Chapter - 1

INTRODUCTION

1.1 Overview of Data Mining

“If you mine the data hard enough, you can also find messages from God”
— Scott Adams, Dilbert

The world today is flooded with data; since human life began hunters seek patterns in animal’s migration behavior. Farmers seek patterns in crop growth, Politicians seek patterns in voter opinion’s, Users create content such as Blog posts, tweets, social-network interactions, and photographs; servers continuously create activity logs, the internet becomes the repository of data.

Over the last few decades, data growth has been on a massive scale that has become problematic with regard to scalability. This brought about a challenging situation for the handling of this huge some of the data. Exponentially increasing volumes of data with lots of complexity require effective and automatic approaches to inference knowledge from this voluminous data are important.

Extraction methods with high efficiency are needed to gain valuable knowledge from the hidden or undiscovered patterns amongst colossal amounts of data. Efficient means of storing, retrieving, and manipulating data, as a revolution in information availability and exchange via the internet helps the database technologists to focus on developing techniques for learning and acquiring knowledge from the data. The efficient decision-making process is the key to a successful organization that is based on timely and valuable knowledge.

Data within a Database needs to be analyzed to extract knowledge, with the implementation of one or more than one computer learning techniques, through the process called Data Mining (DM). Through Data Mining sessions the patterns and trends within data could be identified. In DM, it is important to understand the difference between a model and a pattern that helps in understanding the structure, relationships to a relatively small part of the data or the space in which the data would occur “The nontrivial extraction of implicit, previously unknown, and potentially useful information from data” (Brawley et al., 1996) is the definition of Data Mining. It is also a science of extracting useful information from large data sets or databases.

The definition of Data Mining as per Hand et al is “a well-defined procedure that takes data as input and produces output in the forms of models or patterns”. Data mining is being used in many fields such as Healthcare system, Biological systems, Medical sciences, Biometric applications, Pharmaceutical industries, banking, marketing analysis, Multimedia system, Retail and e-commerce, Spatial data analysis, Security system, String mining, fraud and intrusion detection, image processing, telecommunication industry, scientific applications, etc (Han et al. ,2011).

Raw data is extracted through data mining techniques which is useful in the process of prediction analysis, clustering that helps in the generation of many data mining techniques & tools (Dunham, 2011). When an abundance of data is available, it can be the past & future data, which gets analyzed by data analysis. Data mining is a multi-disciplinary field includes the combination of statistics, machine learning, artificial intelligence, and database technology and applications of it is very high (Fu, 1997)

Table 1.1 Data Mining Applications summarized

Applications	**Usage**
Communications	In communication for the prediction of customer's behavior for the relevant and targeted campaigns.
Insurance	Helps improve their profits through the implementation of offers to the existing & new customers
Education	In this sector, it helps in obtaining a profile of students, so that weak students in a particular subject can be sorted out easily.
Manufacturing	Manufacturers get the predictive approach towards the working & maintenance of machineries that helps them prevent major damages.
Banking	Data mining is useful to banks in the identification of defaulters, loans, etc, and also helps them to keep track of the banking details.
Retail	In malls & retail sectors of groceries, it helps in the sorting out of the highest selling commodity that attracts customer's attention and thus increases its profits.
Service Providers	Prediction of customer behavior through their billing details, interaction at service centers & complaints will improve the approach of mobile & utility industries in offering incentives to the customers.

E-Commerce	E-commerce websites use Data Mining for the promotion of their sales so that more customers are attracted towards e-commerce. This improves their sales strategy.
Super Markets	It allows supermarkets to predict the purchasing patterns of their shoppers & target or offer specific products according to their needs.
Crime Investigation	In crime investigation, it helps the police force to get deployed based on the likeliness of crime in a particular area.
Bioinformatics	In this field, it helps in the segregation of specific data from the biological & medicinal data sets.
Health care	Here in this sector data mining is highly helpful as large amounts of data are generated regularly, here it helps in processing and analyzing the data. Thus, helping in decision making about treatment, healthcare, Customer Relationship. Apart from this, it is helpful in various related sectors of pharmaceuticals industries, medical device industries, etc (Dura raj, 2013)

1.2 The Data Mining Process

As we have seen the amount of data in databases is increasing at a tremendous rate. This growing need gives birth to a new research field to aid humans to intelligently and automatically analyze huge data sets called Knowledge Discovery in Databases (KDD). Researchers start giving attention to many different fields including pattern recognition, database design, machine learning, statistics, and data visualization (Fayyad et al., 1996).

