DATA MINING

PAST PAPER SOLUTIONS

"2015"

1. **(a) What is Data mining? Name different technologies integrated in Data Mining?**

A1: Data mining refers to extracting or mining" knowledge from large amounts of data. There are many other terms related to data mining, such as knowledge mining, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD.

It is a multi-disciplinary skill that uses machine learning, statistics, Artificial Intelligence and database technology.

1(b).A human can also find meaningful information from given data then why do we need to determine data through machine?

A2: Data mining refers to extracting knowledge from a large amount of data. Data mining is the process to discover various types of patterns that are inherited in the data and which are accurate, new and useful. Data mining is the subset of business analytics, it is similar to experimental research. The origins of data mining are databases, statistics. Machine learning involves an algorithm that improves automatically through experience based on data. Machine learning is a way to discover a new algorithm from the experience. Machine learning involves the study of algorithms that can extract information automatically. Machine-learning uses data mining techniques and another learning algorithm to build models of what is happening behind some data so that it can predict future outcomes.

1. To implement data mining techniques, it used two-component first one is the database and the second one is machine learning. The Database offers data management techniques while machine learning offers data analysis techniques. But to implement machine learning techniques it used algorithms.
2. Data mining uses more data to extract useful information and that particular data will help to predict some future outcomes for example in a sales company it uses last year data to predict this sale but machine learning will not rely much on data it uses algorithms, for example, OLA, UBER machine learning techniques to calculate the ETA for rides.
3. Self-learning capacity is not present in data mining, it follows the rules and predefined. It will provide the solution for a particular problem but machine learning algorithms are self-defined and can change their rules as per the scenario, it will find out the solution for a particular problem and it resolves it by its own way.

4. The main and foremost difference between data mining and machine learning is, without the involvement of human data mining can't work but in machine learning human effort is involved only the time when algorithm is defined after that it will conclude everything by own means once implemented forever to use but this is not the case with data mining.
5. The result produces by machine learning will be more accurate as compared to data mining since machine learning is an automated process.
6. Data mining uses the database or data warehouse server, data mining engine and pattern evaluation techniques to extract the useful information whereas machine learning uses neural networks, predictive model and automated algorithms to make the decisions.

1(c). Name any four algorithms normally used in Data mining?

A3: Following are the name of Algorithms normally used in Data Mining:

1. Decision Tree Algorithm
2. FP Growth Algorithm.
3. Apriori Algorithm
4. Naïve Bayes Algorithm.
5. KNN Algorithm.
6. K means Clustering Algorithm.
7. Regression Algorithm.
8. Association Rule based of Decision tree.

2(a). Name the Steps followed to discover the knowledge from a database?

A4: Knowledge discovery as a process consists of an iterative sequence of the following steps:

1. **data cleaning:** to remove noise or irrelevant data
2. **data integration:** where multiple data sources may be combined
3. **data selection:** where data relevant to the analysis task are retrieved from the database
4. **data transformation:** where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations
5. **data mining**: an essential process where intelligent methods are applied in order to extract data patterns
6. **pattern evaluation** to identify the truly interesting patterns representing knowledge based on some interestingness measures
7. **knowledge presentation:** where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

2(b). Describe the multidimensional views of data mining?

A5: Multi-Dimensional View of Data Mining

- **Data to be mined:** Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

2(c). Describe a data mining example from market analysis and decision support?

A6: Consider a marketing head of telecom service provides who wants to increase revenues of long distance services. For high ROI on his sales and marketing efforts customer profiling is important. He has a vast data pool of customer information like age, gender, income, credit history, etc. But its impossible to determine characteristics of people who prefer long distance calls with manual analysis. Using data mining techniques, he may uncover patterns between high long distance call users and their characteristics.

For example, he might learn that his best customers are married females between the age of 45 and 54 who make more than $80,000 per year. Marketing efforts can be targeted to such demographic.

Another Answer in detail for other domains

The predictive capacity of data mining has changed the design of business strategies. Now, you can understand the present to anticipate the future. These are some examples of data mining in current industry.

