AFTER M.S.E.

★ Multilevel Association Rules

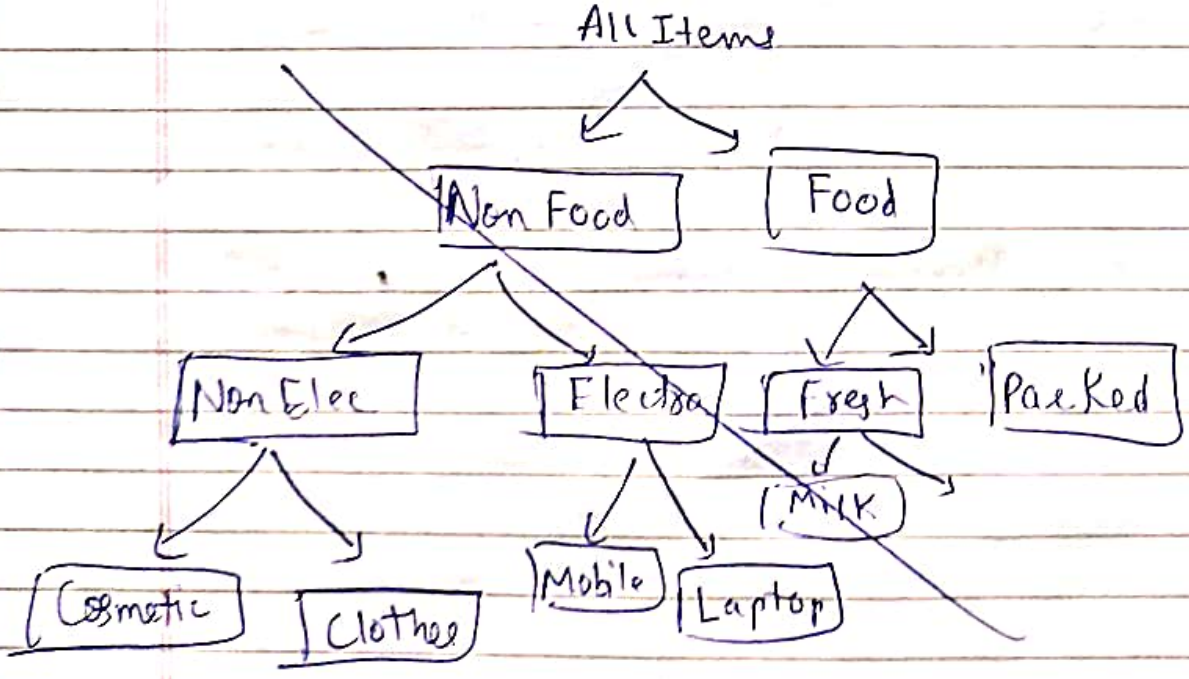
So when transaction data is taken for link analysis. It is present at the low level of abstraction that is detail form. It is difficult to form Rules at the low level of abstraction as data scarcity is there.

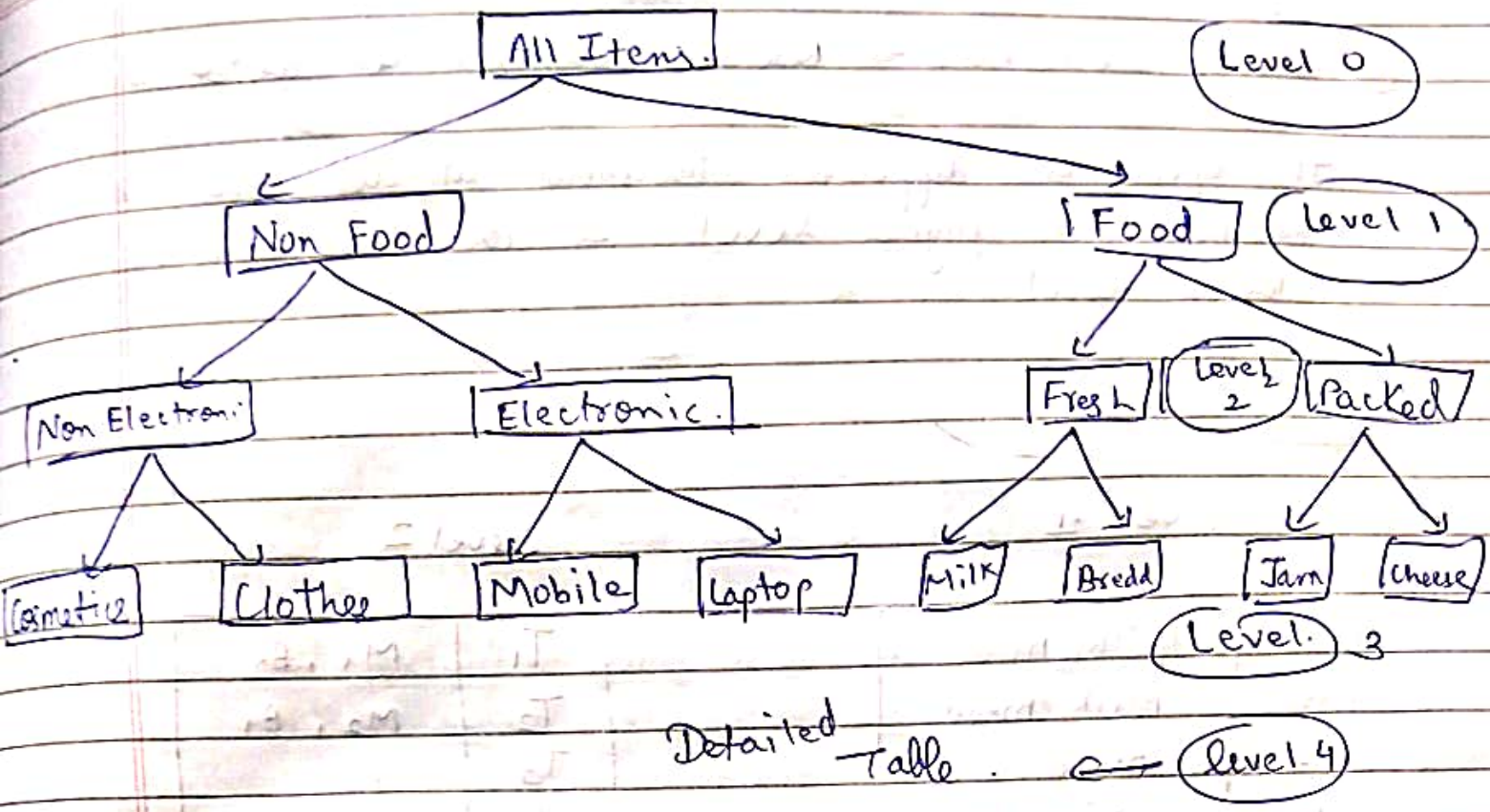
∴ Using concept hierarchies, transaction data can be generated at various levels of abstraction.

