Analyzing Stock Market Data Using Clustering Algorithm

R. Suganthi and P. Kamalakannan

***Abstract*—In order to analyze the various sectors of the stocks in stock market field, need to use machine learning algorithms to determine the particular sectors of the stocks in intraday trade. In this paper, we compared different types of clustering algorithm with the help of data mining tool WEKA. This paper will demonstrate the strength and accuracy of each algorithm for clustering in terms of performance, efficiency and time complexity required.**

***Index Terms*—Machine learning, clustering, weka tool, multi database, stock market data.**

I. Introduction

Machine learning has confirmed to be of good quality value in different kind of application domains. It is specifically useful in data mining problems where big databases may have valuable enclosed regularities that can be discovered automatically. Designing a machine learning approach contains no of design options like choosing the typing of training experience. Target function from training ex: various commonly used machine learning algorithms [1] are Artificial neural network, decision tree, genetic algorithm, apriori algorithm, Rule induction etc...The application of machine learning approach to stock market data is a recent trend in research. The stock market daily trade result in stock market field has been still a source of great concern and research interest to process the analyzing stock market data and buyers to find the better stocks in different stock sectors. The discovery of this process can help the customer to take their own decision on daily basis trade where it can analyze the buying habit of the customers.

The literature is replete with various works in a machine learning area on stock market data. The focus of this work is on applying machine learning algorithms to stock market data for predictive and analyzing data purposes in stock market field.500 records of dataset has to be used the training data will be measured by clustering algorithm. The comparison of various algorithm will be produced the performance and efficiency of dataset.

II. Aim and Objective

The aim of this study is to use machine learning algorithms to determine the specific sector stocks in intraday trade of stock market, the general objectives is to demonstrate how machine learning algorithms can be applied to stock market data.

1) To create database representing the specific sector of stock market
2) To model the daily trade data based on dataset.
3) To compare models generated from the machine learning algorithms (K-means, optics, EM, Cobweb) using the accuracy level and Time taken and identify which model is most appropriate for taking decision on daily trade data.
4) To make the model.
5) To study the sectors to see what can be learned from that
6) To intend a system framework based on efficiency of algorithm.

III. Literature Review

There has been significant progress in the field of machine learning bordering on its application to the areas of medicine, natural language processing, software development and inspection, financial investing, stock market applications and so on. Abul, L *et al.*, 2003 [2] also projected the concept of various clusters and validate the parameters by using the sub sampling method. Cluster dissimilarity is minimized. Except for the first operation, the other three operations are repeatedly performed in the algorithm until the algorithm converges. The essence of the algorithm is to minimize the cost function where n is the number of objects in a data set [3] *X*, *Xi* €*X*, Ql is the mean of cluster l, and Yi l is an element of a partition matrix Yn × l as in (Hand 1981). d is a dissimilarity measure usually defined by the squared Euclidean distance. There exist a few variants of the k-means algorithm [4] which differ in selection of the initial k means, dissimilarity calculations and strategies to calculate cluster means. The sophisticated variants of the k-means algorithm include the well-known ISODATA algorithm and the fuzzy k-means algorithms and Guha, S *et al.*, 1998 proposed the efficient clustering algorithm for huge databases.

Edwin Lughofer *et al.*, 2012 [5] uses cluster analysis to group related documents for browsing, to find genes and proteins that have similar functionality, and to provide a grouping of spatial locations prone to earthquakes. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data

Manuscript received October 9, 2014; revised December 12, 2014.

R. Suganthi is with the Department of Computer Applications, Valluvar college of Science and Management, Karur, India (e-mail: mcasuganthi2011@gmail.com).

P. Kamalakannan is with the Department of Computer Applications, Government Arts College, Salem, India (e-mail: kamal_karthi96@yahoo.com).