- **Marketing.** Data mining is used to explore increasingly large databases and to improve market segmentation. By analysing the relationships between parameters such as customer age, gender, tastes, etc., it is possible to guess their behaviour in order to direct personalised loyalty campaigns.Data mining in marketing also predicts which users are likely to unsubscribe from a service, what interests them based on their searches, or what a mailing list should include to achieve a higher response rate.
- **Retail.** Supermarkets, for example, use joint purchasing patterns to identify product associations and decide how to place them in the aisles and on the shelves. Data mining also detects which offers are most valued by customers or increase sales at the checkout queue.
- **Banking.** Banks use data mining to better understand market risks. It is commonly applied to credit ratings and to intelligent anti-fraud systems to analyse transactions, card transactions, purchasing patterns and customer financial data. Data mining also allows banks to learn more

- about our online preferences or habits to optimise the return on their marketing campaigns, study the performance of sales channels or manage regulatory compliance obligations.
- **Medicine.** Data mining enables more accurate diagnostics. Having all of the patient's information, such as medical records, physical examinations, and treatment patterns, allows more effective treatments to be prescribed. It also enables more effective, efficient and cost-effective management of health resources by identifying risks, predicting illnesses in certain segments of the population or forecasting the length of hospital admission. Detecting fraud and irregularities, and strengthening ties with patients with an enhanced knowledge of their needs are also advantages of using data mining in medicine.
- **Television and radio.** There are networks that apply real time data mining to measure their online television (IPTV) and radio audiences. These systems collect and analyse, on the fly, anonymous information from channel views, broadcasts and programming. Data mining allows networks to make personalised recommendations to radio listeners and TV viewers, as well as get to know their interests and activities in real time and better understand their behaviour. Networks also gain valuable knowledge for their advertisers, who use this data to target their potential customers more accurately.

3(a). Data mining has potential application in risk analysis and management, name any eight application?

A7:

Applications	Usage
Communications	Data mining techniques are used in communication sector to predict customer behavior to offer highly targetted and relevant campaigns.
Insurance	Data mining helps insurance companies to price their products profitable and promote new offers to their new or existing customers.
Education	Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in maths subject.
Manufacturing	With the help of Data Mining Manufacturers can predict wear and tear of production assets. They can anticipate maintenance which helps them reduce them to minimize downtime.
Banking	Data mining helps finance sector to get a view of market risks and manage regulatory compliance. It helps banks to identify probable defaulters to decide whether to issue credit cards, loans, etc.
Retail	Data Mining techniques help retail malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owners to comes up with the offer which encourages customers to increase their spending.
Service Providers	Service providers like mobile phone and utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyze billing

	details, customer service interactions, complaints made to the company to assign each customer a probability score and offers incentives.
E-Commerce	E-commerce websites use Data Mining to offer cross-sells and up-sells through their websites. One of the most famous names is Amazon, who use Data mining techniques to get more customers into their eCommerce store.
Super Markets	Data Mining allows supermarket's develope rules to predict if their shoppers were likely to be expecting. By evaluating their buying pattern, they could find woman customers who are most likely pregnant. They can start targeting products like baby powder, baby shop, diapers and so on.
Crime Investigation	Data Mining helps crime investigation agencies to deploy police workforce (where is a crime most likely to happen and when?), who to search at a border crossing etc.
Bioinformatics	Data Mining helps to mine biological data from massive datasets gathered in biology and medicine.

3(b). What are the types of data sources? Give an example for health related services.

8A: Data mining can be performed on following types of data

- Relational databases
- Data warehouses
- Advanced DB and information repositories
- Object-oriented and object-relational databases
- Transactional and Spatial databases
- Heterogeneous and legacy databases
- Multimedia and streaming database
- Text databases
- Text mining and Web mining

Health care domain and insurance domain:

The data mining related applications can be used to efficiently track and monitor a patient's health condition and also can help in efficient diagnosis based on the past sickness record. In a similar manner,

the growth of the insurance industry depends on the ability to convert the data into knowledge form or by providing various details about the customers, markets and the prospective competitors and therefore all those companies who have applied the data mining techniques efficiently have reaped the benefits. This is applied over the claims and their analysis i.e. identification of the medical procedures which are claimed together. It enables the forecasting of new policies, helps in the detection of risky customer behavior patterns and also helps in the detection of fraudulent behavior.

