Study Material
on

Data Mining For Businees Analysis
Sub Code 18MBA 301E

By

Mr. Bibhu Prasad Mishra
Asst . PROFESSOR
Rajdhani College Of Engg & Management

Module-1

1.1 WhatIsDataMining?

Data mining refers to extracting or mining knowledge from large amountsof data. The term isactuallyamisnomer.Thus,dataminingshouldhavebeenmoreappropriatelynamedasknowledgemining whichemphasis on miningfrom largeamountsof data.
It is the computational process of discovering patterns in large data sets involving methods at theintersectionofartificialintelligence, machine learning,statistics, anddatabasesystems.
The overallgoalof thedata mining processistoextractinformationfromadata setandtransformit into an understandable structureforfurtheruse.

Thekeyproperties ofdata miningare •
Automaticdiscoveryofpatterns•
Prediction oflikelyoutcomes
- Creationofactionableinformation
- Focusonlargedatasetsanddatabases

1.2 TheScopeofDataMining

Data mining derives its name from the similarities between searching for valuable businessinformation in a large database — for example, finding linked products in gigabytes of storescanner data — and mining a mountain for a vein of valuable ore. Both processes require eithersifting through an immense amount of material, or intelligently probing it to find exactly wherethe value resides. Given databases of sufficient size and quality, data mining technology cangeneratenew business opportunities byprovidingthesecapabilities:

Automated prediction of trends and behaviors. Data mining automates the process of findingpredictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of apredictive problem is targeted marketing. Data mining uses data on past promotional mailings toidentify the targets most likely to maximize return on investment in future mailings. Otherpredictive problems include forecasting bankruptcy and other forms of default, and identifyingsegmentsofapopulationlikelyto respond similarlytogiven events.

Automated discoveryof previously unknown patterns. Data mining tools sweep throughdatabases and identify previously hidden patterns in one step. An example of pattern discovery isthe analysis of retail sales data to identify seemingly unrelated products that are often purchasedtogether. Other pattern discovery problems include detecting fraudulent credit card transactionsand identifyinganomalous data that could represent dataentrykeyingerrors.

1.3 Tasks ofDataMining

Datamininginvolves sixcommonclasses oftasks:

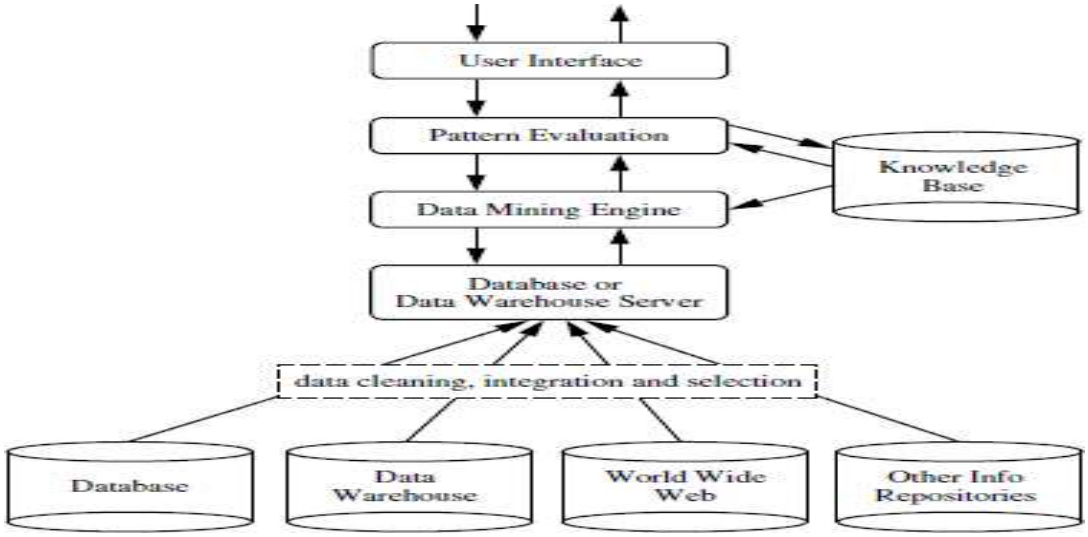
- **Anomalydetection(Outlier/change/deviationdetection)**–
  Theidentificationofunusualdatarecords,thatmightbeinterestingordataerrorsthatrequirefurtherinvestigation.

