International Journal of Computer and Communication Technology

Volume 5 | Issue 4 Article 13

October 2014

Exploiting Data Mining Techniques For Improving the Efficiency of Time Series Data

TARUN DHAR DIWAN
DR. C.V.RAMAN University, Kota, Bilaspur (C.G.), India,, taruncsit@gmail.com

PRADEEP CHOUKSEY
Technocrats Institute of Technology Bhopal M.P. India, dr.pradeep.chouksey@gmail.com

R. S. THAKUR
MANIT Bhopal M.P. India, rsthakur@gmail.com

BHARAT LODHI
BIRT, Bhopal M.P. India, bharatlodhi@gmail.com

Recommended Citation
DIWAN, TARUN DHAR; CHOUKSEY, PRADEEP; THAKUR, R. S.; and LODHI, BHARAT (2014) "Exploiting Data Mining Techniques For Improving the Efficiency of Time Series Data," *International Journal of Computer and Communication Technology*: Vol. 5 : Iss. 4 , Article 13.
DOI: 10.47893/IJCCT.2014.1260
Available at: https://www.interscience.in/ijcct/vol5/iss4/13

Exploiting Data Mining Techniques For Improving the Efficiency of Time Series Data

TARUN DHAR DIWAN[1], PRADEEP CHOUKSEY[2], R. S. THAKUR[3] & BHARAT LODHI[4]

[1] DR. C.V.RAMAN University, Kota, Bilaspur (C.G.), India, [2] Technocrats Institute of Technology Bhopal M.P. India
[3] MANIT Bhopal M.P. India & [4]BIRT, Bhopal M.P. India
E-mail : [1]taruncsit@gmail.com, [2]dr.pradeep.chouksey@gmail.com

Abstract -The research work in data mining has achieved a high attraction due to the importance of its applications This paper addresses some theoretical and practical aspects on Exploiting Data Mining Techniques for Improving the Efficiency of Time Series Data using SPSS-CLEMENTINE. This paper can be helpful for an organization or individual when choosing proper software to meet their mining needs. In this paper, we propose utilizes the famous data mining software SPSS Clementine to mine the factors that affect information from various vantage points and analyse that information. However the purpose of this paper is to review the selected software for data mining for improving efficiency of time series data. Data mining techniques is the exploration and analysis of data in order to discover useful information from huge databases. So it is used to analyse a large audit data efficiently for Improving the Efficiency of Time Series Data. SPSS- Clementine is object-oriented, extended module interface, which allows users to add their own algorithms and utilities to Clementine's visual programming environment. The overall objective of this research is to develop high performance data mining algorithms and tools that will provide support required to analyse the massive data sets generated by various processes that is used for predicting time series data using SPSS- Clementine. The aim of this paper is to determine the feasibility and effectiveness of data mining techniques in time series data and produce solutions for this purpose.

Keywords *- Time series data, Data mining, Forecasting, classification, SPSS-Clementine.*

I. INTRODUCTION

Classification algorithm has discrete allowing predicting the relationship between input data sets. Commercial data mining software's are considerably expensive to purchase and the cost of training involved is high. For this, the best software that fits business needs is very important, crucial and difficult to decide the forecasting of any type of data set.

As well as modern data analysis has to cope with tremendous amounts of data. The modern economy has become more and more information-based. The widespread uses of information technology, a large number of data are collected which results in massive amounts of data. Such **time-ordered** data typically can be aggregated with an appropriate time interval, yielding a large volume of equally spaced **time series** data. Such data can be explored and analysed using many useful tools and methodologies developed in modern time series analysis.

Data mining is the exploration and analysis of data in order to discover meaningful patterns. Data mining techniques have been used to uncover hidden patterns and predict future trends. The competitive advantages achieved by data mining include increased revenue, reduced cost, and much improved marketplace responsiveness and awareness. The term data mining refers to information elicitation. It is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, data-base management, data visualization, mathematics algorithms and statistics. This technology provides different methodologies for decision-making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning and innovation.

The approach presented in this paper is a general one and can be applied to any time series data sequence. An improvement of technological process control level can be achieved by time series analysis in order to prediction of their future behaviour using SPSS- Clementine. SPSS Clementine is very suitable as a mining engine with its interface and manipulating modules that allow data exploration, manipulation and exploration of any interesting knowledge patterns. The paper deals with the utilization of data mining using SPSS-Clementine to fix best the prediction of time series. We can find an application of this prediction by the control in production of energy, heat, and etc.

SPSS Clementine software for data mining to understand Time series patterns for share marketing better curricula offerings used an classification mining tool to assist instructors in changing their pedagogical strategies and interventions by analyzing huge volumes of time series data.Classification finds patterns or a set of models in "training" data that describe and distinguish data cases or concepts. Classification constructs a model to predict the class of objects whose class type is known. Time series data accounts for a large fraction of the data stored in financial, medical and scientific databases. Recently there has been an explosion of interest in data mining time series, with researchers attempting to index, cluster, classify and mine association rules from increasing massive sources of data. For prediction and description of time series data we are using different data mining techniques. Here prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. The second goal which leads to descriptive model, describes patterns in existing data which may be used to guide decisions as opposed to making explicit predictions.