In Multilevel Association Rules, association rules are generated at Multiple levels of abstraction. Instead of going at lowest level of abstraction, association rules are generated from higher level of Abstraction which represents common sense knowledge & be used efficiently.

abstraction.

Eg. → low level, transactions chya basis var association rules banavane is diff mahnun we go for all levels (Multi) level).





: level 0 → high level of abstraction
level 4 → low level of abstraction

The main aim of Multi-level ASRM is to find hidden information between levels of abstraction.

Three types.

① Uniform Support → same support at all levels. Simple to implement.

Draw back.

① If the min. support threshold is set too high, it could miss some meaningful associations occuring at low abstr. levels
② If the threshold is set too low, it may generate uninteresting associations occuring at high abstraction levels

(2) Reduced threshold → Reduced support at lower levels.

It specifies different threshold at diff level. High threshold at higher level & low threshold at lower level.

Ex

Level 1

T_1	Milk, Bred
T_2	Milk, Bread
T_3	Bread
T_4	Milk, Bread
T_5	Milk

Level 2.

T_1	M_1, B_2
T_2	M_2, B_1
T_3	B_2
T_4	M_2, B_1
T_5	M_2

So for level 1 Min Support is 50%. i.e 3
If we have same for level 2 we will not get anything hence reduced support 40%. i-e 2 so that we get frequent pattern.

(3) Group based support

Here we set support threshold based on input by user as to which group is imp.
For that we set low support so that we can get maximum buying patterns of these items.

* Multidimensional Association Rule.

① Single dimensional Association Rule.

↳ Only 1 predicate with Multiple occurence.

buys (x, "digital camera") ⇒ buys (x, "Printer"),

Also referred as intradimensional association rule.

② Multidimensional. → Association rules that involve two or more predicates ~~with no repetations but multiple~~ are called called Multidimensional association rules. Those with no repeated predicates are called interdim. association rules.

If Multidimensional Association rules have repeated predicates then they are called hybrid-dimensional association rules.

Multidimensional

→ age (X, "20..29") ∧ occupa' (x, "stu") ⇒ buys (x, "laptop")

Hybrid

→ age (x, "20.... 29") ∧ buys (x, "laptop") ⇒ buys (x, "HP")

From book

* Single-dimensional Association Rule ⇒ These are Intra-dimension association rules. Association rule that imply a single predicate are called Single-dimensional Association Rules.

Multidimensional Association Rules.

These are Inter-dimension Association rules. Association rules that imply two or more dimensions ~~are~~/predicates are called Multidimension Association Rules.

Inter-dimension → no repeated predicates

Hybrid-dimension → repeated predicates.

* Datal aee attributes can be categorical or quantitative.

① Categorical → Attributes have a finite number of possible values & no ordering among values are called Categorical Attributes. (name of things)

② Quantitative Attributes → Attributes are numeric & have an implicit ordering among values are called Quantitative Attributes. (Numeric)

* Techniques for mining Multidimensional Association Rules can be categorized according to three basic approaches regarding the treatment of quantitative attributes.

① In the first approach, quantitative attributes are discretized using predefined concept hierarchies. The discretization occures prior to mining. The discretization is static & predetermined hence called as static discretization of quantitative attributes.

(2) In the second approach, quantitative attributes are discretized into "bins" based on the distribution of the data.
This discretization is dynamic & established so as to satisfy some Mining criteria.

(3) In the third approach, quantitative attributes are discretized so as to capture the semantic meaning of such interval data. This dynamic discretization procedure considers the distance betⁿ data points.
Hence such quantitative association rules are referred as distance based association rules

* Association Rule Clustering System (ARCS)
ARCS is a method to mine quantitative association rules have two quantitative attributes on the left hand side of the rule & one categorical attribute on the rule.

$$A_{q_1} \wedge A_{q_2} \rightarrow A_{cat}$$

~~ARCS works as by~~
The following steps are involved in ARCS

① ~~Bins~~ Binning:
Quantitative attributes can have a very wide range of values defining their domain. To keep grids down to a manageable size, partition the ranges of quantitative attributes into intervals. This partitioning is referred as binning.

Three common binning strategies are
① Equiwidth → Interval size of each bin is same
② Equidepth → where each bin has exactly same number of tuples assigned
③ Homogeneity based → where bin size is ~~defined~~ determined so that the tuples in each bin are uniformly distrib.

② Then from the 2-D array which has the count distribution for each category we can find frequent predicate sets ~~with~~ that satisfy minimum confidence.
Then using these frequent predicates generate strong Association Rules.