- **Associationrulelearning(Dependencymodelling)**–Searchesforrelationshipsbetween variables. For example a supermarket might gather data on customer purchasinghabits. Using association rule learning, the supermarket can determine which products arefrequently boughttogetherandusethisinformationfor marketing purposes.Thisissometimesreferred to asmarket basketanalysis.
- **Clustering** – is the task of discovering groups and structures in the data that are in somewayor another"similar",without usingknownstructures in thedata.

- **Classification** – is the task of generalizing known structure to apply to new data. Forexample, an e-mail program might attempt to classify an e-mail as "legitimate" or as"spam".
  Regression –attemptstofindafunctionwhichmodelsthedatawiththeleasterror.
- **Summarization**–
- providingamorecompactrepresentationofthedataset,includingvisualizationand report generation.

1.4 ArchitectureofDataMining

Atypical dataminingsystem mayhavethe followingmajor components.



1. **KnowledgeBase:**
   Thisisthedomainknowledgethatisusedtoguidethesearchor evaluatetheinterestingnessofresultingpatterns.Suchknowledgecanincludeconcept hierarchies, used to organize attributes or attribute values into different levels of abstraction.Knowledgesuchasuserbeliefs,whichcanbeusedtoassessapattern'sinteresting ness based on its unexpectedness, may also be included. Other examples ofdomainknowledgeareadditionalinterestingnessconstraintsorthresholds,andmetadata(e.g., describingdatafrom multiple heterogeneous sources).
2. **DataMiningEngine:**
   This is essential to the data mining system and ideally consists of set of functionalmodulesfortaskssuchascharacterization,associationandcorrelation

analysis,classification,prediction,clusteranalysis,outlieranalysis,andevolution analysis.

3. **PatternEvaluationModule:**
   This component typically employs interestingness measures interacts with the datamining modules so as to focus the search toward interesting patterns. Itmay useinterestingness thresholds to filter out discovered patterns. Alternatively, the patternevaluationmodulemaybeintegrated withtheminingmodule,dependingontheimplementation of the datamining method used. For efficient data mining, it is highlyrecommended to push the evaluation of pattern interestingness as deep as possible intotheminingprocess so as to confine thesearch to onlythe interestingpatterns.
4. **Userinterface:**
   This module communicates between users and the data mining system, allowing theuser to interact with the system by specifying a data mining query or task, providinginformation to help focus the search, and performing exploratory datamining based onthe intermediate data mining results. In addition, this component allows the user tobrowsedatabaseanddatawarehouseschemasordatastructures, evaluateminedpatterns,and visualizethe patterns in different forms.

1.5 DataMining Process:

DataMiningisaprocessofdiscoveringvariousmodels,summaries,andderivedvaluesfromagiven collection ofdata.

Thegeneralexperimentalprocedureadaptedtodata-miningproblemsinvolvesthefollowingsteps:

1. Statetheproblemandformulatethehypothesis

Mostdata-basedmodelingstudiesareperformedinaparticularapplicationdomain.Hence, domain-specific knowledge and experience are usually necessary in order to comeup with a meaningful problem statement. Unfortunately, many application studies tend tofocus on the data-mining technique at the expense of a clear problem statement. In thisstep, a modeler usually specifies a set of variables for the unknown dependency and, ifpossible, a general form of this dependency as an initial hypothesis. There may be severalhypotheses formulated for a single problem atthis stage. The first step requires thecombined expertise of an application domain and a data-mining model. In practice, itusually means a close interaction between the data-mining expert and the applicationexpert. In successful data-mining applications, this cooperation does not stop in the initialphase;it continues duringtheentiredata-mining process.