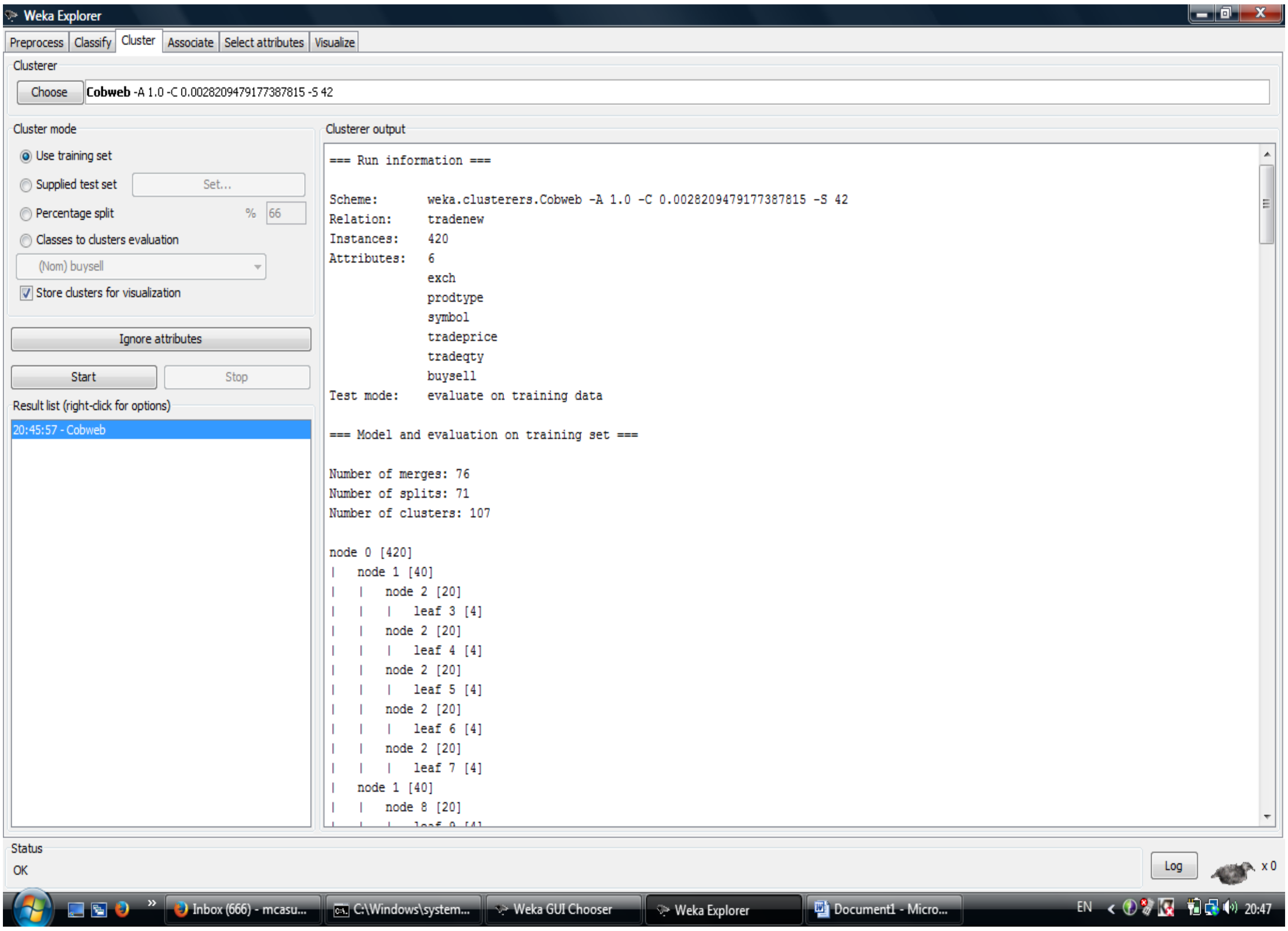
DOI: 10.7763/IJFCC.2015.V4.366

sets grow and their properties and data interrelationships change. Cluster validation is an indispensable process of cluster analysis, because no clustering algorithm can guarantee the discovery of Genuine clusters from real datasets and that different clustering algorithms often impose different cluster structures on a data set even if there is no cluster structure present in it . Cluster validation is needed in data mining to solve the following problems:

1) To measure a partition of a real data set generated by a clustering algorithm
2) To identify the genuine clusters from the partition.
3) To interpret the clusters.

