Medical Tracker

Project Scope

- Information Intensive Industry.
- Data keeps growing on a daily basis.
- An acute care hospital may generate five terabytes of data a year.
- Pattern recognition from such huge data is important for the diagnosis of diseases.

- Computer assisted information retrieval.
- Human decision-making is poor when there are huge amounts of data to be classified.
- This lead to the use of data mining in medical informatics.

Problem Definition

- Medical information systems are designed to generate simple statistics.
- They can answer simple queries.
- However, they cannot answer complex queries.
- Existing system does not have provision for classification and prediction.

- Clinical decisions are often made based on doctors intuition and experience.

- The knowledge rich data hidden in the database is not analysed.

- This practice leads to unwanted biases, errors.

Project Objectives

- Highlight the importance of data mining in medicine and public health.
- To find data mining techniques used in other fields that may also be applied in the health sector.
- To identify issues and challenges in data mining as applied to the medical practise.
- To outline some recommendations for discovering knowledge in electronic databases.

Methodology

- The first phase focuses on understanding the project objectives and requirements, then converting this knowledge into a data mining problem definition.

- Second phase starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data.

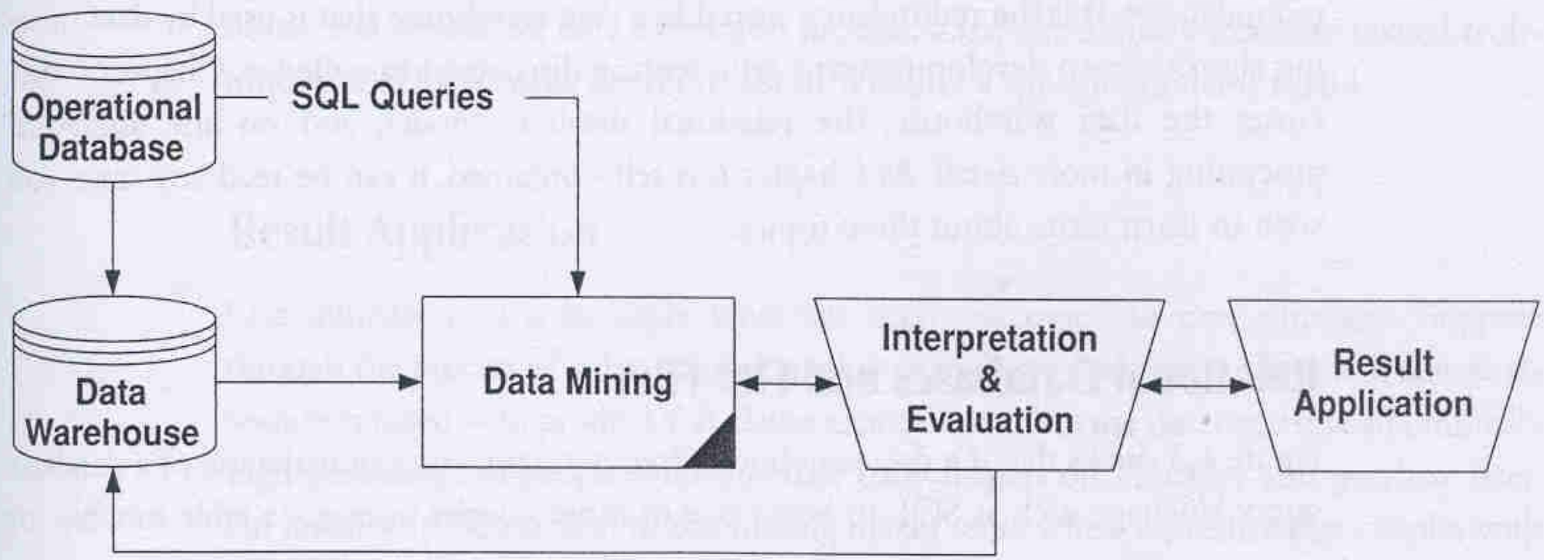
- Various modelling techniques are selected and applied and optimal solutions are chosen. The model is thoroughly evaluated and reviewed.

- The knowledge gained will need to be organized and presented in a way so that it be used in decision making.