2. Collectthedata

This step is concerned with how the data are generated and collected. In general, there aretwodistinctpossibilities.Thefirstiswhenthedata-generationprocessisunderthecontrol of an expert (modeler): this approach is known as a designed experiment. Thesecond possibility is when the expert cannot influence the data- generation process: this isknown as the observational approach. An observational setting, namely, random datageneration,isassumedinmostdata-miningapplications.Typically,thesampling distribution is completely unknown after data are collected, or it is partially and implicitlygiven in the data-collection procedure. It is very important, however, to understand howdata collection affects its theoretical distribution, since such a priori knowledge can bevery useful for modeling and, later, for the final interpretation of results. Also, it isimportant to make sure that the data used for estimating a model and the data used laterfor testing and applying a model come from the same, unknown, sampling distribution. Ifthis is not the case, the estimated model cannot be successfully used in a

final applicationofthe results.

3. Preprocessingthedata

In the observational setting, data are usually "collected" from the existing databases, datawarehouses, and data marts. Data preprocessing usually includes at least two commontasks:

1. **Outlier detection (and removal)** – Outliers are unusual data values that are notconsistentwithmostobservations.Commonly,outliersresultfrommeasurementerrors, coding and recording errors, and, sometimes, are natural, abnormal values.Such nonrepresentative samples can seriously affect the model produced later. Therearetwo strategies fordealingwith outliers:

a. Detectandeventuallyremove outliersasapart ofthe preprocessingphase,or

b. Developrobust modelingmethods thatareinsensitive to outliers.

2. **Scaling, encoding, and selecting features** – Data preprocessing includes several stepssuch asvariable scaling and differenttypesof encoding.For example, one feature withthe range [0, 1] and the other with the range [−100, 1000] will not have the same weightsin the applied technique; they will also influence the final data-mining results differently.Therefore,itisrecommendedtoscalethemandbringbothfeaturestothesameweightf orfurtheranalysis.Also,application-specificencodingmethodsusuallyachieve dimensionalityreductionbyprovidingasmallernumberofinformativefeaturesforsubsequentd atamodeling.

Thesetwoclassesofpreprocessingtasksareonlyillustrativeexamplesofalargespectrumof preprocessingactivities in adata-miningprocess.

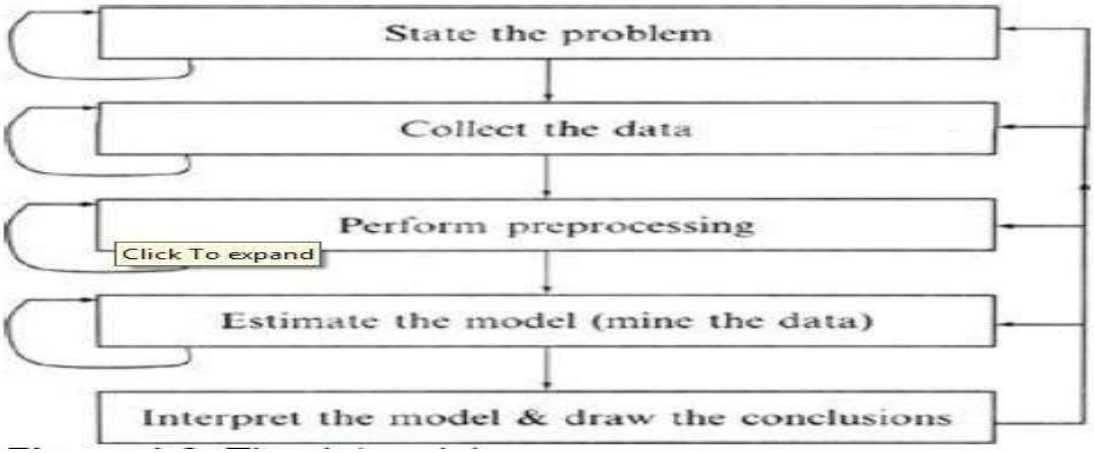
Data-preprocessing steps should not be considered completely independent from otherdata-mining phases. In every iteration of the data-mining process, all activities, together,could define new and improved data sets for subsequent iterations. Generally, a goodpreprocessing method provides an optimal representation for a data-mining technique byincorporatingapriori knowledgeintheformofapplication-specificscalingandencoding.

