**Assignment 01**

**KDD Analysis**

Name: Puja Kawale

SJSUID: 0125-06156

Date: 09/29/2018

**Knowledge discovery in Databases (KDD)**: KDD is similar concept like Data Mining. **Data mining** is finding hidden, often useful data from larger Datasets or Datawarehouse’s to generate models/patterns/rules in order to achieve some intelligence into the business.

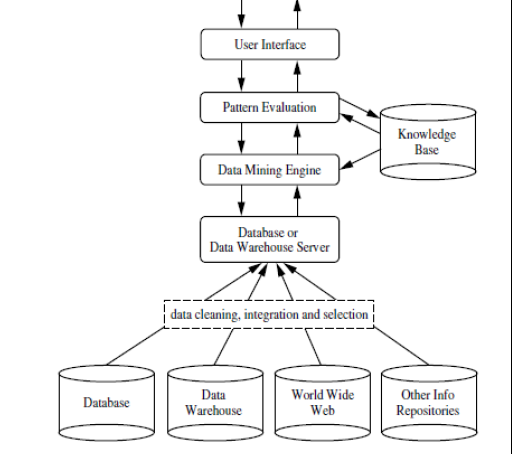


Fig 1. Architecture Diagram of Datamining/KDD

Source 1: Data Mining Concepts and Techniques JaiweiHan and MachelineKambler

Source 2: Lecture Slides

**Steps to achieve KDD:**

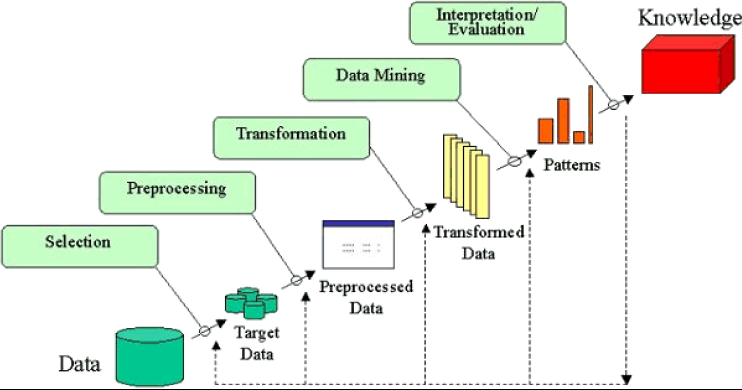


Fig 2: Steps to achieve KDD

*Source: 1* Overview of the KDD process -http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1\_kdd.html/

Source 2. Lecture Slides

**Problem Statement:**

After completion of your master’s program at San Jose State University, you have secured a job at local financial institution that issues credit cards to qualified customers. During the first week of your job, your supervisor, a senior data analytics manager, asked you to attend a regular internal data committee meeting that is developing web-based application to identify fraud process in credit card usage. The data committee has listed following attributes for the Credit card process:

**Demo-graphical Data** about Credit Card (CC) User: Name, age, location, address, phone number, email, Driving License Details, Income details

**Customer Behavior**: Credit Card Location, Amount Spent, Time of Usage, Date of Usage, Day of Usage, Items purchased, Quantity

**Social Media Data**: CC User Email, Web Posts, Sentiment of Post, Location of Post

Based on the above details, your manager wanted you to develop KDD process to identify usage behaviors that could be fraudulent? How do you approach it?

Please note: a credit company considers a transaction to be fraudulent, until unless proven otherwise, under following conditions:

* Credit Card transaction performed at odd hours of day (for instance mid-night 2 am)
* Credit Card transaction performed with unusual large amounts (for instance, you tend to spend in the range of $25 - $300 per transaction, any transaction above $500 by your spending behavior considers to be fraudulent)
* Customer behavior pattern

Solution:

According to the KDD steps mentioned in our description I would follow below approach :

1. Developing an understanding of

* the application domain
* the relevant prior knowledge
* the goals of the end-user

[*https://en.wikipedia.org/wiki/Credit\_card*](https://en.wikipedia.org/wiki/Credit_card)

[*https://en.wikipedia.org/wiki/Payment\_card\_industry*](https://en.wikipedia.org/wiki/Payment_card_industry)

[*https://en.wikipedia.org/wiki/Payment\_gateway*](https://en.wikipedia.org/wiki/Payment_gateway)

1. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

* Demo-graphical Data about Credit Card (CC) User: Name, location, address, phone number, email
* Customer Behavior: Credit Card Location, Amount Spent, Time of Usage, Date of Usage, Day of Usage, Items purchased, Quantity
* Social Media Data: CC User Email, Web Posts, Sentiment of Post, Location of Post

1. Data cleaning and preprocessing.

* Removal of noise or outliers.

This would include preparing my data set with removal of additional meaningless information which is corrupted. This would be the values/cells containing unstructured texts. Statistical analysis would also help me with removing the noise out of data. In this step of KDD I would also consider eliminating Outliers not belonging to the dataset. This is generally due to human error (mislabeling, programming defects, transposing).

* Collecting necessary information to model or account for noise.

In our use case, I would consider –

Demo-graphical Data about Credit Card (CC) User: Name, location

Customer Behavior: Credit Card Location, Amount Spent, Time of Usage, Date of Usage, Day of Usage, Items purchased, Quantity

* Strategies for handling missing data fields.

Know if the values are missing.

-get a count of the missing values for each column.

-evaluate an LDA model on the dataset using k-fold cross validation: Here, an error is expected and user should be prevented from using any such algorithm on the dataset with missing values.

Handle missing values.

-Removal of rows with missing values

-Impute missing values(meaningful constant value, value from another random record, mean/median/mode of other values, estimated value)

-Algorithm that supports missing values: k-Nearest Neighbors, classification and regression trees.

* Accounting for time sequence information and known changes.

**Demo-graphical Data** about Credit Card (CC) User: Name, age, location, address, phone number, email, Driving License Details, Income details

**Customer Behavior**: Credit Card Location, Amount Spent, Time of Usage, Date of Usage, Day of Usage, Items purchased, Quantity

**Social Media Data**: CC User Email, Web Posts, Sentiment of Post, Location of Post

1. Data reduction and projection.

* Finding useful features to represent the data depending on the goal of the task.

In this step, I would basically make choice of features within the data frame to visualize on the plot. Example –

1. Age Vs Item purchased
2. location Vs credit card location
3. Name Vs amount spent
4. Name Vs Quantity
5. Name Vs Time of Usage

* Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

Using R, python, pandas to create a subset of dataset used. This would reduce the number of features considered to represent data.

Eg:

features\_to\_be\_included = [‘Name’, ‘Age’, ‘Location’]

dataframe = dataframe[features\_to\_be\_included]

dataframe [:2]

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Name** | **Age** | **Location** |
| **0** | Kent | 37 | 4529, 1062 N |
| **1** | John | 42 | 1114, 1065 N |

1. Choosing the [data mining task](http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/2_tasks.html).

* Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

As per my understanding, in this use case I would go with Clustering (K-means, GMM) and Classification to detect any fraudulent transaction using the credit card distributed by my company.

1. Clustering: partitioning/grouping data into meaningful subclasses.
2. Classification: categorize data to cater analysis or prediction to take decisions in business.
3. Choosing the [data mining algorithm(s)](http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/3_algs_and_methods.html).

* Selecting method(s) to be used for searching for patterns in the data.
* Deciding which models and parameters may be appropriate.
* Matching a particular data mining method with the overall criteria of the KDD process.

1. Data mining.

* Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

1. Interpreting mined patterns.
2. Consolidating discovered knowledge.

References:

1. Lecture Slides from
2. <http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html>