The KDD process can be decomposed into the following steps as illustrated in figure 1.1.

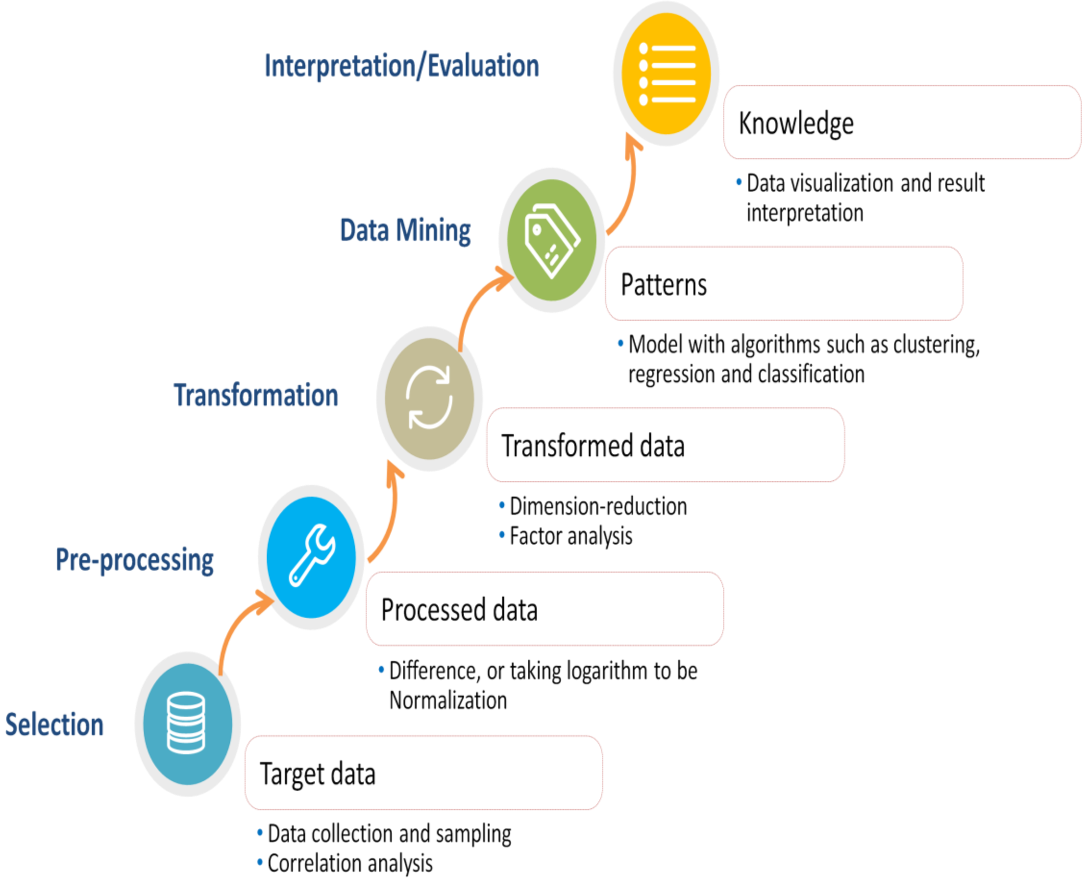
Interpretation/Evaluation
Knowledge
• Data visualization and result interpretation
Data Mining
Patterns
• Model with algorithms such as clustering, regression and classification
Transformation
Transformed data
• Dimension-reduction
• Factor analysis
Pre-processing
Processed data
• Difference, or taking logarithm to be Normalization
Selection
Target data
• Data collection and sampling
• Correlation analysis

Figure 1.1 KDD process

Source: Fayyad (1996)

The above diagram is a representation of the knowledge discovery process.