The research work done is about the share market forecasting of SBI time series dataset. The situation for trading changes every second. A time series database consists of sequence of values or events changing with time. Time series databases are used for studying the daily fluctuation of share market.

II METHODOLOGY OF DATA MINING

A. Definition

Data mining may be defined as "the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules". Hence, it may be considered mining knowledge from large amounts of data since it involves knowledge extraction, as well as data/pattern analysis

B. Tasks

Some of the tasks suitable for the application of data mining are classification, estimation, prediction, affinity grouping, clustering, and description. Some of them are best approached in a top-down manner or hypothesis testing while others are best approached in a bottom-up manner called knowledge discovery either directed or undirected.

As for Classification, it is the most common data mining task and it consists of examining the features of a newly presented object in order to assign it to one of a predefined set of classes. While classification deals with discrete outcomes, estimation deals with continuously-valued outcomes. In real life cases, estimation is often used to perform a classification task. Prediction deals with the classification of records according to some predicted future behavior or estimated future value.

Both Affinity grouping and market basket analysis have as an objective to determine the things that can go together. Clustering aims at segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters that are not predefined. Description is concerned with describing and explaining what is going in a complicated database so as to provide a better understanding of the available data.

C. The Various Cycle of Data Mining

The four stages of the virtuous cycle of data mining are:

- Identifying the problem: where the goal is to identify areas where patterns in data have the potential of providing value.
- Using data mining techniques to transform the data into actionable information: for this purpose, the produced results need to be understood in order to make the virtuous cycle successful. Numerous pitfalls can interfere with the ability to use the results of data mining. Some of the pitfalls are bad data formats, confusing data fields, and lack of functionality. In addition, identifying the right source of data is crucial to the results of the analysis, as well as bringing the right data together on the computing system used for analysis.
- Acting on the information: where the results from data mining are acted upon then fed into the measurement stage.
- Measuring the results: this measurement provides the feedback for continuously improving results. These measurements make the virtuous cycle of data mining *virtuous.* Even though the value of measurement and continuous improvement is widely acknowledged, it is usually given less attention than it deserves.

D. DATA MINING TECHNIQUES

Data mining can be described as "making better use of data". Every human being is increasingly faced with unmanageable amounts of data; hence, data mining or knowledge discovery apparently affects all of us.

There are two different types of tools used in data mining which are classification and prediction. Classification and prediction is the process of identifying a set of common features and models that describe and distinguish data classes or concepts. The models are used to predict the class of objects whose class label is unknown. A large number of classification models have been developed for

predicting future trends of stock market indices and foreign exchange rates.

Classification - is the task of generalizing known structure to apply to new data. Classification tools tend to segment data into different segments. The process of classification starts with a classification algorithm, which is applied to a set of so called training data. The training data is fed through the classification algorithm. When the classification rules have been defined a set of non related test data can be run through the classification rules. With the result from the test data it can be estimated whether the rules work and are able to classify segments. If they show that the classification does not work to with in a desired confidence interval a new classification algorithm can be implemented to improve the results of the classification rules.The purpose of data classification is organizing and allocating data to detached classes. In this process, a primary model is established according to the distributed data. Then this model is used to classify new data. Thus, applying the obtained model, it can be determined that to which class the new data belongs.

Classification is used for discrete values and foretelling. In the process of classification, the existing data objects are classified into detached classes with partitioned characteristics (separate vessels) and are presented as a model. Then considering features of each class, the new data object is allocated to them; its label and kind becomes determinable.

In classification, the established model is obtained based on some training data (data objects that their class's label is determined and identified). The obtained model can be presented in different forms like: classification rules (If- Then), decision trees, and neural networks. Marketing, disease diagnosis, analysis of treatment effects, find breakdown in industry, credit designation and many cases related to prediction are among applications of classification.

2.1. Types of classification methods

Classification is possible through the following methods:

- Bayesian classification
- Decision trees
- Nearest neighbor
- Regression
- Genetic algorithms
- Neural networks
- Support vector machine (SVM)

Prediction- Data mining techniques provides with a level of confidence about the predicted solutions in terms of the consistency of prediction and in terms of the frequency of correct predictions. The most extensively used tools in prediction are linear and multiple regression. Linear regression is the simplest form of regression analysis where there is only one predictor variable. Where as the multiple regressions is a more complex regression analysis where there are two or more predictor variables. Also non linear regression is used in cases where there are no linear relationships with data.

Time Series: A time series is a sequence of values that a randomly varying attribute accumulates over time. A time series does not use any mechanism to adapt its values, and this makes it very different from other series. Common time series examples are stock markets, weekly weather reports, annual precipitation or weekly pizza sales. Real world time series data tend to be continuous, and are usually a sequence of observations or values separated by equal time intervals.

