prepare financial data of each corporation. We seperate 4 sectors each corporations Figure 1. Each data for related company was taken from New York Stock Market.

Sector	Company name
Telecommunication	• AT&T • Verizon • Vodafone
Energy	• BP • EXXON • Shell
E-Commerce	• Amazon • Bestbuy • EBay
Technology	• Apple • IBM • Microsoft

Figure 1. 4 Different Sector and 3 Companies in Each Sector

It was monthly value of each company and all of these taken found in Yahoo Finance Page. The data is related with 2007-2012, 5 year experience data. As show in Figure (2), for enriching data we add some information about companies. Firm age, employee numbers, incomes, gain information are added to original data.

Class	Category	Range
Number of Employee	Small Middle Big Very Big	0 - 100.000 100.000 - 200.000 200.000 - 300.000 300.000 >
Age	Young Middle Age Old Very Old	0 - 25 25 - 50 50 - 75 75 >
Income	Poor Middle Rich Very Rich	0 - 100 100 - 200 200 - 300 300 >
Gain- Revenue	Weak Middle Strong Very Strong	0 - 10 10 - 20 20 - 30 30 >

Figure 2. How can we analyses the data, we decide the range and class of each data value

Inside all of these we add also extra information to each one of our data sets to make a good comparison and also gets a good result as show in figure (3)

Company Name	Foundation Year	Employees	Revenue (Billion $)	Net Income (Billion $)
-				
Vodafone	1991	83900	60	16
Verizone	1983	188200	110	2
At&T	1983	245350	127	7
Exxon	1999	76900	453	44
BP	1954	96200	388	11
Shell	1907	87000	481	26
Amazon	1994	109800	61	0,6
BestBuy	1989	180000	49	1,2
Ebay	1995	27770	14	2,8
Apple	1976	72800	170	37
IBM	1911	426751	104	17
Microsoft	1976	88596	77	21

Figure 3. Extra Information that we added to our original data, extra information about companies

Data mining algorithms

The most important used data mining algorithms are the following , In our projects we used many Data Mining algorithms, we applied these algorithms for our data sets then we found the results and compared them, In the following, there are some definitions for the Algorithms that we used in our project [3,4]:

Decision tree (DT), is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm [5].

Bayes Network A Bayesian network, Bayes network, belief network, Bayes (ian) model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). An example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Formally , Bayesian networks are directed acyclic graphs whose nodes represent random variables in the Bayesian sense: they may be observable quantities , latent variables , unknown parameters or hypotheses. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node [6].

Artificial Neural Network (ANN), often just called a neural network, one of the most important algorithm used in data mining, it is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases a neural network is an adaptive system that changes its structure during a learning phase. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data [7].

Expectation–maximization (EM), algorithm is an iterative method for finding maximum likelihood or maximum posterior (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These Parameter - estimates are then used to determine the distribution of the latent variables in the next E step [8].

Further Clustering Algorithm, Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a

collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters [9].

K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells [10].

Experimental Results

In the following we have to mention our results that come from applied the data mining algorithms under weka environment using the mentioned algorithms before:

Appling the Decision tree algorithm, when we applied this algorithms for our data sets , we got the following results in figure (4):

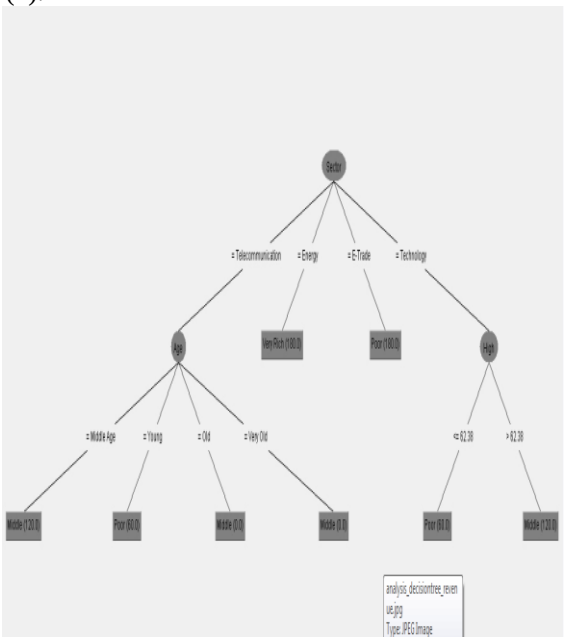

Figure 4. Decision tree applied algorithm

Applied Clustering Algorithm, under weka algorithm, and then we got the following results in figure (5):

