**Ch. 6**

1. Frequent pattern: a that occurs frequently in a data set

2. Why Is Freq. Pattern Mining Important?

An intrinsic and important property of datasets.

Foundation for many essential data mining tasks:

Association, correlation, and causality analysis

Sequential, structural patterns

Classification

3. Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!

4. Further Improvement of the Apriori Method?

* Major computational challenges
  + Multiple scans of transaction database
  + Huge number of candidates
  + Tedious workload of support counting for candidates
* Improving Apriori: general ideas
  + Reduce passes of transaction database scans
  + Shrink number of candidates
  + Facilitate support counting of candidates

5. Benefits of the FP-tree Structure ?

* Completeness
  + Preserve complete information for frequent pattern mining
  + Never break a long pattern of any transaction
* Compactness
  + Reduce irrelevant info—infrequent items are gone
  + Items in frequency descending order: the more frequently occurring, the more likely to be shared Never be larger than the original database

6. What about if FP-tree cannot fit in memory?

* + DB projection

7. Parallel projection vs. partition projection techniques ?

* + Parallel projection
    - Project the DB in parallel for each frequent item
    - Parallel projection is space costly
    - All the partitions can be processed in parallel
  + Partition projection
    - Partition the DB based on the ordered frequent items
    - Passing the unprocessed parts to the subsequent partitions

8. Advantages of the Pattern Growth Approach ?

Divide-and-conquer

A good open-source implementation and refinement of FPGrowth

**Ch. 7**

9. Redundancy Filtering: Some rules may be redundant due to “ancestor” relationships between items

10. Categorical Attributes: finite number of possible values, no ordering among values—data cube approach

11. Quantitative Attributes: Numeric, implicit ordering among values—discretization, clustering, and gradient approaches

12. Techniques can be categorized by how numerical attributes?

* Static discretization based on predefined concept hierarchies
* Dynamic discretization based on data distribution
* Clustering: Distance-based association
* One dimensional clustering then association

13. Negative and Rare Patterns

* Rare patterns: Very low support but interesting

E.g., buying Rolex watches

* Negative patterns Since it is unlikely that one buys Ford Expedition

14 . Constraint-Based Frequent Pattern Mining

* Pattern space pruning constraints
  + Anti-monotonic: If constraint c is violated, its further mining can be terminated
  + Monotonic: If c is satisfied, no need to check c again
  + Succinct: c must be satisfied, so one can start with the data sets satisfying c
  + Convertible: c is not monotonic nor anti-monotonic, but it can be converted into it if items in the transaction can be properly ordered
* Data space pruning constraint
  + Data succinct: Data space can be pruned at the initial pattern mining process
  + Data anti-monotonic: If a transaction t does not satisfy c, t can be pruned from its further mining

Why Is Pattern-Fusion Efficient?

* A bounded-breadth pattern tree traversal
* Ability to identify “short-cuts” and take “leaps”
* Gearing toward colossal patterns
* Catching outliers

**Ch. 8**

15. **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

16. **Classification—A Two-Step Process ?**

* Model construction: describing a set of predetermined classes
  + Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  + The set of tuples used for model construction is training set
  + The model is represented as classification rules, decision trees, or mathematical formulae
* Model usage: for classifying future or unknown objects
  + Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set (otherwise overfitting)
  + If the accuracy is acceptable, use the model to classify new data
* Note: If *the test set* is used to select models, it is called validation (test) set

**17. Algorithm for Decision Tree Induction**

* **Basic algorithm (a greedy algorithm)**
* Tree is constructed in a top-down recursive divide-and-conquer manner
* At start, all the training examples are at the root
* Attributes are categorical (if continuous-valued, they are discretized in advance)
* Examples are partitioned recursively based on selected attributes
* Test attributes are selected on the basis of a heuristic or statistical measure
* **Conditions for stopping partitioning**
* All samples for a given node belong to the same class
* There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
* There are no samples left

**18. Comparing Attribute Selection Measures ?**

The three measures, in general, return good results but

* **Information gain:**
* biased towards multivalued attributes
* **Gain ratio:**
* tends to prefer unbalanced splits in which one partition is much smaller than the others
* **Gini index:**
* biased to multivalued attributes
* has difficulty when # of classes is large
* tends to favor tests that result in equal-sized partitions and purity in both partitions

**19. Other Attribute Selection Measures ?**

* **CHAID:** a popular decision tree algorithm, measure based on χ2 test for independence
* **C-SEP**: performs better than info. gain and gini index in certain cases
* **G-statistic**: has a close approximation to χ2 distribution
* **MDL** (Minimal Description Length) principle (i.e., the simplest solution is preferred):

20. **Overfitting and Tree Pruning** ?