Time series are often presumed to consist of components that enable us to predict future patterns. These components are:

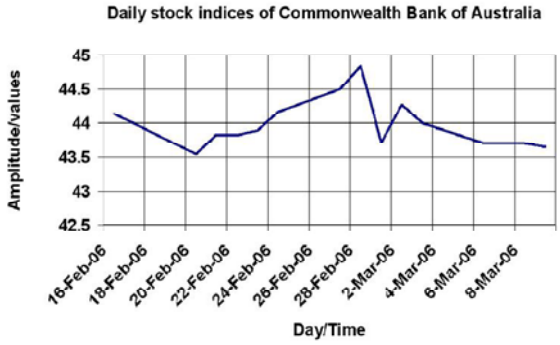
- Trend,
- Cycle,
- Seasonal variations, and
- Irregular fluctuations.

Time series prediction - Time series prediction/ forecasting is the process of studying known past events and extrapolating the results to predict future events, or in other words, the process of predicting future data points before they actually exist to confirm the measurements. Prediction of future values is complex and often difficult owing to the inherently volatile and non-linear nature of time series. Usually, prediction methods predict time series one or more steps ahead by evaluating historic data values alongside related data that may have influenced the series itself. In this thesis, we use the term prediction and forecast interchangeably to mean the forecast the future value.

Date Closing	Price
16-Feb-2006	44.13
17-Feb-2006	43.99
20-Feb-2006	43.54
21-Feb-2006	43.83
22-Feb-2006	43.82
23-Feb-2006	43.89
24-Feb-2006	44.16
27-Feb-2006	44.50

28-Feb-2006	44.85
01-Mar-2006	43.71
02-Mar-2006	44.27
03-Mar-2006	44.00
06-Mar-2006	43.70
08-Mar-2006	43.70
09-Mar-2006	43.65

Time series data: Daily stock price of Commonwealth Bank of Australia.



23-Feb-2006	43.89
24-Feb-2006	44.16
27-Feb-2006	44.50
28-Feb-2006	44.85
01-Mar-2006	43.71
02-Mar-2006	44.27
03-Mar-2006	44.00
06-Mar-2006	43.70
08-Mar-2006	43.70
09-Mar-2006	43.65

Time series data: Daily stock price of Commonwealth Bank of Australia.

2.2 Clementine software

SPSS Clementine data mining software is one of the most prominent software in data mining domain. This software is from famous SPSS software series and like previous statistical software has many facilities in data analysis domain.

The last version of this software is 12 that after its publication, next version named PASW Modeler was published. Among advantages of this software, the following cases can be mentioned:

- Consisting highly various methods for data analysis
- Very high speed in doing calculations and using database's information
- Consisting graphical environment for user's more comfort in doing analytic tasks

In new version, data cleanup and preparation are accomplished fully automatically. This software supports

All famous database software like Microsoft Office, SQL, etc.

Modules existing in this software are:

PASW Association

PASW Classification

PASW Segmentation

PASW Modeler Solution Publisher

This software can be installed on both personal computer and server; and supports 32-bit and 64-bit Windows too.

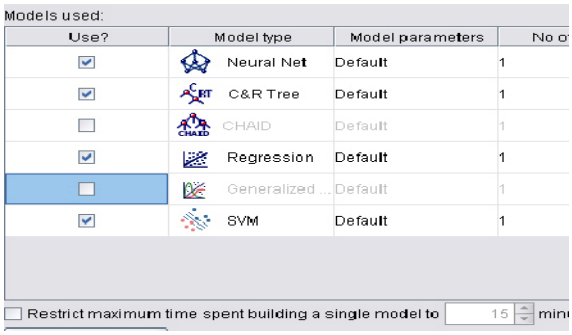
Data prepared for software

In order to make table of instructional data for classification algorithms, first we transfer column of **Max in** completely to an excel column and then transfer the same values to an opposite column in conditions that the first record is deleted.

Max in pre field is the same field that the algorithm should be finally able to predict one of them. Thus, the graph resulted from cleaned up values is presented. As it can be seen, generally, the trend of used bandwidth amount has been increasing but in some months, it has a remarkable decline.

The model created in Clementine software

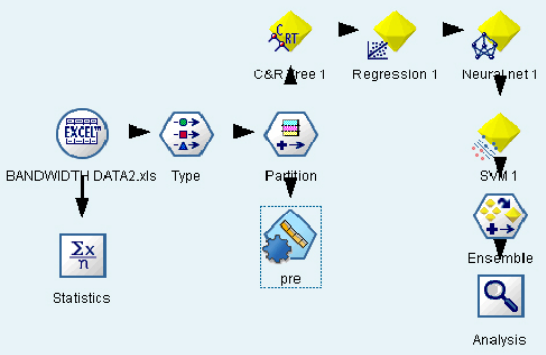
Now, applying Clementine software and partition node, we define eighty percent of data as instructional data and ten percent as training data and the ten remained percent as evaluation data from the final table prepared for software. Then we connect the node related to numerical prediction algorithms to related data and in its settings part, we activate the intended method.