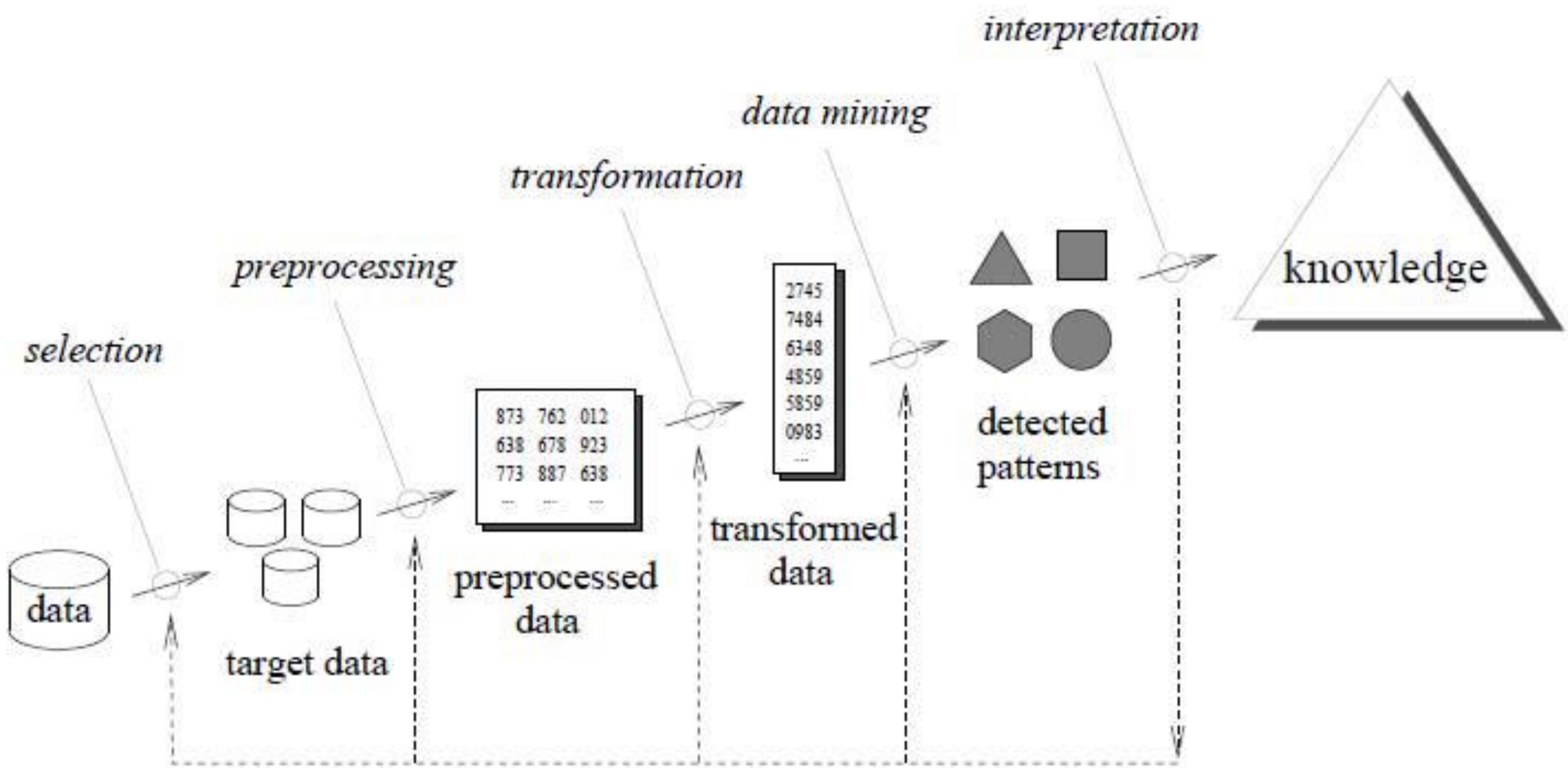
- Diagnosis :- Assist in decision making with a large number of inputs and in stressful situations.
- Therapy :- Based on modeled historical performance, select best intervention course.
- Prognosis :- Accurate prognosis and risk assessment are essential for improved disease management and outcome.
- Hospital Management :- Forecasting patient volume, ambulance run volume, etc.

Data Mining Process

Figure 1.3 ● **A simple data mining process model**



Knowledge Discovery in Databases



Supervised vs. Unsupervised Learning

- **Supervised learning** ***(classification)***
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- **Unsupervised learning** ***(clustering)***
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data.