1.2.1 Data Selection

It is a step in which data required for analytical tasks are retrieved from the database through which discovery can be done. A target dataset required for the domain application can be selected/ generated from different kinds of data sources.

1.2.2 Data Preprocessing

In this technological transformation of raw data into an understandable format is processed. As real-world data is not full – proof, it has many errors within it.

I) Data Cleaning: In this step, the inconsistent data, missing data fields & noise are removed. Data of high quality is required to produce reliable knowledge in KDD processing. In this stage tasks such as removal of noise, missing value imputation, mapping of feature values onto appropriate domains, and attribute. Thus, fill in the missing values, smoothen noisy data, identifying or removing the outliers, and resolving the inconsistencies can be carried out in this step.

II) Data Integration: A combination of multiple data sources is involved in this step, Data from heterogeneous sources are put together and conflicts are resolved.

III) Data reduction: The task of data reduction is to remove non-relevant data from the dataset by preserving the knowledge involved in the dataset to afar extent possible. This process helps in attribute reduction that significantly reduces the complexity of the data mining operations. Thus, the quality of the mined results increases. Finding significant features to represent the data is included in Data reduction and projection. It also uses dimensionality reduction or transformation methods that help not only in reducing the effective number of variables under consideration but also to find invariant representations for the data.

IV) Data Transformation − Data transformation or consolidation into appropriate forms for mining is carried out by performing summary or aggregation operations. The high-dimensional dataset does not have all attributes of equal significance.

V) Data Discretization: Is used to divides the range of attributes into intervals thus reduces the number of instances within a dataset. It also improves the quality of data contained in the dataset.

1.2.3 Data Mining

Data patterns are extracted by intelligent methods. The functions of data mining include

- The purpose of the model (e.g., classification, clustering, regression, or summarization).
- Selection of algorithm(s) for searching for patterns in data.

Decision making on appropriate models & parameters.

1.2.4 Pattern Evaluation

Patterns for different data are evaluated based on the performance evaluation parameters.

1.2.5 Knowledge Presentation

Knowledge representation is carried out in this step. It searches for patterns of interest in a specific representational form by using classification rules, line analysis, regression, clustering, dependency, and sequence modeling.

1.2.6 Interpretation

Interpretation of the discovered patterns and referring back the previous steps, along with visualization of the extracted patterns, removing irrelevant or repetitive patterns, and translating the usage patterns as user friendly.

1.2.7 Use of discovered knowledge

Incorporation of the knowledge into the performance system so that actions could be taken based on the extracted knowledge, along with the resolution of potential conflicts with previously extracted knowledge.