3(c). Describe the major issues in mining methodology?

A9. Major issues in Mining Methodology:

Mining methodology and user-interaction issues:

_ Mining different kinds of knowledge in databases: Since different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

_ Interactive mining of knowledge at multiple levels of abstraction: Since it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive.

_ Incorporation of background knowledge: Background knowledge, or information regarding the domain under study, may be used to guide the discovery patterns. Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

_ Data mining query languages and ad-hoc data mining: Knowledge in Relational query languages (such as SQL) required since it allow users to pose ad-hoc queries for data retrieval.

_ **Presentation and visualization of data mining results:** Discovered knowledge should be expressed in high-level languages, visual representations, so that the knowledge can be easily understood and directly usable by humans

_ **Handling outlier or incomplete data:** The data stored in a database may reflect outliers: noise, exceptional cases, or incomplete data objects. These objects may confuse the analysis process, causing over fitting of the data to the knowledge model constructed. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods which can handle outliers are required.

_ **Pattern evaluation: refers to interestingness of pattern**: A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, representing common knowledge or lacking novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns,

4(a). Why do we need to pre process a data before mining it?

A10: Data in the real world is dirty. It can be in incomplete, noisy and inconsistent from. These data needs to be preprocessed in order to help improve the quality of the data, and quality of the mining results.

- If no quality data , then no quality mining results. The quality decision is always based on the quality data.
- If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult

Incomplete data: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., occupation=“ ”.
Noisy data: containing errors or outliers data. e.g., Salary=“-10”
Inconsistent data: containing discrepancies in codes or names. e.g., Age=“42” Birthday=“03/07/1997”

4(b). Name different measures taken to improve data quality, briefly describe any two?

A11. Measures taken to improve data quality are:

Data cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

- Integration of multiple databases, data cubes, or files

Data transformation

- Normalization and aggregation

Data reduction

- Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization

- Part of data reduction but with particular importance, especially for numerical data

4(c). What is mean by incomplete data?

A12: Incomplete data: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., occupation=“ ”.

- Incomplete data may come from
- "Not applicable" data value when collected
- Different considerations between the time when the data was collected and when it is analyzed.
- Human/hardware/software problems

5(a). Name the major tasks of data pre processing? Describe data cleaning.

A13: Major tasks in data pre processing are:

Data cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

- Integration of multiple databases, data cubes, or files

Data transformation

- Normalization and aggregation

Data reduction

- Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization

- Part of data reduction but with particular importance, especially for numerical data

5(b). What are the main causes of data missing?

A14: Main causes of data missing are:

- Incomplete data may come from

➢"Not applicable" data value when collected
➢Different considerations between the time when the data was collected and when it is analyzed.
➢Human/hardware/software problems

- Noisy data (incorrect values) may come from
- ➢ Faulty data collection by instruments
- ➢ Human or computer error at data entry
- ➢ Errors in data transmission

5(c). How do we handle missing data?

A15: The various methods for handling the problem of missing values in data tuples include:
(a) Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
(b) Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.
(c) Using a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like "Unknown," or $-\infty$. If missing values are replaced by, say, "Unknown," then

the mining program may mistakenly think that they form an interesting concept, since they all have a value in common — that of "Unknown." Hence, although this method is simple, it is not recommended.
(d) Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

(e) Using the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

6(a). Define the equal width and equal depth discretization of data?

A16:

- **Equal-width (distance) Discretization:**

– Divides the range into N intervals of equal size: uniform grid

– if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B – A)/N.

– The most straightforward, but outliers may dominate presentation – Skewed data is not handled well

● **Equal-depth (frequency) Discretization:**

– Divides the range into N intervals, each containing approximately same number of samples

– Good data scaling

– Managing categorical attributes can be tricky

6(b). What is mean by data integration? How does conflict in data values resolved?