Here, in order to implement the above algorithms, we should define columns **target** and

input. We present column **max in** as **input** and column **max in pre** as **target**.

After implementation of algorithms and creating intended models, we create combinational model of used algorithms and compare them to each other. Figure shows the procedure for implementation of the combinational algorithm.



CLEMENTINE TOOL:

Clementine (IBM SPSS Modeler) mines useful patterns out of scattered data. It makes it easy to discover insights into your data. Its high performance increases analyst productivity. It quickly discovers patterns and trends in data more easily using a unique visual interface supported by advanced analytics.

Clementine supports various features:-

1. It allows automated data preparation- saves time when preparing the data and get to analysis phase faster.
2. Comments- document the thoughts or processes used in the creation of a model.
3. Nearest Neighbor- Quickly and easily group similar cases. Eg valuable customers or donors using prediction or segmentation techniques.
4. Statistics Integration.
5. Improved Visualization- Generate graphs from a subset of model data create rich custom graph types and style sheets seamlessly with VizDesigner.
6. It provides faster and greater return on analytical investments. Automated Modeling helps us quickly identify best-performing models and combine multiple predictions for most accurate results.
7. Proven performance and Scalable architecture- Perform data mining within existing databases and score millions of records in a matter of minutes without additional hardware requirements.
8. Improved Visualization- It enables us to visualize the breakdown of clusters, automatically generate graphs from a subset of your model data, and create custom graph types.
9. Auto cluster node- It enables users to create, sort, browse and prioritize models faster.
10. Clementine uses a visual approach to data mining that provides a tangible way to work with data. Working in Clementine is like using a visual metaphor to describe the world of data, statistics and complex algorithms.

Applications of CLEMENTINE TOOL: -

Clementine is used to mine vast repositories of data. It offers a strategic approach to finding useful relationships in large data sets.

1) **Public Sector**:- Governments around the world use data-mining to explore massive data- stores, improve citizen relationships, detect occurrences of frauds example money-laundering and tax evasion, detect crime and terrorist patterns and enhancing the expanding realm of e-govt.

2) Drug discovery and Bioinformatics: Data mining aids both pharmaceutical and genomics research by analyzing the vast data- stores resulting from increased lab automation. Clementine□s clustering and classifications models help generate leads from compound libraries while sequence detection aids the discovery of patterns.

3) Web Mining:- With powerful sequencing and prediction algorithms, Clementine contains the necessary tools to discover exactly what guests do at a Web site and deliver exactly the products or information they desire. From data preparation to modeling, the entire data mining process can be managed inside of Clementine.

4) Clementine provides templates for many data mining applications. Clementine Application

Templates CATs are available for the following activities web mining, fraud-detection, Analytical CRM, Micro array analysis, crime detection and prevention, etc.

5) Customer Relationship Management: CRM can be improved thanks to smart classification of customer types and accurate predictions of churn. Clementine has successfully helped businesses attract and retain the most valuable customers in a variety of industries.

2.3 USAGE OF CLEMENTINE IN VARIOUS PHASES OF DATA MINING

Data Mining offers a strategic approach to finding useful relationships in large data sets. In contrast to traditional statistical methods, you do not need to know what you are looking for. You can

explore your data, fitting different models and investigate different relationships until you find useful information. There are various phases of data mining in which Clementine can help.

VISUALIZATION:-Clementine helps us gain an overall picture of our data. We can create plots and charts to explore relationships among the fields in our data set and generate hypotheses to explore during modeling.

MANIPULATION: - It lets you clean and prepare the data for modeling. We can sort or aggregate data, filter out fields, discard or replace missing values and derive new fields.

MODELING: - It gives us the broadest range of insight into the relationships among data fields.

Models perform a variety of tasks e.g. predict outcomes, detect sequences and group similarities.

III . RESULT AND DISCUSSION

In this section, CRISP-DM methodology has been implemented on real data set. Focus of this research is on Data Preparation and Modeling phases of this Standard for a Classification problem in order to SBI share market data set. The followings are the explanation for both of these steps:

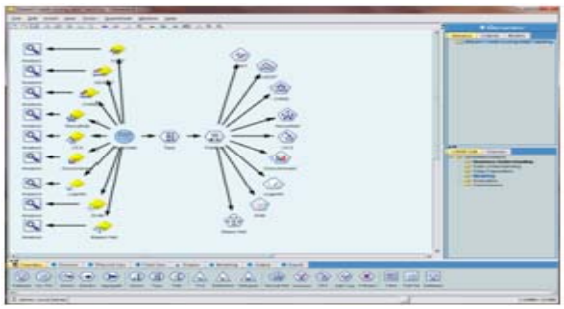
A. Data Preparation Step:

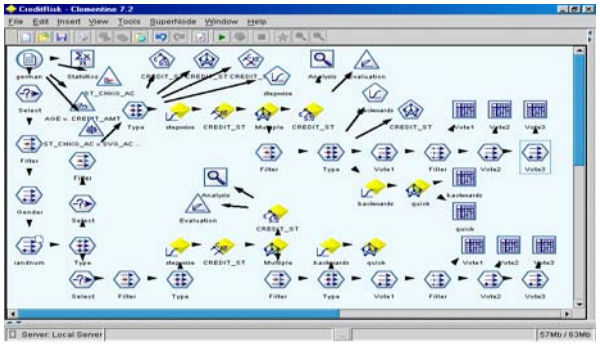
The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool) from the initial raw data. At First, the researchers have removed the noisy data from all the available records. The description of the data fields and the sample data gives us detailed information about the experimental data.