1.3 Data mining task

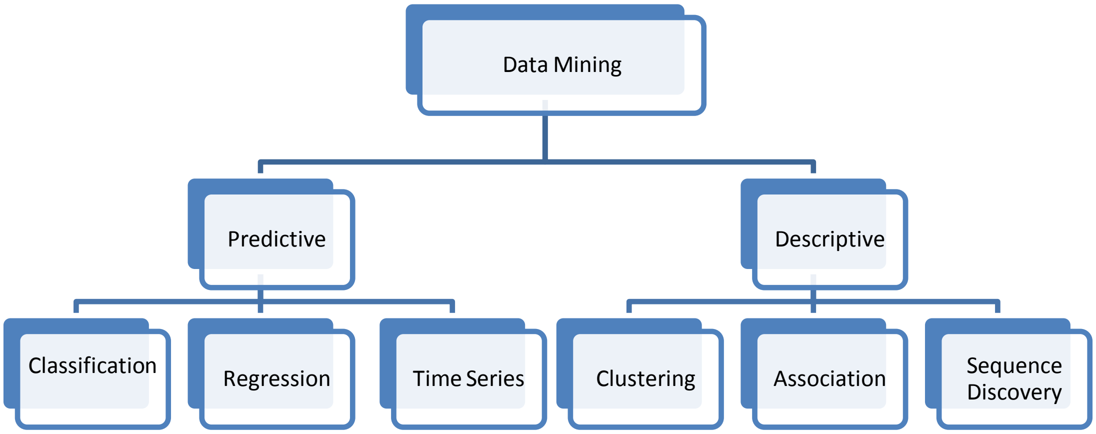

Figure 1.2 Data Mining Task

In general, the data mining task is of two types Predictive & Descriptive.

1.3.1 Predictive mining task

It is a task wherein we obtain inference on the present data to make predictions. The most popular predictive data mining task is Classification. The process of finding a model depends on the analysis of training data whose class label is known. The purpose of this model is to predict the unknown class label from the class of objects. The target dataset is randomly divided into two mutually exclusive and exhaustive sets called training set and test set to carry out a Predictive task.

The relationship between the conditional attributes and the decision attribute is derived from the training data, by which a model (or a function) is derived to describe the concepts. Such a model can be represented in various types such as mathematical formulae, decision trees, prediction (if … then) rules, or neural networks.

For the prediction of the class label of each data instance, this model is used in the test set and the trained model is assumed to be applicable in future cases. Decisions can be made with the help of predictions made. Knowledge discovery in databases predicts unknown or future values of some attributes based on other attribute values in the database through predictive modeling (Masand & Shapiro, 1996). Predefined classes of any instance could be predicted through classification in data mining techniques.

1.3.2 Descriptive task

The objective of this task is to derive patterns to summarize the underlying relationship amongst data.

Clustering is studied in the field of machine learning as a descriptive learning process, as it is “learning from observation” and not “learning from examples.” The pattern proximity matrix is being measured as a distance function defined on pairs of patterns (Jain &Dubes, 1988; Duda et al., 2001). Clustering gives an overview of a given data set. Preprocessing for other data mining algorithms is another important use of clustering algorithms.

1.4 Models for Data mining Tasks

There are three models to accomplish the data mining task.

1) **Single model**

Where the problem is being solved by applying a single technique that is classification, clustering, or association.

2) **Hybrid methods**

Heterogeneous Machine Learning approaches like, clustering and classification or clustering are combined in Hybrid learning.

3) **Ensemble methods**

Homogenous Machine learning approaches are combined in Ensemble learning, using various merging methods.

A Different number of techniques are used in data mining, but different types of data need different techniques that are specifically applicable to them. Hence, only a single technique cannot be used for proper and thorough data mining studies, as every technique has its specific application. The most widely used data mining techniques are Classification and Clustering.

Clustering gives an overview of a given data set. Clustering does not require a specification of a set of examples, which is its special feature. Thus, clustering is applied in applications in which no or little prior knowledge of the groups or classes in a database is available. The clustering usefulness is often associated with individual interpretation and on the selection of suitable similarity measure.

Classification is necessary to segregate data into predefined classes. It depends on the attributes and features present in the data. Users can get a better understanding and description of the data for each class of the database.

1.5 Data Mining Tools

Through evolution like the data & advanced analytical techniques are being continuously developed. This led to the assortment of tools designed for the performance of either comprehensive data mining or a specific data mining task. Such tools are available in a wide range of formats (Rangra & Bansal, 2014).

The data mining algorithms development and application require the use of very powerful software tools. The choice of the tools keeps on increasing because the availability of them grows continuously; this makes it difficult to select a suitable tool (Frank et al. ,2004).

1.5.1 WEKA

The WEKA workbench is a machine learning and data preprocessing tool, under the GNU General Public License. WEKA, acronym is Waikato Environment for Knowledge Analysis, was developed at the University of Waikato in New Zealand. It is written in Java and can run on Linux, Windows, and Macintosh operating systems. The current stable version, 3.8.0, is compatible with Java 1.7.