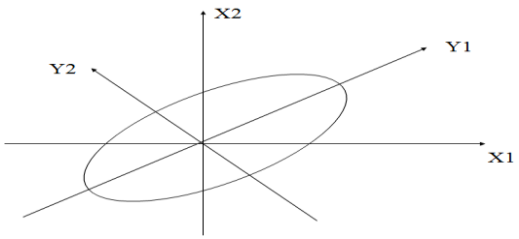
A17: Integration of multiple databases, data cubes, or files is known as Data Integration. It combines data from multiple sources into a coherent store.
v'
These techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Data
reduction includes,
1. **Data cube aggregation**, where aggregation operations are applied to the data in the construction of a data cube.
2. **Attribute subset selection**, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
3. Dimensionality reduction, where encoding mechanisms are used to reduce the data set size.
Examples: Wavelet Transforms Principal Components Analysis

4. **Numerosity reduction**, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.
5. Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data Discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies.

6(c). Define principal component analysis?

A18: Principal Component Analysis (PCA) -also called as Karhunen-Loeve (K-L) method Procedure
• Given *N* data vectors from *k*-dimensions, find *c* <= *k* orthogonal vectors that can be best used to represent data
– The original data set is reduced (projected) to one consisting of N data vectors on c principal components (reduced dimensions)
• Each data vector is a linear combination of the *c* principal component vectors • Works for ordered and unordered attributes
• Used when the number of dimensions is large

The principal components (new set of axes) give important information about variance. Using the strongest components one can reconstruct a good approximation of the original signal.



“2016”

1A: Classification vs. Clustering

- In general, in classification you have a set of predefined classes and want to know which class a new object belongs to.
- Clustering tries to group a set of objects and find whether there is *some* relationship between the objects.
- In the context of machine learning, classification is *supervised learning* and clustering is *unsupervised learning*.

1B: A frequent itemset is one which is made up of one of these patterns, which is why frequent pattern mining is often alternately referred to as frequent itemset mining. Frequent pattern mining is most easily explained by introducing market basket analysis, a typical usage for which it is well-known.

A market basket is a collection of items purchased by a customer in a single transaction, which is a well-defined business activity. For example, a customer's visits to a grocery store or an online purchase from a virtual store on the Web are typical customer transactions. Retailers accumulate huge collections of transactions by recording business activities over time. One common analysis run against a transactions database is to find sets of items, or *itemsets*, that appear together in many transactions. A business can use knowledge of these patterns to improve the Placement of these items in the store or the layout of mail- order catalog page and Web pages. An itemset containing *i* items is called an *i-itemset*. The percentage of transactions that contain an itemset is called the itemsets *support*. For an itemset to be interesting, its support must be higher than a user-specified minimum. Such itemsets are said to be frequent.

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for association Rule means that 2% of all the transactions under analysis show that computer and financial management software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold.

1C: Five primitives for specifying a data mining task

• **Task-relevant data:** This primitive specifies the data upon which mining is to be performed. It involves specifying the database and tables or data warehouse containing the relevant data, conditions for selecting the relevant data, the relevant attributes or dimensions for exploration, and instructions regarding the ordering or grouping of the data retrieved.

• **Knowledge type to be mined:** This primitive specifies the specific data mining function to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. As well, the user can be more specific and provide pattern templates that all discovered patterns must match. These templates or meta patterns (also called meta rules or meta queries), can be used to guide the discovery process.

• **Background knowledge:** This primitive allows users to specify knowledge they have about the domain to be mined. Such knowledge can be used to guide the knowledge discovery process and evaluate the patterns that are found. Of the several kinds of background knowledge, this chapter focuses on concept hierarchies.