Modeling Step:

In this phase, various modeling techniques are available to be applied and their parameters are calibrated to optimal values.

Here, we have done the Classification Data Mining Algorithms with the Clementine tool on these data. Clementine data stream is shown in Fig. **Clementine Data Stream for Implementing Classification Algorithms**





Data Mining Diagram

B. Data Collection

Sources of Data

-The data set is taken for forecasting the future values consists of the financial data i.e., stock market data. The historical data set for share market gives the trends and seasonality patterns that help us to decide the accurate model for forecasting the future values and thus helps the investors to make better decision to buy or sell the share to gain profit in their Business.
We can compare the results with the ones from other models as well. Having the good results with this Hybrid model is not so important in this survey and to find a solution to forecast the problems that the output field (Class field) is being determined by multi criteria, is a proper work and overcomes the exist challenges and vagueness.

Table 1:- Sample Data Table

Record#	Date	Open	High	Low	Close	Volume
1	2011-05-02	2811.5	2819.5	2320.0	2327.6	798900.0
2	2011-04-01	2772.0	2959.9	2707.0	2805.6	349000.0
3	2011-03-01	2651.0	2888.0	2523.55	2767.9	445000.0
4	2011-02-01	2651.9	2813.4	2478.6	2632.0	662100.0
5	2011-01-03	2830.0	2852.4	2468.8	2641.0	819700.0
6	2010-12-01	2998.0	3172.0	2655.7	2811.0	682500.0
7	2010-11-01	3187.0	3515.0	2777.0	2994.1	694200.0
8	2010-10-01	3250.0	3322.0	3077.0	3151.2	280700.0
9	2010-09-01	2772.0	3268.0	2738.75	3233.2	478900.0
10	2010-08-02	2520.0	2884.0	2511.0	2764.8	482200.0
11	2010-07-01	2290.0	2519.9	2254.4	2503.8	343500.0
12	2010-06-01	2260.0	2402.5	2201.0	2302.1	403300.0
13	2010-05-03	2291.0	2348.8	2138.0	2268.3	505800.0
14	2010-04-01	2085.0	2318.8	2015.0	2297.9	452200.0
15	2010-03-02	1990.0	2120.0	1978.0	2079.0	357400.0
16	2010-02-01	2045.0	2059.9	1863.0	1975.8	550300.0
17	2010-01-04	2265.0	2315.2	1957.0	2058.0	583500.0
18	2009-12-01	2253.0	2374.7	2126.2	2269.4	585800.0
19	2009-11-03	2190.0	2394.0	2059.1	2238.1	725500.0
20	2009-10-01	2180.1	2500.0	2048.2	2191.0	855900.0
21	2009-09-01	1760.0	2235.0	1710.1	2195.7	508000.0
22	2009-08-03	1825.0	1886.9	1670.0	1743.0	428700.0
23	2009-07-01	1737.9	1840.0	1512.0	1814.0	650300.0
24	2009-06-01	1875.0	1935.0	1612.0	1742.0	651200.0
25	2009-05-04	1300.0	1891.0	1225.0	1869.1	924500.0
26	2009-04-01	1079.7	1355.0	980.0	1277.7	1100200.0
27	2009-03-02	1010.0	1132.2	894.0	1066.5	1214400.0
28	2009-02-02	1141.0	1205.9	1008.3	1027.1	814100.0
29	2009-01-01	1294.4	1376.4	1031.0	1152.2	1106800.0
30	2008-12-01	1095.0	1325.0	995.0	1288.2	1504000.0
31	2008-11-03	1155.0	1375.0	1025.0	1086.8	1426500.0
32	2008-10-01	1480.0	1569.9	991.1	1109.5	1220600.0
33	2008-09-01	1376.0	1618.0	1353.0	1465.6	993100.0
34	2008-08-03	1396.0	1638.9	1302.0	1403.6	840500.0
35	2008-07-01	1120.0	1567.5	1007.0	1414.7	674000.0
36	2008-06-02	1450.0	1496.7	1101.1	1111.4	474100.0
37	2008-05-02	1796.0	1840.0	1438.2	1443.3	368500.0
38	2008-04-01	1611.0	1819.9	1592.0	1776.3	369400.0

The data consist of historical data i.e. Stock market related fields such as open, high, low, close, volume and adjacent close and date field. The sample data is helpful in analysing the overall data set. It shows the various fields used in the dataset as well as the time interval at which the data are recorded.