③ Clustering → The strong association rules obtained in the previous step are then mapped to a 2D grid.
For ex.

age (X, 34) ∧ income (X, "31K ... 40K") → buys (X, "High resolution TV")

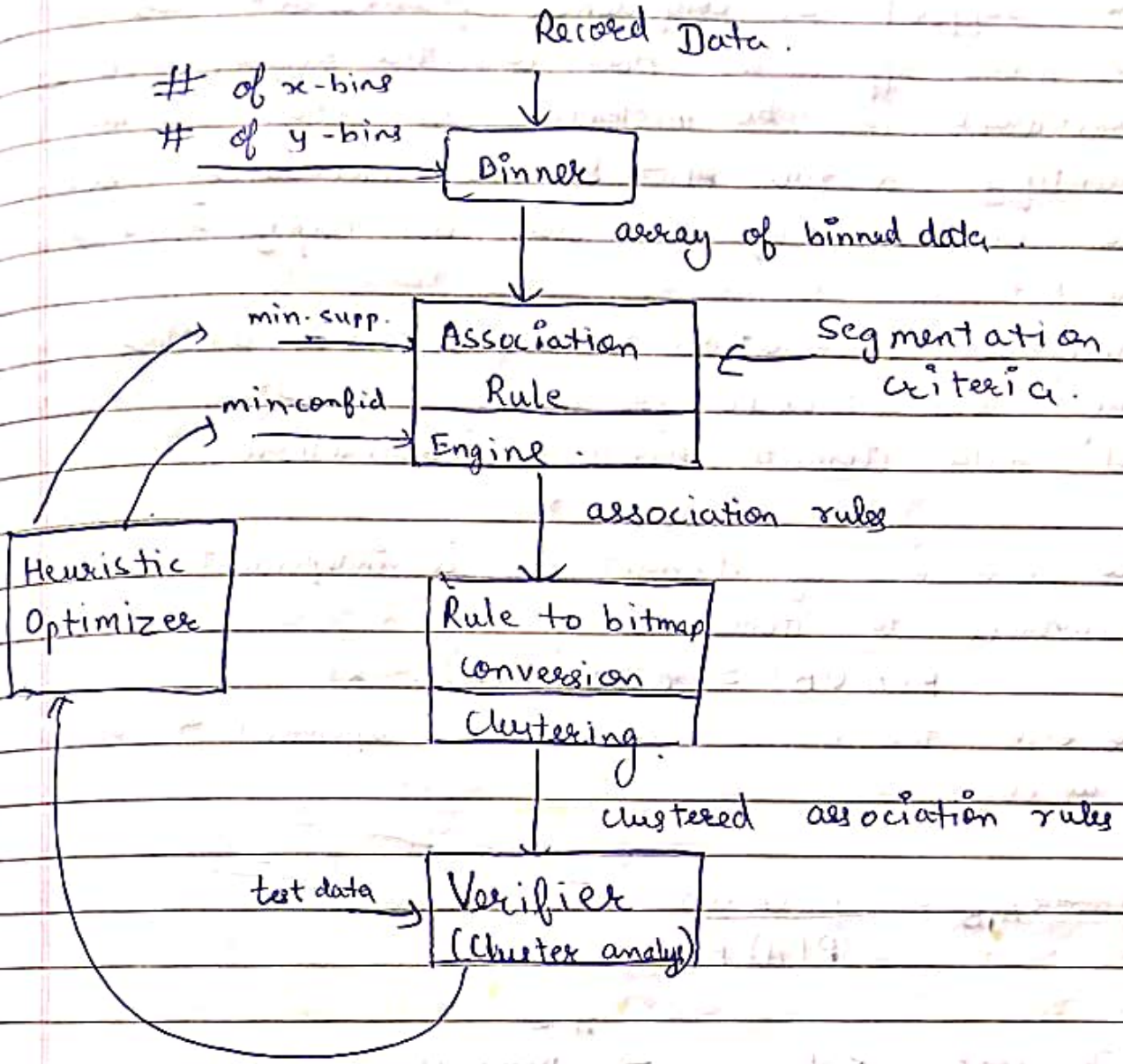
age (X, 35) ∧ income (X, "31K ... 40K") → buys (X, "High resolution TV")

age (X, 34) ∧ income (X, "41K ... 50K") → buys (X, "High resolution TV"),

age (X, 35) ∧ income (X, "41K ... 50K") → buys (X, "high resolution TV"),

So ~~a~~ after clustering all we can form a simple rule.

age (X, "34 ... 35") ∧ income (X, "31K ... 50K") → buys (X, "high resolution TV").



* Limitations of ARCS
  - Only quantitative attributes on LHS of rules.
  - Only 2 attributes on LHS (2D Limitation)

* The support & confidence framework are useful for many applications However the support - confidence framework can be misleading in that it may identify a rule $A \rightarrow B$ as interesting, even the occurence of A does not imply the occurence of B.

(Correlation Analysis)

Therefore Correlation Analysis was another framework which was introduced for finding interesting relationships betⁿ data itemsets based on correlation.

* The occurence of itemset A is independent of the occurence of itemset B if

$$P(A \cup B) = P(A) \cdot P(B).$$

Or else the itemsets A & B are dependent & correlated as events.

$$corr_{A,B} = \frac{P(A \cup B)}{P(A)\,P(B)}$$

→ More related to Probability

If $corr_{A,B} < 1$ → A is negatively correlated with occurence of B.

$corr_{A,B} > 1$ → " —" +vely correlated.

$= 1$ → Independent.

* Constraint Based Association Mining.

For a given set of task relevant data, the data mining process may uncover thousands of rules many of which are uninteresting to the user. In constraint based

mining. mining is performed, under the guidance of various kinds of constraints provided by the user.

These constraints include the following

1. Knowledge type constraints
2. Data constraints
3. Dimension / level constraints.
4. Interestingness constraints
5. Rule constraints.

* Regression Analysis.

$$\frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

* Correlation analysis → The primary goal of correlation is ~~to identify~~ to understand if changes in one variables are associated with changes in another.
Correlation is measured by the correlation coefficient.