• **Pattern interestingness measure:** This primitive allows users to specify functions that are used to separate uninteresting patterns from knowledge and may be used to guide the mining process, as well as to evaluate the discovered patterns. This allows the user to confine the number of uninteresting patterns returned by the process, as a data mining process may generate a large number of patterns. Interestingness measures can be specified for such pattern characteristics as simplicity, certainty, utility and novelty.
• **Visualization of discovered patterns:** This primitive refers to the form in which discovered patterns are to be displayed. In order for data mining to be effective in conveying knowledge to users, data mining systems should be able to display the discovered patterns in multiple forms such as rules, tables, cross tabs (cross-tabulations), pie or bar charts, decision trees, cubes or other visual representations.

1D: Major issues in data mining

Major issues in data mining is regarding mining methodology, user interaction, performance, and diverse data types

1 Mining methodology and user-interaction issues:

_ Mining different kinds of knowledge in databases: Since different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

_ Interactive mining of knowledge at multiple levels of abstraction: Since it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive.

_ Incorporation of background knowledge: Background knowledge, or information regarding the domain under study, may be used to guide the discovery patterns. Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

_ Data mining query languages and ad-hoc data mining: Knowledge in Relational query languages (such as SQL) required since it allow users to pose ad-hoc queries for data retrieval.

_ Presentation and visualization of data mining results: Discovered knowledge should be expressed in high-level languages, visual representations, so that the knowledge can be easily understood and directly usable by humans

_ Handling outlier or incomplete data: The data stored in a database may reflect outliers: noise, exceptional cases, or incomplete data objects. These objects may confuse the analysis process, causing over fitting of the data to the knowledge model constructed. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods which can handle outliers are required.

_ Pattern evaluation: refers to interestingness of pattern: A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, representing common knowledge or lacking novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns,

2. Performance issues. These include efficiency, scalability, and parallelization of data mining algorithms.

_ Efficiency and scalability of data mining algorithms: To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.

_ Parallel, distributed, and incremental updating algorithms: Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

3. Issues relating to the diversity of database types

_ Handling of relational and complex types of data: Since relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important.

_ Mining information from heterogeneous databases and global information systems: Local and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining.

1E: Difference between Data Mining and KDD:

- Data mining is one of the steps (seventh) and the KDD process is basically the search for patterns of interest in a particular representational form or a set of these representations.
- KDD is a non-trivial process for identifying valid, new, potentially useful and ultimately understandable patterns in dat. It consists of nine steps that begin with the development and understanding of the application domain to the action on the knowledge discovered. Data mining is one of the steps (seventh) and the KDD process is basically the search for patterns of interest in a particular representational form or a set of these representations.

2.DATA MINING FUNCTIONALITIES:

Characterization is a summarization of the general characteristics or features of a target class of data. For example, the characteristics of students can be produced, generating a profile of all the University first year computing science students, which may include such information as a high GPA and large number of courses taken.

Discrimination is a comparison of the general features of target class data objects with the general

features of objects from one or a set of contrasting classes. For example, the general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.

Association is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like

major(*X; \computing science*"") *) owns*(*X; \personal computer*") [*support* = 12%*; confidence* = 98%]

where *X* is a variable representing a student. The rule indicates that of the students under study, 12% (support) major in computing science and own a personal computer. There is a 98% probability (confidence, or certainty) that a student in this group owns a personal computer.

Classification differs from prediction in that the former constructs a set of models (or functions) that describe and distinguish data classes or concepts, whereas the latter builds a model to predict some missing or unavailable, and often numerical, data values. Their similarity is that they are both tools for prediction: Classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.

Clustering analyzes data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Each cluster that is formed can be viewed as a class of objects. Clustering can also facilitate *taxonomy formation*, that is, the organization of observations into a hierarchy of classes that group similar events together.

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

3: The method that mines the complete set of frequent itemsets with candidate generation. Apriori property & The Apriori Algorithm.

To find the frequent itemsets from a transaction data set and derive association rules the Apriori algorithm is widely used. To find frequent item sets is not difficult because of its combinatorial explosion. Once we get the frequent itemsets then it is clear to generate association rules for larger or equal specified minimum confidence. Apriori is an algorithm which helps in finding frequent data sets by making use of candidate generation. It assumes that the item set or the items present are sorted in a lexicographic order. After the introduction of Apriori data mining research has been specifically boosted. It is simple and easy to implement. The basic approach of this algorithm is as below:

Join: The whole database is used for the hoe frequent 1 item sets.