**Overfitting:** An induced tree may over fit the training data

**Two approaches to avoid overfitting**

**Prepruning:** Halt tree construction early ̵ do not split a node if this would result in the goodness measure falling below a threshold

**Postpruning**: Remove branches from a “fully grown” tree—get a sequence of progressively pruned trees

21. **Enhancements to Basic Decision Tree Induction?**

Allow for continuous-valued attributes

Handle missing attribute values

Attribute construction

22. Why is decision tree induction popular?

* + relatively faster learning speed (than other classification methods)
  + convertible to simple and easy to understand classification rules
  + can use SQL queries for accessing databases
  + comparable classification accuracy with other methods

23. **Bayesian Classification: Why?**

**A statistical classifier**: performs probabilistic prediction, i.e., predicts class membership probabilities

**Foundation**: Based on Bayes’ Theorem.

**Performance:** A simple Bayesian classifier, *naïve Bayesian classifier*, has comparable performance with decision tree and selected neural network classifiers

**Incremental:** Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data

**Standard**: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

24. **Naïve Bayes Classifier ?**

**Advantages**

Easy to implement

Good results obtained in most of the cases

**Disadvantages**

Assumption: class conditional independence, therefore loss of accuracy

Practically, dependencies exist among variables

25. **How to Learn-One-Rule?**

Start with the most general rule possible..

Adding new attributes by adopting a greedy depth-first strategy..

Rule-Quality measures: consider both coverage and accuracy

Rule pruning based on an independent set of test tuples..

**26.** **How can we measure accuracy? Other metrics to consider?**

* Use **validation test set** of class-labeled tuples instead of training set when assessing accuracy
* Methods for estimating a classifier’s accuracy:

Holdout method, random subsampling

Cross-validation

Bootstrap

* Comparing classifiers:

Confidence intervals

Cost-benefit analysis and ROC Curves

**27. Evaluating Classifier Accuracy: Bootstrap ?**

**Bootstrap**

Works well with small data sets

Samples the given training tuples uniformly *with replacement*

**28. Issues Affecting Model Selection ?**

* **Accuracy**
  + classifier accuracy: predicting class label
* **Speed**
  + time to construct the model (training time)
  + time to use the model (classification/prediction time)
* **Robustness**: handling noise and missing values
* **Scalability**: efficiency in disk-resident databases
* **Interpretability**
  + understanding and insight provided by the model

**29**. Popular ensemble methods ?

* + Bagging: averaging the prediction over a collection of classifiers
  + Boosting: weighted vote with a collection of classifiers
  + Ensemble: combining a set of heterogeneous classifiers

**30. How boosting works?**

* Weights are assigned to each training tuple
* A series of k classifiers is iteratively learned
* After a classifier Mi is learned, the weights are updated to allow the subsequent classifier
* The final M\* combines the votes of each individual classifier, where the weight of each classifier's vote is a function of its accuracy

**31. Typical methods for imbalance data in 2-class classification?**

* + **Oversampling**: re-sampling of data from positive class
  + **Under-sampling**: randomly eliminate tuples from negative class
  + **Threshold-moving**: moves the decision threshold, t, so that thrare class tuples are easier to classify, and hence, less chance of costly false negative errors
  + **Ensemble techniques:** Ensemble multiple classifiers introduced above

**Ch. 9**

**32. How Are Bayesian Networks Constructed?**

* Subjective construction: Identification of (direct) causal structure
* Synthesis from other specifications
* Learning from data

**33. Classification by Backpropagation?**

**Backpropagation:** A neural network learning algorithm

Started by psychologists and neurobiologists to develop and test computational analogues of neurons

**A neural network:** A set of connected input/output units where each connection has a weight associated with it During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input

**34. Neural Network as a Classifier ?**

* Weakness
  + Long training time
  + Require a number of parameters typically best determined empirically, e.g., the network topology or “structure.”
  + Poor interpretability: Difficult to interpret the symbolic meaning behind the learned weights and of “hidden units” in the network
* Strength
  + High tolerance to noisy data
  + Ability to classify untrained patterns
  + Well-suited for continuous-valued inputs *and outputs*
  + Successful on an array of real-world data, e.g., hand-written letters
  + Algorithms are inherently parallel
  + Techniques have recently been developed for the extraction of rules from trained neural networks