~~Types of Correlation Co~~

① Pearson's Correlation Coefficient (r) → Measures the linear relationship bet^n two continuous variables.

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{(n\Sigma x^2 - (\Sigma x)^2)\,(n\Sigma y^2 - (\Sigma y)^2)}}$$

* Regression Analysis.

Regression Analysis builds on correlation by not only assessing the strength of the relationship but also modeling it to make predictions.
In regression we try to find an equation that describes how Y changes as X changes.

$$y = a + bx$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a = \frac{\Sigma y - b\Sigma x}{n}$$

* Correlation quantifies the strength & direction of a relationship
* Regression models the relationship to predict values.

★ Classification & Prediction

Data classification is a 2 step process.
There are two sets of data ① training set ② test set.

In the first step a model is built ~~using~~ ~~dis~~ describing a predetermined set of data classes or concepts.
That means it is used for deriving classifier.
The second step is used to measure the accuracy of the classifier. The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified.

★ Applications of Classification & prediction

① Credit Approval, medical diagnosis, perfomance prediction
② Classifying credit card transaction as legitimate or fraudulent
③ Classifying secondary structures of protein as alpha-helix, beta-sheet or random coil.
④ Categorizing news stories as finance, weather, sports & ~~et~~ entertainments.

★ Classification & prediction methods can be compared & evaluated according to the foll criteria.

① Predictive accuracy → This refers to the ability of the
② ~~Speed~~ model to correctly predict the class label of
③ new or previously unseen data

② Speed → This refers to the computation cost involved in generating & using the model.

③ Robustness → This is the ability of the model to make correct predictions given in noisy data or data with values.

④ Scalability → This refers to the ability to construct the model efficiently given in large amounts of data.

⑤ Interpretability → This refers to the level of understanding & insight that is provided by the model

★ We categorize the attributes of the records into two diff types. Attributes whose domain is numerical are called numerical Attribute, & the attributes whose domain is not numerical are called categorical Attribute.

The goal of classification is to build a concise model that can be used to predict the class of the records whose class label is not known.

★ Advantages & Shortcomings of Decision Tree Classifications.

The major strengths of the decision tree Methods are the foll.

① Decision trees are able to generate understandable rules.

② They provide a clear indication of which fields are most important for prediction or classification.

③ They are able to handle both numerical & categorical attributes.

Disadvantages

① Only some decision trees can deal with binary-valued target classes.

② The process of ~~gen~~ growing a decision tree is computationaly expensive.

★ Tree Construction Principle.

The qualifying condition on the splitting attribute for data set splitting at a node is called the splitting criterion at that node.

The construction of the decision tree involves the foll. three main phases.

① Construction phase → The initial decision ~~de~~ tree is constructed in this phase, based on the entire training data set. It requires recursively partitioning the training set into two or more sub partition using a splitting criteria until a stopping criteria is met.

② Pruning phase → The tree constructed in the previous phase may not result in the best possible set of rules due to over-fitting. The pruning phase removes some of the lower branches & nodes to improve its performance.

③ ~~Optim~~ Optimization phase.

→ Processing the pruned tree to improve understandability.

* The Generic Algorithm:

Let the training data set with be T with class labels $\{C_1, C_2 \ldots C_K\}$.

- T is Homogeneous: T contains cases all belonging to a single class C. The decision tree for T is a ~~class~~ leaf identifying class C.
- T is not Homogeneous: T contains cases that belong to mixture of classes. The decision tree for T consists of a decision node identifying the test & one branch for each possible outcome.
- T is trivial: T contains no cases. The decision tree T is a leaf, but the classes to be associated with the leaf must be determined from information other than T.

* The generic algorithm of decision tree construction outl

* Guillotine Cut → Normally the splitting is done for a single attribute at any stage & if the attribute is numeric then the splitting test is an inequality. Geometrically each splitting can be viewed as a plane parallel to one of the axes. We call this the Guillotine cut phenomenon.

* Overfit → A decision tree T is said to overfit the training data if there exists some other tree T' which is a simplification of T, such that T has smaller error than T' over the training set T' but T' has smaller error than T over

the entire distribution of instances.

* Attribute Selection Error : Most of the D-T algorithms exhibit a systematic, unwanted preference for certain types of variables. Some decision tree algorithms are far more likely to construct models that use discrete variables with many values, than discrete variables with relatively few values.

Disadvantages of an overfitted decision tree.

- ✓ Overfitted models are incorrect
- ✓ Overfitted decision trees require more space & more computational resources
- ✓ Overfitted model require the collection of unnecessary features
- ✓ They are more difficult to comprehend.

* The best split is defined as one that does the best job of separating the records into groups where a single class predominates.

* To select best split we have two Indices
  1. one index is based on the information theory. That is information gain based entropy.
  2. The other one is derived from economics as measure of diversity This is called the gini Index.

* The information gain represents the diff bet^n the information needed to identify an element of T & the information needed to identify an element of T after the value of attribute X is obtained.

$$Info(X,T) = \sum_{i=1}^{n} \left|\frac{T_i}{T}\right| Info(T_i)$$

$$Gain(X,T) = Info(T) - Info(X,T)$$

$$Gain\ Ratio(X,T) = \frac{Gain(X,T)}{Info(X,T)}$$

* Binary splits for Numerical Attributes.

	C_1	C_2
L	a_1	a_2
R	b_1	b_2

← Class Histogram.

$$Gini = \frac{(a_1+a_2)}{n}\left[1 - \left(\frac{a_1}{a_1+a_2}\right)^2 - \left(\frac{a_2}{a_1+a_2}\right)^2\right] + \frac{b_1+b_2}{n}\left[1 - \left(\frac{b_1}{b_1+b_2}\right)^2 - \left(\frac{b_2}{b_1+b_2}\right)^2\right]$$

* Binary splits for Categorical Attributes

Here we have to draw a count Matrix.