Prune: This item set must satisfy the support and confidence to move to the next round for the 2 item sets.

Repeat: Until the pre-defined size is not reached till then this is repeated for each itemset level.

Apriori property

• All nonempty subsets of a frequent item set most also be frequent.

– An item set I does not satisfy the minimum support threshold, min-sup, then I is not frequent, i.e., support(I) < min-sup

– If an item A is added to the item set I then the resulting item set (I U A) can not occur more frequently than I.

• Monotonic functions are functions that move in only one direction.
• This property is called anti-monotonic.
• If a set cannot pass a test, all its supersets will fail the same test as well.
• This property is monotonic in failing the test.

The Apriori Algorithm

• Join Step: C_k is generated by joining $L_{k-1}$with itself
• Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

D, of transactions; minimum support threshold, min_sup.
ent itemsets in D.

_frequent_1-itemsets(D);
; $L_{k-1} \neq \phi$; k++) {
apriori_gen(L_{k-1}, min_sup);
ch transaction $t \in D$ { // scan D for counts
= subset(C_k, t); // get the subsets of t that are candidates
r each candidate $c \in C_t$
c.count++;

$c \in C_k | c.count \geq min\_sup\}$

$= \cup_k L_k$;

ori_gen(L_{k-1}:frequent *(k-1)*-itemsets; min_sup: minimum support
itemset $l_1 \in L_{k-1}$
ch itemset $l_2 \in L_{k-1}$
$(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \ldots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k$
$c = l_1 \bowtie l_2$; // join step: generate candidates
if has_infrequent_subset(c, L_{k-1}) then
delete c; // prune step: remove unfruitful candidate
else add c to C_k;

k;

infrequent_subset(c: candidate k-itemset; L_{k-1}: frequent $(k-1)$-
$(k-1)$-subset s of c
L_{k-1} then

C_1

Itemset	Sup.
{I1}	2
{I2}	3
{I3}	3
{I4}	3
{I5}	1

Compare candid
support with
minimum suppo
count
⟶

C_2

Itemset
{I1,I2}

Scan D for
count of

C_2

Itemset	Sup.
{I1,I2}	2

Association Rule: Basic Concepts

• Given: (1) database of transactions, (2) each transaction is a list of items (purchased by a customer in a visit)

• Find: all rules that correlate the presence of one set of items with that of another set of items

– E.g., *98% of people who purchase tires and auto accessories also get automotive services done*

• Applications

– * □ *Maintenance Agreement* (What the store should do to boost Maintenance Agreement sales)

– *Home Electronics* □ * (What other products should the store stocks up?)

– Attached mailing in direct marketing
– Detecting “ping-pong”ing of patients, faulty “collisions”

Rule Measures: Support and Confidence

• Find all the rules *X & Y* □ *Z* with minimum confidence and support
– support, *s*, probability that a transaction contains {X □ Y □ Z}
– confidence, *c*, conditional probability that a transaction having {X □ Y} also contains *Z*

Let minimum support 50%, and minimum confidence 50%, we have
– *A* □ *C* (50%, 66.6%)
– *C* □ *A* (50%, 100%)

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

4:Classification through Decision Tree Algorithm:

There are constructs that are used by classifiers which are tools in data mining. These systems take inputs from a collection of cases where each case belongs to one of the small numbers of classes and are described by its values for a fixed set of attributes. The output classifier can accurately predict the class to which it belongs. It makes use of decision trees where the first initial tree is acquired by using a divide and conquer algorithm.