C. Data Cleaning

-Real world data, like data acquired, tend to be incomplete, noisy and inconsistent. Data cleaning routines attempt to fill on missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

1) Missing Values

Many methods were applied to solve this issue depending on the importance of the missing value and its relation to the search domain.

- Fill in the missing value manually
- Use a global constant to fill in the missing value

2) Noisy Data

Noise is a random error or variance in a measured variable. Many techniques were used to smooth out the data and remove the noise.

• Clustering

Outliers were detected by clustering, where similar values are organized into groups, or clusters, values that fall outside of the set of clusters may be considered outliers.

• Combined computer and human inspection

Using clustering techniques and constructing groups of data sets, human can then sort through the patterns in the list to identify the actual garbage ones. This is much faster than having to manually search through the entire database.

3) Inconsistent Data

There may be inconsistencies in the data recorded for some transactions. Some data inconsistency may be corrected manually using external references, for example errors made at data entry may be corrected by performing a paper trace (the most used technique in our search, to guarantee the maximum data quality possible, by reducing prediction factors). Other inconsistency forms are due to data integration, where a given attribute can have different names in different databases. Redundancies may also exist.

D. Data Integration

Data Mining often requires data integration, the merging of data from multiple data sources into one coherent data store.

Careful integration of the data from multiple sources helped reducing and avoiding redundancies and inconsistencies in the resulting data set. This helped improving the accuracy and speed of the subsequent mining process.

E. Data Selection

Selecting fields of data of special interest for the search domain is the best way to obtain results relevant to the search criteria. We carefully selected from the overall data sets, and mining techniques were applied to these specific data groups in order to reduce the interesting patterns reached to the ones that represent an interest for the domain.

F. Data Transformation

In Data Transformation, the data is transformed or consolidated into forms appropriate for mining.

• **Smoothing:** which works to remove the noise form data. Such techniques include binning, clustering, and regression.

• **Aggregation:** where summary or aggregation operations are applied to the data.

• **Generalization of the data:** where low-level data are replaced by higher-level concepts through concept hierarchies.

• **Normalization:** where the attribute data are scaled so as to fall within a small specified range.

• **Attribute construction:** where new attributes are constructed and added from the given set of attributes to help the mining process.

G. Data Mining

1) Choosing the Tool

SPSS Clementine 8.1

As a data mining application, Clementine offers a strategic approach to finding useful relationships in large data sets. In contrast to more traditional statistical methods, you do not necessarily need to know what you are looking for when you start. You can explore your data, fitting different models and investigating different relationships, until you find useful information.

Working in Clementine is working with data. In its simplest form, working with Clementine is a three-step process. First, you read data into Clementine, then run the data through a series of manipulations, and finally send the data to a destination. This sequence of operations is known as a **data stream** because the data flows record by record from the source through each manipulation and, finally, to the destination--either a model or type of data output. Most of your work in Clementine will involve creating and modifying data streams.

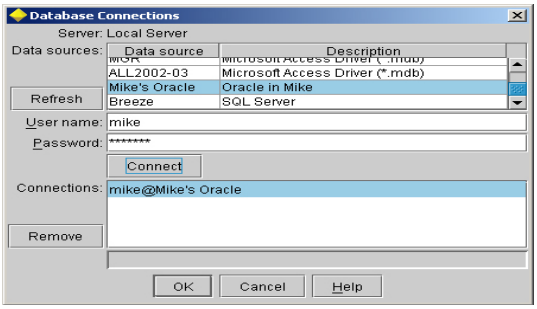
At each point in the data mining process, Clementine's visual interface invites your specific business expertise. Modeling algorithms, such as

prediction, classification, segmentation, and association detection, ensures powerful and accurate models. Model results can easily be deployed and read into databases, SPSS, and a wide variety of other applications. You can also use the add-on component, Clementine Solution Publisher, to deploy entire data streams that read data into a model and deploy results without a full version of Clementine. This brings important data closer to decision makers who need it. The numerous features of Clementine's data mining workbench are integrated by a visual programming interface. You can use this interface to draw diagrams of data operations relevant to your business. Each operation is represented by an icon or node, and the nodes are linked together in a stream representing the flow of data through each operation.

2) Using the Tool

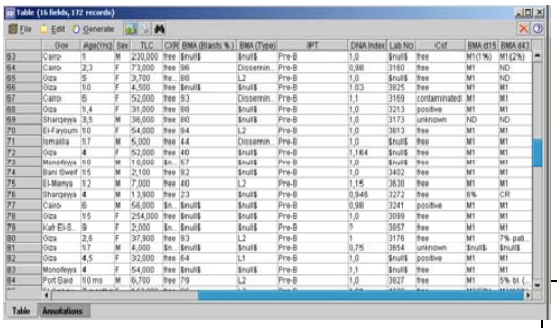
Creating the Data Source to use:

In our case an Oracle 9i Database was set and data fed into it in one table, ODBC was used to link the data source with the Clementine engine.