	Class	
Family	High	1
Family	Low	4
Sports	High	2
Truck	Low	4
Sports	High	3
Family	High	6

→

	H	L
F	2	1
S	2	0
T	0	1

Count Matrix

Attribute list for a categorical Attribute

$$\text{Gini S(T)} = \frac{4}{6}\left[1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right] + \frac{2}{6}\left[1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2\right]$$

$$\Rightarrow \frac{1}{3}$$

* Bayesian Classifiers.

X → attribute, Y → class variable

If there is a non deterministic relationship ~~with the~~ in betⁿ X & Y then we can treat X & Y as random variables & capture their relationship probabilistically using P(Y/X).

$$P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)}$$

$P(y)$ → Prior Probability

$P(y/x)$ → Posterior Probability.

☆ Naive Bayes Classifier.

A Naive Bayes Classifier estimates the class-cond-probability by assuming that the attributes are conditionally independent, given the class label y.

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y)$$

d → No of attributes in X.

For Categorical → Use the regular method.

For Numerical → Use the below given Formula.

$$P(A,C) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{\frac{(A-\mu)^2}{2\sigma^2}}$$

Characteristics of Naive Bayes Classifiers.

① There are ~~robos~~ robust to isolated noise points because such pts are averaged out when estimating conditional probabilites from data.

② There are robust to irrelevant attributes. If X is an irrelevant attributes, then P(X|Y) becomes almost uniformly distributed.

③ Correlated attributes can degrade the performance of naive Bayes classifiers because the conditional independence assumptions no longer holds for such attributes.

Bayesian Belief Networks. (BBN)

BBN or simply Bayesian Network provides a graphical representation of the probabilistic relationship among a set of random variables.

There are two key elements of a Bayesian Network.

① A directed acyclic graph (dag) encoding the dependence relationships among a set of variables.

② A probability table associating each node to its immediate parent nodes

Characteristics of BBN.

① BBN provides an approach for capturing the prior knowledge of a particular domain using a graphical ~~Method~~ Model.

② The Network can also be used to encode causal dependencies among variables.

③ Constructing the Network can be time consuming & requires a large amt of effort. However once the structure is complete adding a new variable is straightforward.

④ Bayesian networks are well suited to dealing with incomplete data.

⑤ ~~Beca~~ Because the data is combined probabilistically with prior knowledge, the method is quite robust to model overfitting

✱ Other Classification Methods.

① K-Nearest Neighbor Classifiers

Here a point is in n-dimensional plane. When given an unknown sample, a K-nearest neighbor classifier searches the pattern space ~~of~~ for the K-training samples that are closest to it. The one with least Euclidean distance is assigned to the ~~most common class amon~~ unknown sample.

② Case-Based Reasoning

③ Genetic Algorithms.

④ Rough sets

⑤ Fuzzy sets.

Prediction

Prediction is similar to classification. It constructs a model. This model is used to predict unknown & missing values.

① Most popular approach for prediction is

* Regression.

$$y = b + wX \quad \rightarrow \text{Covered earlier.}$$

* Classifier Accuracy.

Estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classifier will label a future data that is data on which the classifier has not been trained.

Methods.

① Holdout Method → Here the original data with labeled examples is partitioned into ~~to~~ training & test sets.

② Random Subsampling → The holdout method can be repeated several times to improve the estimation of a classifer's perfomance.

③ Cross - Validation → An alternative to Random subsampling is cross - validation. Here each record is used ~~to~~ the same number of times for training & exactly once for testing.

Bootstrap → The training records are sampled with~~out~~ replacement i.e a record already chosen for training is put back into the original pool of records so that it is equally likely to be redrawn.

Clustering

* What Is Cluster Analysis ?

The process of grouping a set of physical or abstract objects into classes of similar to one another within the same cluster & are dissimilar to the objects in other clusters.

* Requirements of clustering in data mining.

(1) Scalability → Highly scalable clustering algorithm is needed.

(2) Ability to deal with diff types of attributes.

(3) An algorithm that can detect ~~arbitary~~ clusters of arbitary shape.

(4) Minimum requirement for domain knowledge to determine input parameters.

(5) Ability to deal with Noisy data.

(6) Incremental clustering & insensitivity to order of input records.

(7) High dimensionality

(8) Constraint - based Clustering

(9) Interpretability & usability

→ Major - Methods used in Clustering.

i) Partitioning methods: The partition method constructs K - partitions of data, where each partition represents a cluster & $K <= n$ i.e it classifies the data into K - groups, which together satisfy the foll. requirements.

a) Each group must contain at least one object.
b) Each object must belong to exactly 1 group.

A) K means Algorithm. → Here each cluster is represented by mean value of the objects in the cluster.
B) K - medoids, where each cluster is represented by one of the obj. located near the center of the cluster

ii) Hierarchical Methods: - This method creates a Hierarchical decomposition of the given set of data objects.
It can be classified as

Agglomerative (bottom Up)	- divisive (Top - down)
Starts with all pts as clusters & then goes on merging	Here all pts are considered in one cluster & the odd man out is removed in succession.

Hierarchical methods suffer from the fact that once a step is done it can never be undone.

iii) Density - based Methods. →
Their general idea is to continue growing a given cluster as long as the density in the neighborhood exceeds some threshold.

DBSCAN & optics are typical density-based methods.

iv) Grid-based Methods →
Grid based methods quantize the object space into a finite number of cells that form a grid structure.

v) Model - based Methods → Model - based methods hypothesize a model for each of the cluster & find the best-fit of the data to the given model.

vi) Constraint based Clustering → Here clustering is done on the basis of user specified constraints.

* Algorithm : Kmeans. → The Kmeans algorithm for ~~cluster~~ partioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input :
- K: the number of clusters
- D: a data set containg n objects.