4. Estimatethemodel

The selection and implementation of the appropriate data-mining technique is the maintaskinthisphase.Thisprocessisnotstraightforward;usually,inpractice,theimplementatio n is based on several models, and selecting the best one is an additionaltask. The basic principles of learning and discovery from data are given in Chapter 4 ofthis book. Later, Chapter 5 through 13 explain and analyze specific techniques that areapplied to perform a successful learning process from data and to develop an appropriatemodel.

5. Interpretthemodelanddrawconclusions

In most cases, data-mining models should help in decision making. Hence, such modelsneed to be interpretable in order to be useful because humans are not likely to base theirdecisions on complex "black-box" models. Note that the goals of accuracy of the modeland accuracy of its interpretation are somewhat contradictory. Usually, simple models aremore interpretable, but they are also less accurate. Modern data-mining methods areexpected to yield highly accurate results using high dimensional models. The problem ofinterpretingthesemodels,alsoveryimportant,isconsideredaseparatetask,withspecific techniques to validate the results. A user does not want hundreds of pages of numericresults. He does not understand them; he cannot summarize, interpret, and use them forsuccessfuldecision making.



TheDataminingProcess

1.6 ClassificationofData mining Systems:

Thedata miningsystemcan be classified accordingtothefollowingcriteria:

- DatabaseTechnology
- Statistics
- MachineLearning
- InformationScience
- Visualization
- OtherDisciplines

SomeOtherClassificationCriteria:

- Classificationaccordingtokindofdatabasesmined
- Classificationaccordingtokindofknowledgemined
- Classificationaccordingtokindsoftechniquesutilized
- Classificationaccordingto applicationsadapted

Classificationaccordingtokindofdatabasesmined

We can classify the data mining system according to kind of databases mined. Database systemcan be classified according to different criteria such as data models, types of data etc. And thedata mining systemcanbe classifiedaccordingly.For example ifweclassify the databaseaccording to data model then we may have a relational, transactional, object- relational, or datawarehouseminingsystem.

Classificationaccordingtokindofknowledgemined

Wecanclassifythe data miningsystemaccordingtokind ofknowledgemined.Itismeansdataminingsystemare classified on thebasisof functionalities such as:

- Characterization
- Discrimination
- AssociationandCorrelationAnalysis
- Classification
- Prediction
- Clustering
- OutlierAnalysis
- EvolutionAnalysis

Classificationaccordingtokindsoftechniquesutilized

Wecanclassifythedataminingsystemaccordingto kindoftechniquesused.Wecandescribesthese techniques according to degree of user interaction involved or the methods of analysisemployed.

Classificationaccordingtoapplicationsadapted

Wecanclassifythe data miningsystemaccordingtoapplication adapted.Theseapplicationsareas follows:

- Finance
- Telecommunications•

DNA

- StockMarkets•

E-mail

1.7 MajorIssuesInDataMining:

•**Mining different kinds of knowledge in databases.** - The need of different users isnotthesame.AndDifferentusermaybeininterestedindifferentkindofknowledge.Thereforeitis necessaryfor data miningto cover broad rangeof knowledgediscoverytask.

•**Interactive mining of knowledge at multiple levels of abstraction.** - The data mining processneeds to be interactive because it allows users to focus the search for patterns, providing andrefiningdata miningrequests based on returned results.

•**Incorporation of background knowledge.** - To guide discovery process and to express thediscovered patterns, thebackground knowledgecan be used. Background knowledge may beusedtoexpressthediscoveredpatternsnotonly inconcisetermsbutatmultiplelevelofabstraction.

•**Data mining query languages and ad hoc data mining.** - Data Mining Query language thatallowstheusertodescribeadhocminingtasks,shouldbeintegratedwithadatawarehousequerylangu ageand optimized forefficient andflexibledata mining.