Viewing the Data in a Tabular Form

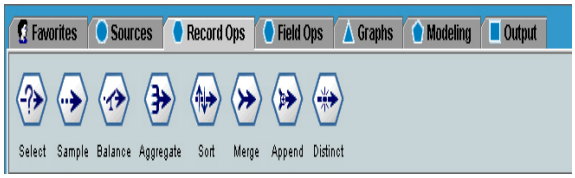
After linking the software with the data source, we view the data from the database in a tabular form by linking the data source to a "table" output



This enables us to assure that the link is successfully built and let us take a look on the form of data read by the software to detect any loss, inconsistency or noise that may have occurred in the linking process.

Manipulating the Data

By using the record Ops, operation concerning the records as a whole can be applied and used to operate on the data, using sampling, aggregation, sorting etc



Record Ops are linked to the data source directly and their output can be in any "output" means or can be directly fed as an input to other functions. Using the Field Ops on specific fields allows us to explore the data deeper, by using type selecting and filtering some fields as input or output fields, deriving new fields and binning fields.

H. Data Evaluation

After applying the data mining techniques comes the job of identifying the obtained results, in form of interesting patterns representing knowledge depending on interestingness measures. These measures are essential for the efficient discovery of patterns of value to the given user. Such measures can be used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre-specified interestingness constraints.

EMAMINING THE MODEL:

We can compare the testing data set for the months Jan 2011, Feb 2011, Mar 2011, Apr 2011, and May 2011, for both Expert Modeller and Exponential Smoothing. We analysed to know which model forecast better. The one, which gives better result and is more acceptable, will used to forecast the share values or the coming five months.

Table 2:- Time series values for validation

Record#	Model	Date	$TS Forecasted values for Testing				
			open	high	Low	Close	volume
97	Expert modeller	Jan 2011	2990.1	3264.4	2748.4	2653.8	733958.8
98		Feb 2011	2909.8	3359.5	2844.4	2573.6	777626.7
99		Mar 2011	2869.5	3457.3	2943.8	2520.6	814683.1
100		Apr 2011	2825.2	3558.0	3046.6	2718.0	846129.1
101		May 2011	2906.7	3661.7	3153.0	2801.5	872814.1
97	Exponential	Jan 2011	2990.1	3134.5	2572.4	2699.0	665905.1

98	**Smoothing**	Feb 2011	2909.8	3043.3	2558.1	2661.2	769676.0
99		Mar 2011	2869.5	3012.2	2470.6	2614.3	598313.2
100		Apr 2011	2825.2	3036.0	2516.8	2690.4	330289.0
101		May 2011	2906.7	3142.0	2526.2	2741.4	523515.1

Table 3: Statistical Measures

odel	**Attributes**	**Stationary R**2**	**Q**	**df**	**Sig**
Expert modeller	**open**	0.555	21.83	15.0	0.112
	high	0.202	19.66	18.0	0.352

Record #	**Model**	**Date**	**Forecasted values**				
			open	**high**	**Low**	**Close**	**volume**
102.	Expert modeller	Jun2011	2816.9	2792.5	2394.1	2299.7	1022627.1
103.		Jul 2011	2768.3	2813.8	2470.6	2415.0	869841.9
104.		Aug2011	2887.3	2835.2	2549.6	2450.5	1018116.7
105.		Sep2011	2926.8	2856.8	2631.1	2651.5	916858.6
106.		Oct2011	3133.0	2878.6	2715.2	2607.7	1015127.5

	Low	0.13	18.93	18.0	0.396
	Close	0.633	46.33	17.0	0.5
	volume	0.761	23.46	17.0	0.135
Exponential Smoothing	**open**	0.555	21.83	15.0	0.112
	high	0.527	17.37	15.0	0.297
	Low	0.536	28.86	15.0	0.017
	Close	0.562	20.72	15.0	0.146
	volume	0.677	27.56	15.0	0.024

This statistics provides an estimate of the proportion of the total variation in the series that is explained by the model. The higher value (to a maximum of 1.0), the better the fit of the model.
From the statistical measures and autocorrelation function/partial autocorrelation function we must conclude that the expert modeller will give better result as compared to exponential smoothing. It forecast the future values for the time series interval with more specificity and accuracy. Thus, for forecasting the share values for the coming five months we can use the expert modeller techniques.

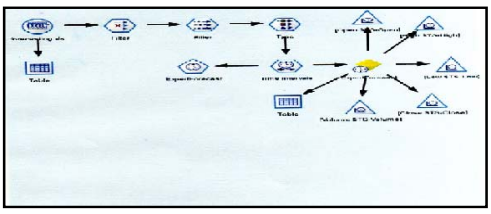
I. Knowledge Representation (outcome)

In this step visualization and knowledge representation techniques are used to present the mined knowledge to the user. All the operations applied on the records and fields, and the mining processes itself are represented in the form visualizations and graphics in this step.