Output : A set of K clusters.

Method.

1) Arbitrary choose K objects from D as the initial cluster centers.
2) repeat.
3) reassign each object to the cluster to which the object is the most similar., based on the mean value of the objects in the cluster.
4) Update the cluster means, that is, calculate the mean value of the objects for each cluster;
5) until no change;

Disadvantages of K means.

① Not applicable for categorical data.
② The necessity for users to specify K, the no. of clusters, in advance.
③ Not suitable for clusters with non convex shapes
④ Sensitive to noise & outlier data pts.

Advantages

① scalable & efficient in processing large data sets.
② Method often terminates at local optimum.

Hierarchical Method.

* A hierarchical clustering method works by grouping data objects into a hierarchy or tree of clusters.

* Agglomerative hierarchical clustering → This bottom-up strategy starts by placing each object in its own cluster & then merges these atomic clusters into larger & larger clusters until all of the objects are in a single cluster. ~~or until~~.

* Divisive hierarchical clustering → This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the cluster into smaller & smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination condition.

A tree structure called a dendogram is commonly used to represent the process of hierarchical clustering. It shows how objects are grouped together step by step.

~~Single Lin~~ ∞

Single linkage → Mini.
Complete → Max
Avg → Avg.

The use of mean or avg distance is a compromise bet^n the min. & max. distances to ~~be~~ overcome the outlier sensitivity problem.

Difficulties with hierarchical Clustering.

There is difficulty regarding the selection of merge or split points. Merge-split decisions, if not well chosen at some step may lead to low-quality clusters.

Method does not Scale well.

* Density-Based Methods.

The Algorithm grows regions with sufficiently high density into clusters & discovers clusters of arbitary shape in spatial database with noise.

Here a cluster is defined as a maximal set of density-connected points.

The neighborhood within a radius ε of a given object is called the ε neighbourhood of the object.

If the ε-neighborhood, ~~within a radius ε of~~ of an obj contains at least a minimum number of MinPts of Objects then the object is called a core object.

Minpts → specifies the density threshold of dense region.

A given a set of objects D we say that an Object P is directly density-reachable from q if p is within the ε-neighborhood of q & q is core objects.

* An object p is density-reachable from object q with respect to ε & MinPts in a set of objects D, If there is a chain of objects $p_1 \dots p_n$ where $p_1 = 1$ & $p_n = p$ such that p_{i+1} is directly density reachable from p_i with respect to ε & MinPts

* Two objects $p_1, p_2 \in D$ are density-connected with respect to ε & MinPts if there is an object $q \in D$ such that both p_1 and p_2 are density reachable from q with respect to ε & MinPts.

Algorithm for the DBSCAN is
[Density Based spatial Clustering of Application with Noise]

- Arbitrarily select a point q.
- Retrieve all points density-reachable from q wrt Eps & MinPts
- If q is a core point, a cluster is formed
- If q is border point no points are density-reachable from q & DBSCAN visits the next point of database
- Continue the process until all of the points have been processed.

★ Grid - Based Methods.

The grid- based clustering approach uses a ~~Multires~~ Multiresolution grid data structure. It quantizes the object space into a finite number of cell that form a grid structure on which all of the operations for ~~co~~ clustering are performed.

The main advantage of this approach is its fast processing time.

STING

Ex → STING (statistical information grid) which ~~clusters~~ explores statistical information stored in the grid cells.

Wave Cluster which clusters objects using a wavelet transform method.

Clique → which represents a grid & density based approach for clustering ~~in~~ in a high dimensional data space.

★ Model - Based Clustering Methods.

Model Based clustering Methods attempt to optimise the fit bet^n the given data & some mathematical model.

Examples → ① Expectation - Maximization

② Conceptual Clustering

③ Neural Network Approach.

Clustering High-Dimensional Data.

Most clustering methods are designed for clustering low-dimensional data & encounter challenges when the dimensionality of the data grows really high. This is because when the dimensionality increases usually only a small number of dimensions are relevant to certain clusters, but in the irrelevant dimensions may produce noise & mask the real clusters to be discovered.

① Self organizing feature maps (SOMs)
Clustering is also performed by having several units competing for the current object. The unit whose weight vector is closest to the current objects wins.
① Feature transformation techniques.
② Feature selection techniques.

* Frequent Pattern Based Clustering Methods.
As the name implies, searches for patterns that occur frequently in large data sets.
Frequent Pattern Based cluster analysis is well suited to high-dimensional data.

* Constraint-Based Cluster Analysis.

Constraint-based clustering finds clusters that satisfy user-specified preferences constraints.

The constraints are as follows.

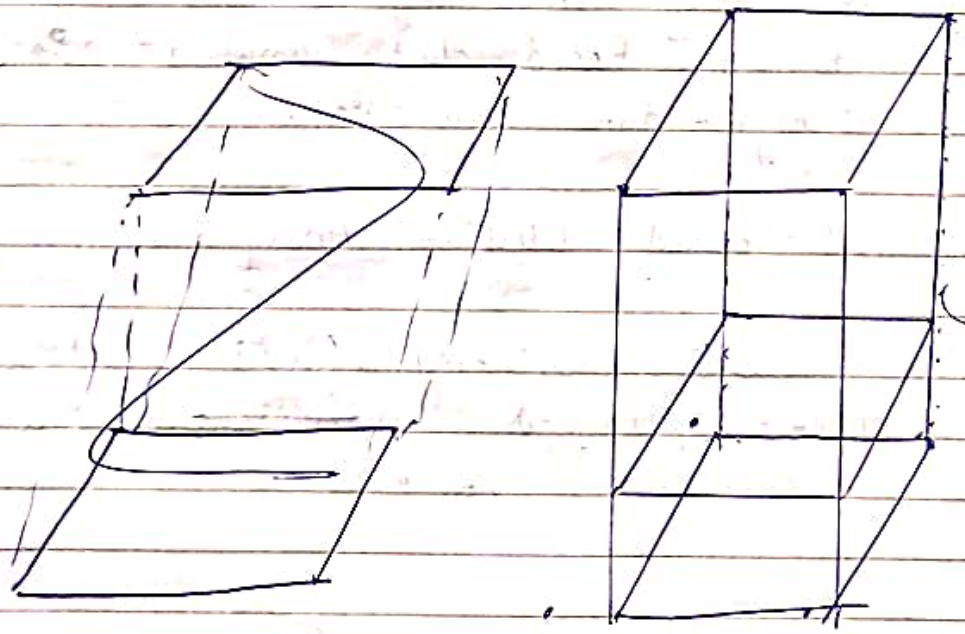
1. Constraints on individual objects
2. Constraints on selection of clustering ~~patterns~~ parameters
3. Constraints on distance of similarity function.
4. User specified constraints on the properties of individual clusters
5. Semi - Supervised clustering based on partial supervision