•**Presentation and visualization of data mining results.** - Once the patterns are discovered itneeds to be expressed in high level languages, visual representations. This representations shouldbeeasilyunderstandablebythe users.

•**Handling noisy or incomplete data.** - The data cleaning methods are required that can handlethe noise, incomplete objects while mining the data regularities. If data cleaning methods are nottherethen theaccuracyof thediscovered patternswill bepoor.

•**Patternevaluation.** -Itreferstointerestingness oftheproblem.Thepatternsdiscoveredshouldbeinterestingbecause either theyrepresent common knowledgeor lack novelty.

- **Efficiency and scalability of data mining algorithms.** - In order to effectively extract theinformation from huge amount of data in databases, data mining algorithm must be efficientandscalable.

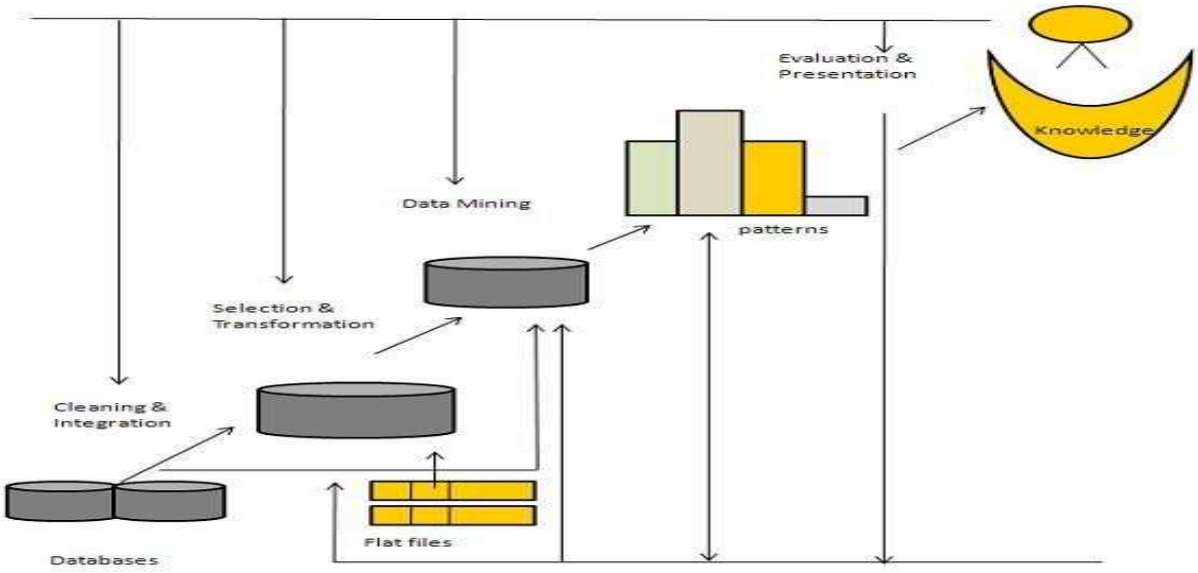
- **Parallel, distributed, and incremental mining algorithms.** - The factors such as huge size ofdatabases,widedistributionofdata, andcomplexity ofdataminingmethodsmotivatethedevelopment of paralleland distributed data mining

algorithms. These algorithms divide thedata into partitions which is further processed parallel. Then the results from the partitions ismerged. The incremental algorithms, updates databases without having mine the data againfromscratch.

1.8 KnowledgeDiscovery inDatabases(KDD)

Some people treat data mining same as Knowledge discovery while some people view datamining essential step in process of knowledge discovery. Here is the list of steps involved inknowledgediscoveryprocess:

- **DataCleaning**- Inthis stepthenoiseand inconsistentdataisremoved.
- **DataIntegration**-Inthisstepmultipledatasourcesarecombined.
- **DataSelection**- Inthissteprelevanttotheanalysistaskareretrieved fromthedatabase.
- **DataTransformation**-Inthisstepdataaretransformedorconsolidatedintoformsappropriate forminingbyperformingsummaryor aggregation operations.
- **DataMining**-Inthisstepintelligentmethodsareappliedinordertoextractdatapatterns.
- **PatternEvaluation**-In thisstep,datapatternsareevaluated.
- **KnowledgePresentation** -Inthisstep, knowledgeisrepresented.