Generally speaking, cluster validation approaches are classified into the following three categories Internal. Approaches, Relative approaches and External approaches.

Kevin Duh [6] has proposed a novel active learning strategy for data-driven classifiers, which is based on unsupervised criterion during off-line training phase, followed by a supervised certainty-based criterion during incremental on-line training. In this se, they call the new strategy hybrid active learning. Sample selection in the first phase is conducted from scratch (i.e.no initial labels/learners are needed) based on purely unsupervised criteria obtained from clusters: samples lying near cluster centres and near the borders of clusters are expected to represent the most informative ones regarding the distribution characteristics of the classes. In the second phase, the task is to update already trained classifiers during on-line mode with the most important samples in order to dynamically guide the classifier to more predictive power.



Cobweb -A 1.0 -C 0.0028209479177387815 -S 42

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) buysell
☑ Store clusters for visualization

20:45:57 - Cobweb

```
=== Run information ===

Scheme:       weka.clusterers.Cobweb -A 1.0 -C 0.0028209479177387815 -S 42
Relation:     tradenew
Instances:    420
Attributes:   6
              exch
              prodtype
              symbol
              tradeprice
              tradeqty
              buysell
Test mode:    evaluate on training data

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

Number of merges: 76
Number of splits: 71
Number of clusters: 107

node 0 [420]
|   node 1 [40]
|   |   node 2 [20]
|   |   |   leaf 3 [4]
|   |   node 2 [20]
|   |   |   leaf 4 [4]
|   |   node 2 [20]
|   |   |   leaf 5 [4]
|   |   node 2 [20]
|   |   |   leaf 6 [4]
|   |   node 2 [20]
|   |   |   leaf 7 [4]
|   node 1 [40]
|   |   node 8 [20]
|   |   |   leaf 9 [4]
```

Status
OK

Fig. 1. Result of cobweb clustering algorithm.

Both strategies are essential for reducing the annotation and supervision effort of operators in off-line and on-line classification systems, as operators only have to label an exquisite subset of the off-line training data representation give feedback only on specific occasions during on-line phase.

Xiaodong Yu *et al*. [7] have proposed a flexible transfer learning strategy based on sample selection. Source domain training samples are selected if the functional relationship between features and labels do not deviate much from that of the target domain. This is achieved through a novel application of recent advances from density ratio estimation. The approach is flexible, scalable, and modular. It allows many existing supervised rankers to be adapted to the transfer learning setting.

R. J. Gil. [8] has proposed a novel updating algorithm based on iterative learning strategy for delayed coking unit (DCU), which contains both continuous and discrete characteristics. Daily DCU operations under different conditions are modelled by a belief rule-base (BRB), which is then, updated using iterative learning methodology, based on a novel statistical utility for every belief rule. Compared with the other learning algorithms, their methodology can lead to a more optimal compact final BRB. With the help of this expert system, a feed forward compensation strategy is introduced to eliminate the disturbance caused by the drum-switching operations.

also R.J. Gil *et al*. have proposed a novel model of an Ontology-Learning Knowledge Support System (OLeKSS) is proposed to keep these KSSs updated. The proposal applies concepts and methodologies of system modelling as well as a wide selection of OL processes from heterogeneous knowledge sources (ontology texts, and databases), in order to improve KSS's semantic product through a process of periodic knowledge updating. An application of a Systemic Methodology for OL (SMOL) in an academic Case Study illustrates the enhancement of the associated ontology through process of population and enrichment.