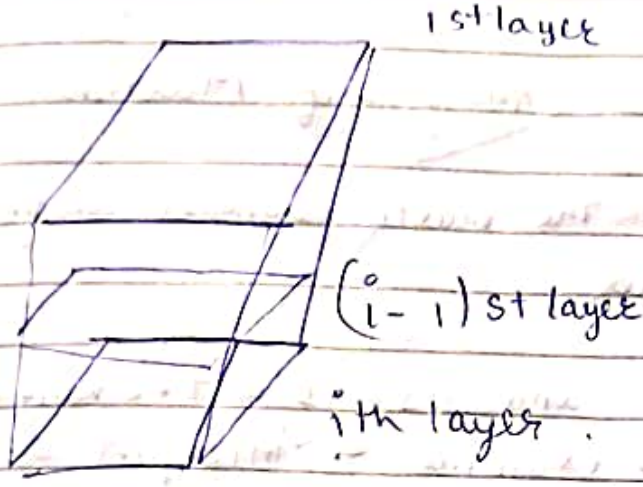
★ Grid Based Methods

STING is a grid-based multiresolution clustering method in which the spatial area is divided into rectangular cells. There are generally several levels of such rectangular cells corresponding to multiple levels of resolution.

Statistical parameters of higher-level cells can simply be calculated from the parameters of the lower-level cells.

15





A hierarchical structure for Sting clustering.

★ Outlier Analysis

There exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data are called outliers.

There are two basic types of procedures for detecting outliers.

① Block procedures. → In this either all the suspect objects are treated as outliers or all of them are accepted as consistent.

② Consecutive procedures → The main idea here is the object least likely to be outlier is tested first. If it is found to be outlier, then all of the more extreme values are also considered outliers otherwise the next most extreme object is tested.
This is more effective than block procedures.

Mining Complex types of Data.

* Multidimensional Analysis of Multimedia Data

To facilitate the multidimensional analysis of large multimedia data base.

Multidimensional analysis is a technique used in data analysis to examine & interpret data from multiple perspective or dimensions simultaneously

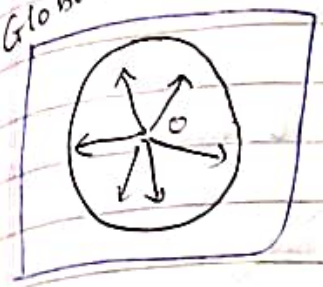
A multimedia data cube can have many dimension. Some examples are: The size of the image or video in bytes, the width & height of the frames, constituting two dimensions, the date on which the image or video was created, the format type of the image or video, the frame sequence duration in seconds,
The multimedia data cube seems to be an interesting model for multidimensional analysis of ~~data~~ multimedia data.

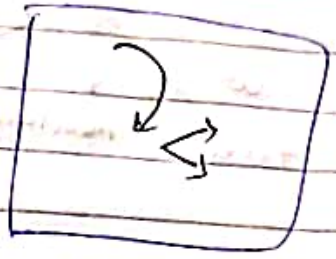
* Spatial Trend Detection.

Spatial Trend detection is as a regular change of one or more non-spatial attributes when moving away from a given start object 'o'.

There are 2 diff types of Algorithms.
1. Global - trend.
2. local trend.

Global trend.