Thefollowingdiagram shows the process of knowledgediscoveryprocess:

ArchitectureofKDD

1.9 DataWarehouse:

Adatawarehouseisasubject-oriented,integrated,time-variantandnon-volatilecollectionofdatain support ofmanagement's decision-makingprocess.

Subject-Oriented: A data warehouse canbe used to analyze a particular subjectarea. Forexample,"sales"can beaparticular subject.

Integrated: A data warehouse integrates data from multiple data sources. For example, source Aand source B may have different ways of identifying a product, but in a data warehouse, therewillbeonlyasingle wayof identifyingaproduct.

Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve

datafrom 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrastswitha transactionssystem,where oftenonly the mostrecentdata iskept.For example,atransaction system may hold the most recent address of a customer, where a data warehouse canholdall addresses associated with a customer.

Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a datawarehouseshould neverbealtered.

1.9.1 DataWarehouseDesignProcess:

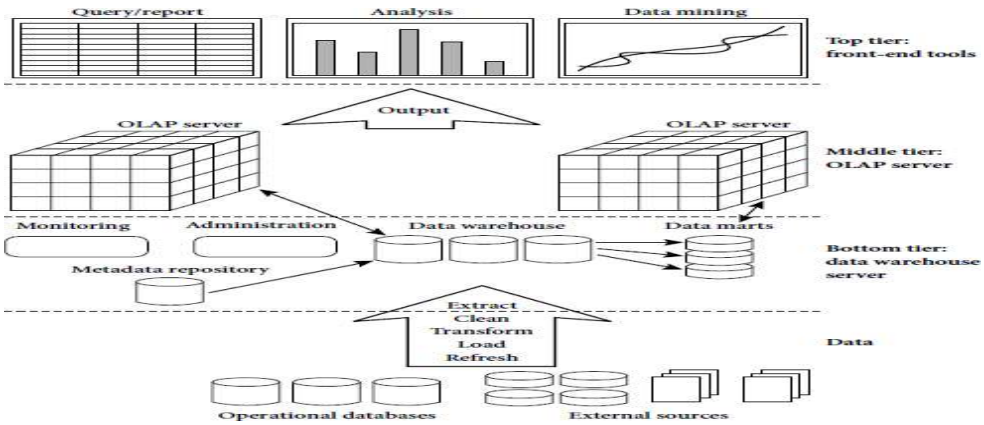
Adata warehouse can be built using a*top-downapproach*, a*bottom-up approach*,or a *combinationof both*.

- The top-down approach starts with the overall design and planning. It is useful in caseswhere the technology is mature and well known, and where the business problems that mustbesolved areclear and well understood.
- The bottom-up approach starts with experiments and prototypes. This is useful in the earlystage of business modeling and technology development. It allows an organization to moveforward at considerably less expense and to evaluate the benefits of the technology beforemakingsignificantcommitments.
- In the combined approach, an organization canexploitthe planned and strategic nature ofthetop-downapproachwhileretainingtherapidimplementationandopportunisticapplicationofthe bottom-up approach.

Thewarehousedesign process consists of thefollowingsteps:

- Choose a business process to model, for example, orders, invoices, shipments, inventory,account administration, sales, or the general ledger. If the business process is organizationalandinvolvesmultiplecomplexobjectcollections,adatawarehousemodelshouldbe followed. However, if the process is departmental and focuses on the analysis of one kind ofbusinessprocess, a datamart model should be chosen.
- Choose the grain of the business process. The grain is the fundamental, atomic level of datato be represented in the fact table for this process, for example, individual transactions,individual dailysnapshots, andso on.
- Choose the dimensionsthat willapply toeachfact table record.Typical dimensions aretime,item, customer, supplier,warehouse, transaction type, andstatus.
- Choose the measures that will populate each fact table record. Typical measures are numericadditivequantities likedollars sold and units sold.