Also they proposed a novel updating algorithm based on iterative learning strategy for delayed coking unit (DCU), which contains both continuous and discrete characteristics. Daily DCU operations under different conditions are modelled by a belief rule-base (BRB), which is then, updated using iterative learning methodology, based on a novel statistical utility for every belief rule. Compared with the other learning algorithms, their methodology can lead to a more optimal compact final BRB. With the help of this expert system, a feed forward compensation strategy is introduced to eliminate the disturbance caused by the drum-switching operations.

M. Ankerst *et al*. [9] have proposed OPTICS is good at investigating the arbitrarily shaped clusters, but its non-linear complexity often makes it only applicable to small or medium datasets. And different kinds of data review was taken to the clustering algorithm by the Jain *et al*., [10] 1998 methods

IV. Methodology

In this research, we focused on comparing the performance of machine learning algorithms that are trained with data relating to intraday trade with the aim of obtaining good stocks sectors for taking decision by their own of customers that will provide the proper alignment of daily basis trading data. For collection data to train the machine learning algorithms, quantitative approach will be used. A tool will be built based on database that interfaces with weka for training the algorithm for finding the optimization of analyzing model of stock market data will make a comparative research on the machine learning algorithms using cross validation and non clustering errors benchmarks. For training 40% date will be used while the remaining 60% will be used to validate. During data collection, the relevant data will be gathered. Once the data has been collected, its quality will be verified. Incomplete data will be eliminated and the data will be cleaned by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving in consistencies. At last, the un noisy data will be stored in different tables and later joined in a single table to remove errors.

V. Various Techniques

There are several clustering algorithm available in weka. But cobweb, DBscan, EM, Optics and K-Means will be used in this study. Attribute importance analysis will be carried out to rank the attributes by significance using information gain. Consistency subset selection and feature subset selection filter algorithm will be used to rank and select the attributes that are mostly used. The COBWEB algorithm was developed by machine learning researchers in the 1980s for clustering objects in a object-attribute data set. The COBWEB algorithm yields a clustering dendrogram called classification tree that characterizes each cluster with a probabilistic Description. Cobweb generates hierarchical clustering, where clusters are described probabilistically. COBWEB uses a heuristic evaluation measure called category utility to guide construction of the tree. It incrementally incorporates objects into a classification tree in order to get the highest category utility.

VI. Experimental Results

The data set is with COBWEB algorithm with evaluation on training set of WEKA tools. The number of merges found was 76, and number of splits are 71. Total time taken to build model on full training data was 27.88 seconds. Numbers of clusters were 107 (see Fig. 1).

VII. Dbscan Clustering Algorithm

DBSCAN (for density-based spatial clustering of applications with noise) is a data clustering Algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. OPTICS can be seen as a generalization of DBSCAN to multiple ranges, effectively replacing the parameter with a maximum search radius.

VIII. Experimental Results

The NSE data set is applied with COBWEB algorithm for evaluation on training set with Epsilon: 0.9, min Points: 6. Distance-type of clusters for DBSCAN Data Objects is Euclidian Data Object. The Number of generated clusters are 4 and elapsed time is 1.14 seconds.

IX. Hierarchical Clustering

In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis.

X. Experimental Results Other Recommendations

The NSE data set is applied with hierarchical clustering algorithm for evaluation on training set with four clusters. The elapsed time is 1.88 seconds. The tree Visualizer can also be analyzed.

XI. The k-Means Clustering

The k-means algorithm is one of a group of algorithms called partitioning methods. The k-means algorithm is very simple and can be easily implemented in solving many practical problems. The k-means algorithm is the best-known squared error-based clustering algorithm.