1. **Discriminative Classifiers ?**

* **Advantages**
  + Prediction accuracy is generally high
  + Robust, works when training examples contain errors
  + Fast evaluation of the learned target function
* **Criticism**
  + Long training time
  + Difficult to understand the learned function (weights)
  + Not easy to incorporate domain knowledge

1. **SVM vs. Neural Network ?**

|  |  |
| --- | --- |
| **Support Vector Machine (SVM )** | **Neural Network** |
| Deterministic algorithm | Nondeterministic algorithm |
| Nice generalization properties | Generalizes well but doesn’t have strong mathematical foundation |
| Hard to learn – learned in batch mode using quadratic programming techniques | Can easily be learned in incremental fashion |
| Using kernels can learn very complex functions | To learn complex functions—use multilayer perceptron (nontrivial) |

1. **Why Associative Classification effective?**

* + It explores highly confident associations among multiple attributes and may overcome some constraints introduced by decision-tree induction
  + Associative classification has been found to be often more accurate than some traditional classification methods

1. **Lazy vs. Eager Learning** ?
   1. **Lazy learning** (e.g., instance-based learning): Simply stores training data (or only minor processing) and waits until it is given a test tuple
   2. **Eager learning** (the above discussed methods): Given a set of training tuples, constructs a classification model before receiving new (e.g., test) data to classify

**39. Genetic Algorithms (GA) ?**

**Genetic Algorithm:** based on an analogy to biological evolution

An initial population is created consisting of randomly generated rules

1. **Semi-Supervised Classification ?**

**Semi-supervised:** Uses labeled and unlabeled data to build a classifier .

**Ch 10**

**41.What is Cluster Analysis?**

**Cluster:** A collection of data objects similar (or related) to one another within the same group dissimilar (or unrelated) to the objects in other groups

**Cluster analysis** (or *clustering*, *data segmentation, …*) Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters .

**42.Basic Steps to Develop a Clustering Task ?**

* Feature selection
* Proximity measure
* Clustering criterion
* Clustering algorithms
* Validation of the results
* Interpretation of the results

**43.What Is Good Clustering ?**

* A good clustering method will produce high quality clusters
  + high intra-class similarity: cohesive within clusters
  + low inter-class similarity: distinctive between clusters
* The quality of a clustering method depends on
  + the similarity measure used by the method
  + its implementation, and
  + Its ability to discover some or all of the hidden patterns

**44.What are the Considerations for Cluster Analysis?**

* Partitioning criteria
* Separation of clusters
* Similarity measure
* Clustering space

**45.Major Clustering Approaches ?**

* Partitioning approach:
* Hierarchical approach:
* Density-based approach:
* Grid-based approach:
* Model-based:
* Frequent pattern-based
* User-guided or constraint-based:
* Link-based clustering

**46.What Is the Problem of the K-Means Method?**

* The k-means algorithm is sensitive to outliers
* K-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster

**47.Distance between Clusters ?**

* Single link: smallest distance between an element in one cluster and an element in the other
* Complete link: largest distance between an element in one cluster and an element in the other
* Average: avg distance between an element in one cluster and an element in the other
* Centroid: distance between the centroids of two clusters
* Medoid: distance between the medoids of two clusters

**48.Algorithmic and Probabilistic hierarchical clustering?**

* Algorithmic hierarchical clustering
  + Nontrivial to choose a good distance measure
  + Hard to handle missing attribute values
  + Optimization goal not clear: heuristic, local search
* Probabilistic hierarchical clustering
  + Use probabilistic models to measure distances between clusters
  + Generative model: Regard the set of data objects to be clustered as a sample of the underlying data generation mechanism to be analyzed
  + Easy to understand, same efficiency as algorithmic agglomerative clustering method, can handle partially observed data

**49.** **Two parameters for the Density-Based Clustering ?**

* ***Eps*:** Maximum radius of the neighbourhood
  + ***MinPts*:** Minimum number of points in an Eps-neighbourhood of that point

**50.Major features of Density-Based Clustering ?**

* + Discover clusters of arbitrary shape
  + Handle noise
  + One scan
  + Need density parameters as termination condition

51. **Strength and Weakness of *CLIQUE*** ?