1.9.2 AThreeTierDataWarehouseArchitecture:



Tier-1:

The bottom tier is a warehouse database server that is almost always a relational databasesystem.Back-endtoolsandutilitiesareusedtofeeddataintothebottom tierfromoperationaldatabasesorotherexternalsources(suchascustomerprofile informationprovidedbyexternalconsultants).Thesetoolsandutilitiesperformdataextraction,cleaning, and transformation (e.g., to merge similar data from different sources into aunified format), as well as load and refresh functions to update the datawarehouse. Thedataareextractedusingapplicationprogram interfacesknownasgateways.Agatewayissupported by the underlying DBMS andallows client programs to generate SQL code tobeexecuted at a server.

Examplesof gateways include ODBC (Open Database Connection) and OLEDB (OpenLinking andEmbeddingforDatabases)byMicrosoftandJDBC(JavaDatabaseConnection).

This tier also contains a metadata repository, which stores information aboutthe datawarehouseand its contents.

Tier-2:

The middle tier is an OLAP server that is typically implemented using either a relationalOLAP(ROLAP) modelor amultidimensional OLAP.

- OLAP model is an extended relational DBMS thatmaps operations on multidimensionaldatato standard relational operations.
- AmultidimensionalOLAP(MOLAP)model,thatis,aspecial-purposeserverthatdirectlyimplements multidimensional data and operations.

Tier-3:

The top tier is a front-end client layer, which contains query and reporting tools,analysistools,and/or dataminingtools (e.g.,trendanalysis, prediction,andso on).

1.9.3 DataWarehouse Models:

Therearethreedatawarehousemodels.

1. Enterprisewarehouse:

- An enterprise warehouse collects all of the information about subjects spanning the entireorganization.
- It provides corporate-wide data integration, usually from one or more operational systemsorexternal information providers, and is cross-functional in scope.

- It typically contains detailed data aswell as summarized data, and can range in size from afew gigabytes to hundredsof gigabytes, terabytes, orbeyond.
- An enterprise data warehouse may be implemented on traditional mainframes, computersuper servers, or parallel architecture platforms. It requires extensive business modelingandmaytakeyears to design and build.

2. **Datamart:**
   - A data mart contains a subset of corporate-wide data that is of value to aspecific group ofusers. The scope is confined to specific selected subjects. For example,a marketing datamart may confine its subjects to customer, item, and sales. Thedata contained in datamartstend to be summarized.
   - Datamartsareusuallyimplementedonlow-costdepartmentalserversthatareUNIX/LINUX- or Windows-based. The implementation cycle of a data mart ismorelikely to be measured in weeks rather than months or years. However, itmay involvecomplexintegration inthelongrun ifits designand planningwe arenot enterprise-wide.
   - Dependingonthesourceofdata,datamartscanbecategorizedasindependentordependent.Independentdatamartsaresourcedfromdatacapturedfromoneormoreoperationalsystemsorexternal informationproviders,orfromdatageneratedlocallywithinaparticulardepartmentorgeographical area.Dependentdatamartsaresourceddirectlyfrom enterprise datawarehouses.

3. **Virtualwarehouse:**
   - Avirtualwarehouseisasetofviewsoveroperationaldatabases.Forefficientqueryprocessing, onlysomeofthe possiblesummaryviews maybe materialized.
   - Avirtualwarehouseiseasytobuildbutrequiresexcesscapacityonoperationaldatabaseservers.

1.9.4 MetaData Repository:

Metadata are data about data.When used in a data warehouse, metadata are the data thatdefinewarehouseobjects.Metadataarecreatedforthedatanamesanddefinitionsofthegivenwarehouse.Additionalmetadataarecreatedandcapturedfortimestampinganyextracteddata,the source of the extracted data, and missing fieldsthat have been added by data cleaning orintegrationprocesses.