Suppose S is a class and the tree is leaf labeled with the most frequent class in S. Choosing a test based on single attribute with two or more outcomes than making this test as root one branch for each outcome of the test can be used. The partitions correspond to subsets S1, S2, etc. which are outcomes for each case. C4.5 allows for multiple outcomes. In the case of complex decision trees, C4.5 has introduced an alternative formula, which consists of a list of rules, where these rules are grouped together for each class. To classify the case the first class whose conditions are satisfied is named as the first one. If no rule is satisfied by the case, then it is assigned a default class. The C4.5 rulesets are formed from the initial decision tree. C4.5 enhances the scalability by multi-threading.

5: Why Data Preprocessing?

Data in the real world is dirty. It can be in incomplete, noisy and inconsistent from. These data needs to be preprocessed in order to help improve the quality of the data, and quality of the mining results.

❖If no quality data , then no quality mining results. The quality decision is always based on the quality data.

❖If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult

Incomplete data: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., occupation=“ ”.

Noisy data: containing errors or outliers data. e.g., Salary="-10"
Inconsistent data: containing discrepancies in codes or names. e.g., Age="42" Birthday="03/07/1997"
❖Incomplete data may come from
➢"Not applicable" data value when collected
➢Different considerations between the time when the data was collected and when it is analyzed.
➢Human/hardware/software problems ❖Noisy data (incorrect values) may come from
➢Faulty data collection by instruments ➢Human or computer error at data entry ➢Errors in data transmission
❖Inconsistent data may come from ➢Different data sources
➢Functional dependency violation (e.g., modify some linked data)

Major Tasks in Data Preprocessing

- ❖Data cleaning
  - ➢Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- ❖Data integration
  - ➢Integration of multiple databases, data cubes, or files
- ❖Data transformation
  - ➢Normalization and aggregation
- ❖Data reduction
  - ➢Obtains reduced representation in volume but produces the same or similar analytical results
- ❖Data discretization
  - ➢Part of data reduction but with particular importance, especially for numerical data

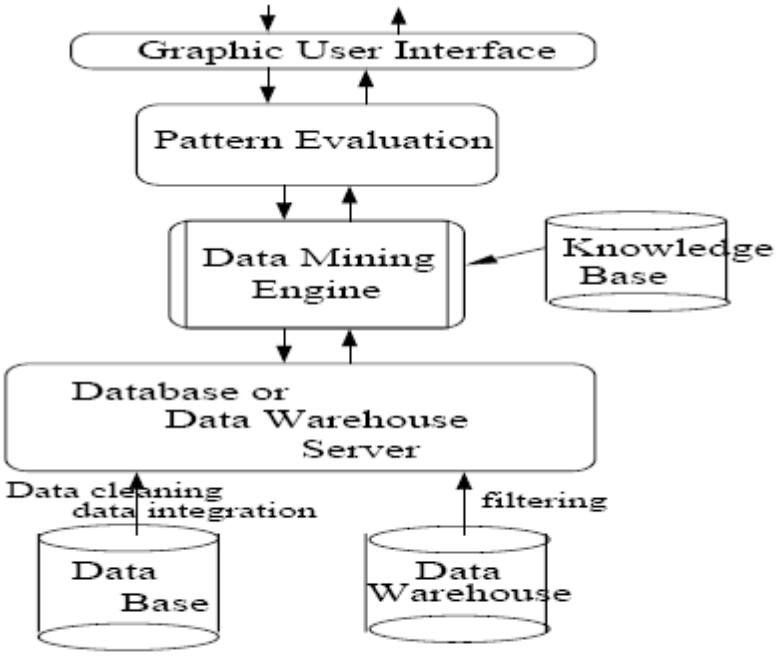
Forms of Data Preprocessing

6: Architecture of a typical data mining system/Major Components

Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Based on this view, the architecture of a typical data mining system may have the following major components:

1. A database, data warehouse, or other information repository, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.
2. A database or data warehouse server which fetches the relevant data based on users' data mining requests.
3. A knowledge base that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple heterogeneous sources.
4. A data mining engine, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.
5. A pattern evaluation module that works in tandem with the data mining modules by employing interestingness measures to help focus the search towards interestingness patterns.

6. A graphical user interface that allows the user an interactive approach to the data mining system.



Architecture of a typical data mining system.