* **Strength** 
  + automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces
  + insensitive to the order of records in input and does not presume some canonical data distribution
  + scales linearly with the size of input and has good scalability as the number of dimensions in the data increases
* **Weakness**
  + The accuracy of the clustering result may be degraded at the expense of simplicity of the method

52. **Measuring Clustering Quality?**

3 kinds of measures: External, internal and relative

**External:** supervised, employ criteria not inherent to the dataset

**Internal:** unsupervised, criteria derived from data itself

**Relative:** directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

**Ch 11**

**53. The k-means algorithm has two steps at each iteration?**

* + **Expectation Step** (E-step): Given the current cluster centers, each object is assigned to the cluster whose center is closest to the object: An object is *expected to belong to the closest cluster*
  + **Maximization Step** (M-step): Given the cluster assignment, for each cluster, the algorithm *adjusts the center* so that *the sum of distance* from the objects assigned to this cluster and the new center is minimized

**54.** **The EM (Expectation Maximization) Algorithm?**

**The (EM) algorithm:** A framework to approach maximum likelihood or maximum a posteriori estimates of parameters in statistical models.

* + **E-step** assigns objects to clusters according to the current fuzzy clustering or parameters of probabilistic clusters
  + **M-step** finds the new clustering or parameters that maximize the sum of squared error (SSE) or the expected likelihood

**55.** **Advantages and Disadvantages of Mixture Models ?**

* Strength
  + Mixture models are more general than partitioning and fuzzy clustering
  + Clusters can be characterized by a small number of parameters
  + The results may satisfy the statistical assumptions of the generative models
* Weakness
  + Converge to local optimal (overcome: run multi-times w. random initialization)
  + Computationally expensive if the number of distributions is large, or the data set contains very few observed data points
  + Need large data sets
  + Hard to estimate the number of clusters

**56.** **Methods of Clustering High-Dimensional Data ?**

**Subspace-clustering:** Search for clusters existing in subspaces of the given high dimensional data space

**Dimensionality reduction approaches:** Construct a much lower dimensional space and search for clusters there.

**57. Graph Clustering: Challenges of Finding Good Cuts ?**

* High computational cost
* Sophisticated graphs
* High dimensionality
* Sparsity

**58. Two Approaches for Graph Clustering ?**

* + Use generic clustering methods for high-dimensional data
  + Designed specifically for clustering graphs

**Ch 12**

**59.** **What Are Outliers and its types?**

**Outlier:** A data object that deviates significantly from the normal objects as if it were generated by a different mechanism

Global outlier (or point anomaly)

*Contextual outlier* (or *conditional outlier*)

*Collective Outliers*

**60. Challenges of Outlier Detection ?**

Modeling normal objects and outliers properly..

Application-specific outlier detection..

Handling noise in outlier detection..

Understandability..

**61.** **Two ways to categorize outlier detection methods?**

a) Based on whether user-*labeled* examples of outliers can be obtained..

b) Based on *assumptions about normal data and outlier*

**62.** **Clustering-Based Method: Strength and Weakness?**

* Strength
  + Detect outliers without requiring any labeled data
  + Work for many types of data
  + Clusters can be regarded as summaries of the data
  + Once the cluster are obtained, need only compare any object against the clusters to determine whether it is an outlier (fast)
* Weakness
  + Effectiveness depends highly on the clustering method used—they may not be optimized for outlier detection
  + High computational cost: Need to first find clusters
  + A method to reduce the cost: Fixed-width clustering

**Ch 13**

**63.** **Statistical data mining methods ?**

**Regression**: predict the value of a response (dependent) variable from one or more predictor (independent) variables where the variables are numeric

**Generalized Linear Model:** allow a categorical response variable (or some transformation of it) to be related to a set of predictor variables

**Analysis of Variance:** Analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables

**Mixed-Effect Models:** For analyzing grouped data, i.e. data that can be classified according to one or more grouping variables

**Factor Analysis**: determine which variables are combined to generate a given factor

**Discriminant Analysis**: predict a categorical response variable, commonly used in social science

**Survival Analysis:** Predicts the probability that a patient undergoing a medical treatment would survive at least to time

**64.** **Integration of visualization and data mining ?**

* + data visualization
  + data mining result visualization
  + data mining process visualization
  + interactive visual data mining

**تمنياتي للجميع بالسداد والتوفيق**