STREAM MODELING FOR FORECASTING DATASET:

The stream designed for forecasting the data set is shown below:

STREAM MODEL FOR FORECASTING



The stream designing and source data is same. The source node for forecasting includes the monthly data from Jan 2003 to May 2011.

The expert modeller method is used for forecasting. The expert modeller method is set in the time series node. The expert modeller automatically select winters additive model or open & close fields; while ARIMA model is chosen for high, low, and volume fields. The golden executable nugget is connected with a table for showing the generated time series forecasted values.

Table 4:- Forecasted Values

Model	**Attributes**	**Stationary R**2**	**Q**	**df**	**Sig**
Expert modeller	open	0.546	21.212	15.0	0.13
	high	0.019	7.599	17.0	0.974
	Low	0.124	17.513	18.0	0.488
	Close	0.519	20.421	15.0	0.159
	volume	0.798	15.372	16.0	0.498

Table 5:- Statistical Measures

Table 4 gives the time series forecasted values of share market for open, high, low, close, and volume having record numbers from 102 to 106. Table 5 show the various statistical error measures.

IV. CONCLUSIONS

In this paper, with the use of Time series data, we assigned the Class to the records of a database and then the Data Mining algorithms have been done on the records with Clementine software. Correct Labeling of records is so essential in Data Mining, otherwise the Data Mining is not an integrated solution and we can not trust on the results.

Based on the previous work, the following conclusions were drawn:

1. Classification, as a data mining technique, is very useful in the process of knowledge discovery in the share market field, especially in the domains where available data have many limitations like inconsistent and missing values.

   In addition, using this technique is very convenient since the Classification is simple to understand, works with mixed data types, models non-linear functions, and most of the readily available tools use it.

2. SPSS Clementine is very suitable as a mining engine with its interface and manipulating modules that allow data exploration, manipulation and exploration of any interesting knowledge patterns

3. Using better quality of data influences the whole process of knowledge discovery, takes less time

in cleaning and integration, and assures better results from the mining process.

4. Using the same data sets with different mining techniques and comparing results of each technique in order to construct a full view of the resulted patterns and levels of accuracy of each technique may be very useful for this application.

So data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of time series data using SPSS- Clementine. However, data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved.We conducted an extensive consolidation on representation methods for time series data.

The application of data mining is extremely important. It is conclusive that the average error for simulations using lots of data is smaller than that using less amount of data. That is more data for training gives better prediction. If the training error is low, predicted values are close to the real values.

ACKNOWLEDGMENT

This work is supported by research grant from MPCST, Bhopal M.P., India, Endt.No. 2427/CST/R&D/2011 dated 22/09/2011.

REFERENCES

[1] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, CA, 2005.

[2] Lon-Mu Liu, Siddhartha Bhattacharyya, Stanley L. Sclove, Rong Chen (*) etal has paper DATA MINING ON TIME SERIES: AN ILLUSTRATION USING FAST-FOOD RESTAURANT FRANCHISE DATA (1-28).

[3] G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). Time Series Analysis: Forecasting and Control. Third Edition. Prentice Hall.

[4] Chaudhuri, S. and Dayal, U. (1997). "An Overview of Data Warehousing and OLAP Technology" ACM SIGMOD Record 26(1), March 1997.

[5] Fayyad, U. M. (1997). "Editorial." Data Mining and Knowledge Discovery 1: 5-10.

[6] Friedman, J. H. (1997). "Data Mining and Statistics: What's the Connection?" Proceedings of Computer Science and Statistics: the 29th Symposium on the Interface.

[7] E. J. Keogh. A Decade of Progress in Indexing and Mining Large Time Series Databases. In VLDB, 2006.

[8] E. J. Keogh and S. Kasetty. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. Data Min. Knowl. Discov. 7(4), 2003.

[9] Weiss, S. M. and Indurkhya, N. (1998). Predictive Data Mining. San Francisco: Morgan Kaufmann Publishers. Widom, J. (1995). "Research Problems in Data Warehousing". Proceedings of 4th International Conference on Information and Knowledge Management (CIKM), November 1995

[10] http://www.jatit.org/volumes/research-papers/Vol4No12/1Vol4No12.pdf,

[11] Proceedings of the 5th National Conference; INDIACom-2011 Copy Right © INDIACom-2011 ISSN 0973-7529 ISBN 978-93-80544-00-7.

[12] IJCSNS International Journal of Computer Science and Network Security, Vol.11, No.6, June 2011 262

[13] International Journal of Information and Education Technology, Vol. 1, No. 2, June 2011.

[14] World Academy of Science, Engineering and Technology 8 2005 309 Spring/Summer 2007.

[15] Computing For Nation Development, March 10 – 11, 2011

[16] Bharati Vidyapeeth□s Institute of Computer Applications and Management, New Delhi.,

[17] Data Mining – Decision Tree Induction in SAS Enterprise Miner and SPSS Clementine – Comparative Analysis Zulma Ramirez 2901 N Juan St. Edinburg, TX 78541 (956)802-6283

◈◈◈