WEKA provides the support for the whole data mining process, *viz.,* and preparation of the input data by data transformation and preprocessing, analyzing the data using learning schemes, and visualizing the data. All the features are accessible by interactive interface; it does not require users to have any prior programming knowledge. WEKA can processes data in Attribute Relation File Format (ARFF), convert other file formats to ARFF format. This support exists for files of XML Relation File Format (XRFF), C4.5 format, LIBSVM format, CSV format, JSON based ARFF format, and MATLAB files.

WEKA's a graphical interface; its Explorer has panels corresponding to different data mining tasks supported by WEKA, like preprocessing, clustering, association rule mining, attribute selection, classification and regression, and data visualization. These panels are accessed by selecting the appropriate tabs. Each panel has menus to select the desired methods. Each method has a property form (or object editor) to assign values to method-specific parameters. The formation of the plug-in from WEKA deploys new features. Several graphical user interfaces are present in WEKA that easy access to the functionalities underlying. The main graphical user interface is the “Explorer.

Three graphical user interfaces are provided in WEKA i.e. the Explorer for exploratory data analysis to support preprocessing, visualization, learning, attribute selection, in the experimental environment of testing and evaluating machine learning algorithms, and the Knowledge Flow for new process model inspired interface for the visual design of KDD process (Frank et al. ,na). A simple Command-line explorer for typing commands is also provided by WEKA.

1.5.2 ORANGE

Orange is a component-based data mining and machine learning software suite, it features visual programming for explorative data analysis that is existing in the front end and helps in visualization, libraries for scripting and Python bindings. A set of components are used for data pre-processing, filtering, feature scoring, model evaluation, modeling, and exploration techniques. C++ and Python are implemented in their study. On a cross-platform framework, the graphical user interface is built up & Developed in 2009.

Orange 2.7, is the latest version, Licensed by GNU General Public License cross-platform GUI that is compatible with Python, C++, & C. Data mining in Orange is done through Python or visual programming. It is an Open source data visualization and analysis tool. It is done through Python scripting. For visual programming and explorative data analysis, a data mining tool is useful. Orange has widgets, supported on macOS, Windows, and Linux platforms (Demsar et al. ,2004).

1.6 Problem Statement and Objectives

1.6.1 Problem statement

The aim of this study is to design and develop an integrated framework using clustering and classification methods to improve the performance of the classifier when the data is imbalanced, high dimensional , Noisy and incomplete .

1.6.2 Research objectives

The objective of this study is to develop an accurate, simple and understandable model by combining supervised (Classification) and unsupervised (Clustering) learning methods. This study has proved that the integrated model does not only improve the accuracy of the classifier but it also handles some challenging issues that have existed in Data mining areas since a long time.

The integration of these methods can serve the following purposes.

1. To develop a model that is capable in handling high dimensional data

Based on clustering criteria features can be clustered together. It reduces dimensionality, complexity, computation time, increases comprehensibility and the overall performance of classification algorithms.

2. To develop a model that is capable in handling Imbalanced data

The main purpose of this approach is to selectively discard majority instances from the dataset to make the distribution balanced, which can be applied to any traditional classifier.

3. To develop a model that can improve the performance of the classifiers when the data set is noisy

The objective is to utilize the strength of one method to complement the weaknesses of another. If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as an integrated model on noisy dataset.

1.6.3 Research Contribution

Three models are presented and implemented that offer a generic and logical solution to meet up the research objectives.

The first proposed model provides a Cluster-based approach using an under-sampling solution to balance the imbalanced data. A study for binary class distribution on 12 data sets Abalone, Cleveland-0, Ecoli-3, Glass-1, Haberman, New-Thyroid-1, Page-Blocks-0, Pima, Wine Quality White, Breast Cancer, Wisconsin, Yeast 1, and Vowel openly available in UCI machine learning repository with different degrees of imbalance nature has been conducted. The proposed framework is capable in giving a three-fold solution.