1) Selection of the initial k means for k clusters
2) Calculation of the dissimilarity between an object and the mean of a cluster
3) Allocation of an object to the cluster whose mean is nearest to the object
4) Re-calculation of the mean of a cluster from the objects allocated to it so that the intra

XII. Experimental Results of k-Means Clustering

The NSE data set is applied with k-means algorithm for evaluation on training set. Distance-type of Clusters for k-means Data Objects is Euclidian Distance. The number of iterations is 17, within cluster sum of squared errors: 127.02034887051619. Missing values globally replaced with mean/mode The Number of generated clusters are 4 and elapsed time is 0.16 seconds

XIII. Comparisons of Various Algorithms

Cobweb:
No of instances: 420
No of attributes: 6
No of merges: 76
No of splits: 71
No of clusters: 107
EM:
No of clusters elected by cross validation: 4
Cluster Instances: 4
Log likelihood: 14.01445
DBSCAN:
Clustering data objects: 420
No of generated clusters: 21
Elapsed time: 3.97
Un clustered instances: 40
K-Means:
No of iterations: 4
Within cluster sum of squared errors: 539.34
Clustered instances: 0 372 89%
1 48 11%

XIV. Conclusion

The main aim of this paper is to provide a detailed introduction of weka clustering algorithms.. It is the simplest tool for classify the data of various types. It is the first model to provide the graphical user interface of the user in order to performing the clusterization with the help of NSE data. It provides some features to analyze the NSE dataset for intraday trading with different clustering algorithms. Comparatively k-Means is the best because of simplicity and fastest than other algorithms that why k-means clustering algorithm is more suitable for stock market data mining applications. This paper shows only the clustering operations in the weka, further research can be performed using other data mining algorithms on Stock market data.

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2000.

[2] L. Abul, R. Alhajj, F. Polat, and K. Barker, “Cluster Validity Analysis Using Sub Sampling,” in *Proc. the IEEE International Conference on Systems*, Washington DC, 2003, vol. 2, pp. 1435-1440.

[3] S. Guha, R. Rastogi, and K. Shim, “CURE: efficient clustering algorithm for large databases,” in *Proc. the ACM SIGMOD Conference*, 1998.

[4] S. Jain, M. A. Aalam, and M. N. Doja, “K-means clustering using weka interface,” in *Proc. the 4th National Conference*, India, Com-2010.

[5] E. Lughofer, “Hybrid active learning for reducing the annotation effort of operators in classification systems,” *Pattern Recognition*, vol. 45, pp. 884–896, 2012.

[6] K. Duh and A. Fujino, “Flexible sample selection strategies for transfer learning in ranking,” *Information Processing and Management*, vol. 48, pp. 502–512, 2010.

[7] X. D. Yu, D. X. Huang, Y. H. Jiang, and Y. H. Jin, “Iterative learning belief rule-base inference methodology using evidential reasoning for delayed coking unit,” *Control Engineering Practice*, vol. 20, pp. 1005–1015, 2012.

[8] R. J. Gil and M. J. M. Bautista, “A novel integrated knowledge support system based on ontology learning: Model specification and a case study,” *Knowledge- Based Systems*, vol. 36, pp. 340–352, 2010.

[9] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “OPTICS: Ordering points to identify the clustering structure,” in *Proc. the Acmsigmod Conference*, pp. 49-60, 2010.

[10] A. Jain, M. N. Murty, and P. J. Flynn, “Data Clustering: A Review,” *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

P. Kamalakannan has completed his B.Sc and M.C.A from Bharathiyar University, Coimbatore, India in 1988 and 1991. He has completed his Ph.D degree in Periyar University, Salem in 2008. He has published more than 50 research articles in International journals and Conferences. He has conducted various seminars / conferences / workshop / FDP. His specialization is computer networks, Manet and distributed computing. He has 18 years teaching and 3 years Industry experience. Currently he is working as the head of the department of computer science department in Government Arts College, Namakkal.

R. Suganthi was worked as an assistant professor in computer application department in Valluvar College of Science and Management, Karur. She is doing research in the area of data mining and received her M.Phil degree in Periyar University, and completed her MCA in Alagappa University, Karaikudi in 2005. Her research interests include Data mining, Soft computing.