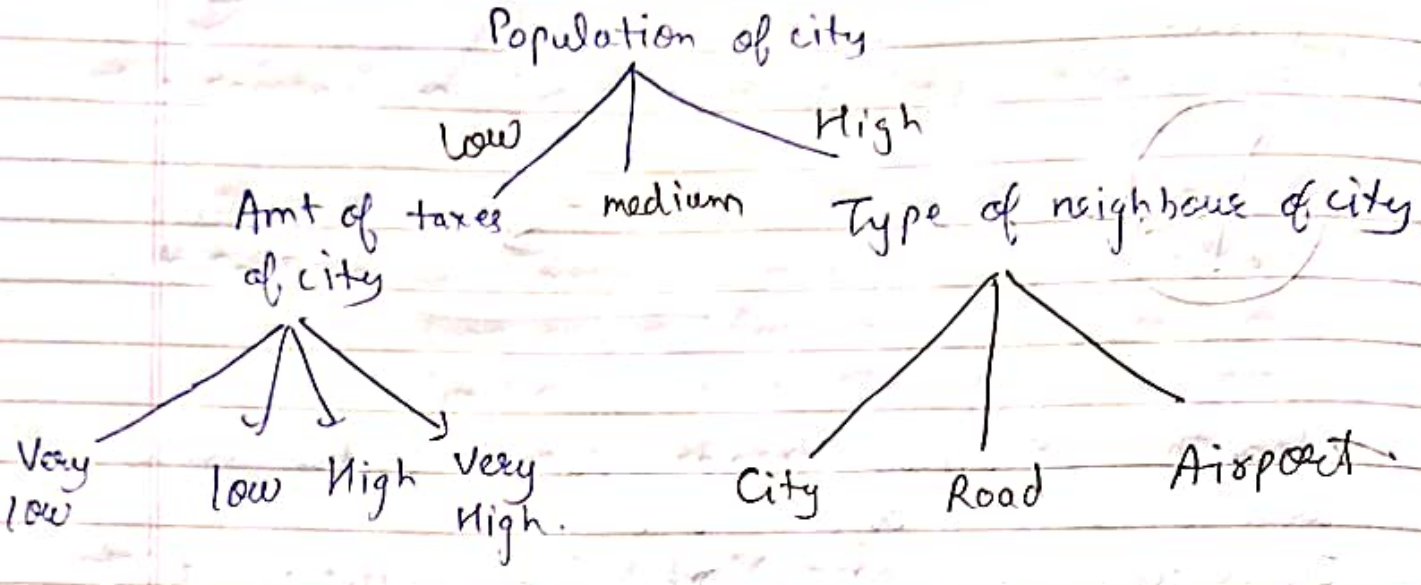


local trend.

Algorithm global-trends detects global trends around a start objects `o'.
The existence of a global trend for a start object `o' indicates that if considering all objects or all paths starting from `o' the values of for the specific attribute in general tend to increase with increasing distance.
Algorithm local trend detects single paths starting from an object `o' & having a certain trend.

* Spatial Classification.

In spatial Classification the attributes values of neighbouring objects may also be relevant for the membership of objects.
Because it is reasonable to assume that the influence of neighbouring objects & their attributes decreases with increasing distance, we can limit the length of the relevant neighbourhood paths by an input parameter max length.



★ Similarity Search in Multimedia Data

When searching similarities in multimedia data, we consider two main families of multimedia indexing & retrieval systems.

① Description based retrieval system

② Content based retrieval system.

Description-based retrieval is labor-intensive if performed manually. If automated the results are typically of poor quality.

Content-based retrieval uses visual features to index images & promotes object retrieval based on feature similarity - ~~which is highly desirable~~ in In this image retrieval system, there are often two kinds of queries.

① Image-sample based queries ② Image features specification queries

Several approach have been proposed & studied for on similarity - based retrieval in image database, based on image signature.

1. Color - histogram based signature
2. Multi feature composed signature
3. Wavelet based signature
4. Wavelet based signature with region - based granularity

* Similarity Search in Time-series Analysis

Time Series Similarity search means finding similarity bet^n two time series. For example a similarity problem having two time series $X = x_1, x_2 \dots x_n$

$Y = y_1, y_2 \dots y_n$; then similarity measure can be computed as SIM (X, Y) which is a measurement of distance bet^n the series.

Ex → Does Stock in X & Y have similar Movements? This type of question can be answered using similarity measures

* Time Series Analysis

Trend → A trend can be viewed as systematic non repetitive changes to the attribute values over time.

Cycle → The long term oscillations about a trend line or curve, which may or may not be periodic.

- seasonal → The detected patterns may be based on time of year or month or day.

- Outliers → Irregular moments occurs in the motion of time series due to events such as labor disputes, floods or announced personal changes in companies.

★ Sequential Pattern Mining

A sequence is an ordered list of item sets or list of attribute values from any domain.

A subsequence of a given sequence is one that can be obtained by removing some items & from the original sequence.

The search of frequent item set subsequences is commonly called sequential pattern mining. Itemset sequences can be searched with the well known AprioriAll algorithm

The task of Sequence matching is divided into two categories.

① Whole Matching → a query time series is matched against a database of individual time series to identify the ones similar to the query.

② Subsequence Matching →
A short query subsequence time series is matched against longer time series by sliding it along the longer sequences, looking for the best matching locat-

* Parameters for Sequential Pattern Mining

(1) Duration of time sequence.

(2) Event Fading Window.

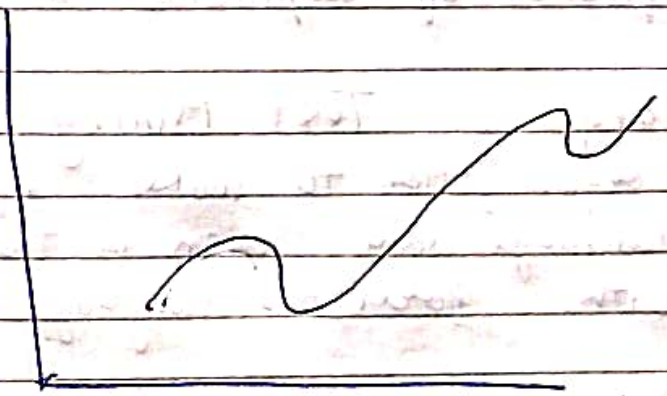
(3) Time Interval.

* Methods for Sequential Pattern Mining

(1) Apriori All
(2) SPADE
(3) GSP
(4) CloSpan
(5) Prefix Span.

* Trend Analysis in Time Series data

Trends are qualitative movements that may exist in a time series dataset. Frequently occuring trends in time series are excellent pointers for understanding the general behavior of the series. They can also be a good start for mining pattern association existing within the time series.



An illustration of repeating trend in time series.

★ Text data Mining.

Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, Stats & computational linguistics.

Types of Text Mining / Application.

① Information extraction
Extracts from free text document entities such as persons, organizations, places, articles as well as events involving these entities.

② Text summarization
Understands document & identifies the most imp. sentences in them.

③ Text Categorization
Automatically organizes documents into user-defined categories or taxonomies

④ Text clustering
Groups together conceptually related documents, enabling identification of duplicates & near duplicates.

⑤ Foreign language Text Mining
Enables an organization to make effective use of foreign language data, even in the absence of staff with foreign language skills

★ Web Mining.

Web mining is mining of data related to the World Wide Web.