Firstly, it balances imbalanced data using the K-means clustering approach. The main purpose of this approach is to selectively discard majority instances from the dataset to make the distribution balanced and can be applied to any traditional classifier. Secondly, it can handle between-class imbalance distribution and within-class distribution. Thirdly it can handle the different degrees of imbalance distribution. The proposed model is simple yet effective in classifying the imbalanced distribution of Data.

The second proposed model is applicable as a Feature compression and Extraction technique to build a better feature space because it can solve several machine learning problems. High dimensional data generally diminish the accuracy and efficiency of Data Mining algorithms. The higher the dimensionality, the higher the computation cost involved in processing. Irrelevant features may exert undue burden on classification evaluation parameters and increases the time and resources needed to build the model.

The advantage of the proposed model is, it first identifies the relevant features that may lead to accurate results. The experiments were conducted on nine benchmark dataset taken from UCI machine repository with diverse degrees of dimensionality they are Heart Disease, Wine quality white, Parkinson dataset, Colon Cancer, Breast Cancer, Image Segmentation, Cleveland-0, Ionosphere, and Squash Harvest Stored. The comparative analysis of the proposed approach with a standard correlation-based Feature Selection approach (RELIEF and Info-Gain Approach) was presented on the aforementioned datasets. The analysis was presented on 20, 40, 60, 80, and 100 % of the features on three performance measuring parameters; Accuracy, F-Measure, and ROC. The K-means algorithm is used to group similar features. The centroid of each cluster has been taken as the individual cluster representative. It reduces dimensionality, complexity, and computation time, and increases comprehensibility and the overall performance of classification algorithms. Therefore, the proposed model first identifies the relevant features that may lead to accurate results.

The third proposed model is to improve the performance of any classifier using a hybrid model (prior clustering to classification). The research experiments observed that the hybrid model is more refined and accurate than a single individual classifier. The objective is to utilize the strength of one method to complement the weaknesses of another. If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as a hybrid model.

The proposed model consists of three steps. Clustering Step: Prior to the classification, the dataset is grouped into the different number of clusters using the K-means algorithm depending upon the value of K (K=2, 3, 4, and 5). Expansion step: The resultant dataset is augmented with one more feature originated from the clustering step named cluster number as an input to the classifier. Classification step: The expanded data is applied to SVM (Support Vector Machine) classifier. The experiment has been conducted on 13 benchmark dataset taken from UCI machine learning repository, these are- Heart Disease dataset, Wine quality white dataset, Parkinson dataset, Colon Cancer dataset, Breast Cancer dataset, Image

Segmentation dataset, Cleveland-0 Dataset, Ionosphere Dataset, Squash Harvest Stored dataset, Bank Dataset, Glass Dataset, Pima dataset, and Haberman Dataset. The results revealed that this integration process generates a more precise and accurate model. The major classifier's accuracy gets affected positively when supervised and unsupervised learning methods are combined.

1.7 Organization of the thesis

The contributions from this study have been presented in the chapters as follows:

Chapter-1: This chapter contains the general introduction of the broad area of the research it contains an overview of Data mining, Data mining application areas, Data mining process, Data mining tasks, models for Data mining tasks, classification and clustering techniques and tools.

Chapter-2: An extensive literature review has been accomplished and presented in this chapter.

Chapter -3: A detailed description of Clustering and Classification and the techniques used to perform these tasks are presented in this chapter.

Chapter-4: This chapter demonstrates the application of this proposed integrated approach in handling Imbalance class distribution in real-world applications which is one of the top challenging problems in data mining.

Chapter -5: This Chapter contains the third proposed model which is applicable as a Feature compression and Extraction technique to build a better feature space because it can solve several machine learning problems.

Chapter-6: This chapter proposed an integrated approach which is applicable in various aspects of data mining in this chapter model is deployed to use clustering as a preprocessing process for classification. It reduces the training data set size and its dimensionality by dividing the whole Dataset into smaller sub-sets leads to a smaller and less complicated classification task becomes quicker and easier to solve.

Chapter-7 Conclusion & Future work

Chapter-8 References.