Acknowledgment

According to our results those got from applied all of the used algorithms under weka environment, we can say that all of the used algorithms have adjacent results and can be used over our data sets and those near to ours. And can get good results.

naïve bayes_revenue, which is one of the most important used algorithms in data mining under weka program , when we applied this algorithm for our data set , we have the following results shown in figure (6) :

```
Missing values globally replaced with mean/mode

Cluster centroids:
                                    Cluster#
Attribute               Full Data          0          1
                          (720)        (420)      (300)
=============================================================
Sector              Telecommunication      Energy    Technolog
Age                         Young         Young  Middle Age
Size                        Small         Small       Small
Revenue                      Poor          Poor      Middle
Income                       Weak          Weak        Weak
Open                      68.3666       56.68      84.7278
High                      72.9254       60.645     90.1179
Low                       63.8566       52.6605    79.5311
Close                     68.9709       56.9239    85.8366
Volume                19462580 11333101.6667 30843849.6667
Adj Close                 62.748        51.027     79.1576



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      420 ( 58%)
1      300 ( 42%)
```

Figure 5. EM Clustering Algorithm

Attribute	Class Middle (0.33)	Poor (0.42)	Very Rich (0.25)
Sector			
Telecommunication	121.0	61.0	1.0
Energy	1.0	1.0	181.0
E-Trade	1.0	181.0	1.0
Technology	121.0	61.0	1.0
[total]	244.0	304.0	184.0
Firm			
AT&T	61.0	1.0	1.0
Verizon	61.0	1.0	1.0
Vodafone	1.0	61.0	1.0
BP	1.0	1.0	61.0
Exxon	1.0	1.0	61.0
Shell	1.0	1.0	61.0
Amazon	1.0	61.0	1.0
Best Buy	1.0	61.0	1.0
EBay	1.0	61.0	1.0
Apple	61.0	1.0	1.0
IBM	61.0	1.0	1.0
Microsoft	1.0	61.0	1.0
[total]	252.0	312.0	192.0
Age			
Middle Age	181.0	61.0	1.0
Young	1.0	241.0	61.0
Old	1.0	1.0	61.0
Very Old	61.0	1.0	61.0
[total]	244.0	304.0	184.0

Figure 6. NAÏVE BAYES Algorithm

Other Results

```
                Age
                Size
                Revenue
                Income
                Date
                Open
                High
                Low
                Close
                Volume
                Adj Close
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

Sector = Energy: Very Rich (180.0)

Age = Young: Poor (240.0)

Volume <= 46656300: Middle (237.0/1.0)

Firm = Microsoft: Poor (59.0)

: Middle (4.0)

Number of Rules  :      5


Time taken to build model: 0.01 seconds
```

Figure 7. revenue

References

[1] Stephen Langdell, PhD, Numerical Algorithms Group, Examples of the use of data mining in financial applications,

[2] Jiawei Han and Micheline Kamber , Data Mining Concepts and Techniques , Second Edition, University of Illinois at Urbana-Champaign ,2006.

[3] Tan. , Steinbach , Kumar , The K-means algorithms ,ICDM: Top Ten Data Mining Algorithms , December , 2006.

[4] MacKay, D.J.C. Information Theory, Inference and Learning Algorithms (Cambridge University Press, Cambridge, UK, 2003).

[5] Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

[6] Pearl, Judea (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press. ISBN 0-521-77362-8. OCLC 42291253.

[7] Hentrich, Michael (2015). "Methodology and Coronary Artery Disease Cure".

[8] Matsuyama, Yasuo (2011). "Hidden Markov model estimation based on alpha-EM algorithm: Discrete and continuous alpha-HMMs". International Joint Conference on Neural Networks: 808–816.
[9] Zhang, et al. "Agglomerative clustering via maximum incremental path integral." Pattern Recognition (2013).
[10] Honarkhah, M; Caers, J (2010). "Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling". Mathematical Geosciences 42: 487–517. doi:10.1007/s11004-010- 9276-7.