Supervised Learning

Table 1.1 • **Hypothetical Training Data for Disease Diagnosis**

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

Unsupervised Learning

Table 1.2 • **Data Instances with an Unknown Classification**

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
11	No	No	Yes	Yes	Yes	?
12	Yes	Yes	No	No	Yes	?
13	No	No	No	No	Yes	?

Naïve Bayes

- Each data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes.
- Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability

Naïve Bayes

- That is, the naive probability assigns an unknown sample X to the class Ci if and only if: $P\ (Ci/X) > P\ (Cj/X)$.
- Thus we maximize $P\ (Ci|X)$. The class Ci for which $P(Ci|X)$ is maximized is called the maximum posteriori hypothesis.

  By Bayes theorem,

  $P\ (Ci/X) = (P(X/Ci)P(Ci))/P(X)$.

Dataset

age	sex	chest_pain	resting_bp	serum_c	fasting_sugar	thalach	result
67	0	3	115	564	0	160	1
57	1	2	124	261	0	141	2
64	1	4	128	263	0	105	1
74	0	2	120	269	0	121	1
65	1	4	120	177	0	140	1
56	1	3	130	256	1	142	2
59	1	4	110	239	0	142	2
60	1	4	140	293	0	170	2
63	0	4	150	407	0	154	2
59	1	4	135	234	0	161	1
53	1	4	142	226	0	111	1
44	1	3	140	235	0	180	1
61	1	1	134	234	0	145	2
57	0	4	128	303	0	159	1
71	0	4	112	149	0	125	1
46	1	4	140	311	0	120	2
53	1	4	140	203	1	155	2
64	1	1	110	211	0	144	1
40	1	1	140	199	0	178	1
67	1	4	120	229	0	129	2
48	1	2	130	245	0	180	1
43	1	4	115	303	0	181	1
47	1	4	112	204	0	143	1
54	0	2	132	288	1	159	1
48	0	3	130	275	0	139	1

Output

```
Enter The Data :Age
56
Enter The Data : Sex 0:Female 1:Male
1
Enter The Data :Chest pain
 1:typical type 1 angina
 2:typical type angina
 3:non-angina pain
 4:asymptomatic
3
Enter The Data :Resting Blood Pressure:
130
Enter The Data :Serum Cholestoral in mg/dl
256
Enter The Data :Fasting Blood Sugar
 1:>120mg/dl.
 0:<120 mg/dl
1
Enter The Data :Maximum heart rate achieved
142
Connection Established
Current Status:
Result:2-Has HEART disease
54.126144% HEART RISK
```

```
Enter The Data :Age
63
Enter The Data : Sex 0:Female 1:Male
0
Enter The Data :Chest pain
 1:typical type 1 angina
 2:typical type angina
 3:non-angina pain
 4:asymptomatic
4
Enter The Data :Resting Blood Pressure:
140
Enter The Data :Serum Cholestoral in mg/dl
293
Enter The Data :Fasting Blood Sugar
 1:>120mg/dl.
 0:<120 mg/dl
0
Enter The Data :Maximum heart rate achieved
170
Connection Established
Current Status:
Result:1-NO HEART disease
47.838203% HEART RISK
```

- Some publicly available datasets
  - UCI Machine Learning Repository
  - KDD Cup 2008 -Siemens (*Requires registration*)
  - MIT-BIH Arrhythmia Database
  - ECML/PKDD discovery challenge dataset.
  - Healthcare Cost and Utilization Project (H-CUP)
  - HIV Prevention Trials Network - Vaccine Preparedness Study/Uninfected Protocol Cohort
  - National Trauma Data Bank (NTDB)
  - Behavioral Risk Factor Surveillance System (BRFSS)
  - Link to National Public Health Data Sets

(http://www-users.cs.umn.edu/~desikan/pakdd2011/datasets.html)

References

- [1] Awang, R. & Palaniappan, S., “Intelligent heart disease predication system using data mining technique”. IJCSNS International Journal of Computer Science and Network Security. Vol. 8, No. 8, 2008.
- [2] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction” IJCSE Vol. 3 No. 6 June 2011.
- [3] G. Parthiban, A. Rajesh, S.K.Srivatsa “Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method”.
- [4] Eapen, A. G. (2004). Application of Data mining in Medical Applications. Ontario, Canada, 2004: University of Waterloo.