Ametadata repositoryshould containthefollowing:

- A description of the structure of the data warehouse, which includes the warehouseschema, view, dimensions, hierarchies, and derived data definitions, as well as data martlocationsand contents.
- Operationalmetadata,whichincludedatalineage(historyofmigrateddataandthesequence of transformations applied to it), currency of data (active, archived, or purged),andmonitoringinformation(warehouseusagestatistics,errorreports,and audittrails).
- Thealgorithmsusedforsummarization,whichincludemeasureanddimensiondefinition algorithms,dataongranularity,partitions,subjectareas,aggregation,summarization, andpredefined queriesand reports.
- Themappingfromtheoperationalenvironmenttothedatawarehouse,whichincludes source databases and their contents, gateway descriptions, data partitions, dataextraction, cleaning, transformation rules and defaults, data refresh and purging rules,and

security(user authorization and accesscontrol).

- Data related to system performance, which include indices and profiles that improve dataaccess and retrieval performance, in addition to rules for the timing and scheduling ofrefresh,update,and replication cycles.
- Business metadata, which include business terms and definitions, dataownership information,and chargingpolicies.

1.10 OLAP(OnlineanalyticalProcessing):

- OLAPisanapproachtoansweringmulti-dimensionalanalytical(MDA)queriesswiftly.
- OLAPispartofthebroadercategoryofbusinessintelligence,whichalsoencompassesrelational database, report writingand datamining.
- OLAPtoolsenableuserstoanalyzemultidimensionaldatainteractivelyfrommultipleperspectives.

OLAPconsistsofthreebasicanalyticaloperations:

- Consolidation(Roll-Up)
- Drill-Down
- SlicingAnd Dicing

- Consolidation involves the aggregation of data that can be accumulated and computed inoneormoredimensions.Forexample,allsalesofficesarerolleduptothesalesdepartmentor sales division to anticipatesales trends.
- Thedrill-downisatechniquethatallowsuserstonavigatethroughthedetails.Forinstance,userscanviewthesales byindividualproducts thatmakeup aregion's sales.
- Slicinganddicingisafeaturewherebyuserscantakeout(slicing)aspecificsetofdataofthe OLAP cubeand view (dicing)theslices from differentviewpoints.

1.10.1 TypesofOLAP:

1. RelationalOLAP(ROLAP):

- ROLAP works directly with relational databases. The base data and the dimensiontables are stored as relational tables and new tables are created to hold the aggregatedinformation.It depends on aspecializedschemadesign.
- This methodology relies on manipulating the data stored in the relational database togive the appearance of traditional OLAP's slicing and dicing functionality. In essence,each action of slicing and dicing is equivalent to adding a "WHERE" clause in theSQLstatement.
- ROLAP tools do not use pre-calculated data cubes but instead pose the query to thestandard relational database and its tables in order to bring back the data required toanswerthequestion.
- ROLAPtoolsfeaturetheabilitytoaskanyquestionbecausethemethodologydoesnot limit to the contents of a cube. ROLAP also has the ability to drill down to thelowestlevel of

detail in the database.

2. **MultidimensionalOLAP(MOLAP):**
   - MOLAPisthe 'classic'formofOLAPandissometimesreferredtoasjustOLAP.
   - MOLAP stores this data in an optimized multi-dimensional array storage, rather thanin a relational database. Therefore it requires the pre-computation and storage ofinformationin the cube- the operation known as processing.
   - MOLAPtoolsgenerallyutilizeapre-calculateddatasetreferredtoasadatacube.

   Thedatacubecontainsall thepossibleanswers toagivenrangeofquestions.
   - MOLAPtoolshaveaveryfastresponsetimeandtheabilitytoquicklywritebackdatainto thedata set.
3. **HybridOLAP(HOLAP):**
   - There is no clear agreement across the industry as to what constitutes Hybrid OLAP,exceptthatadatabasewilldividedata betweenrelationalandspecializedstorage.
   - For example, for some vendors, a HOLAP database will use relational tables to holdthe larger quantities of detailed data, and use specialized storage for at least someaspectsof the smallerquantities of more-aggregateor less-detailed data.
   - HOLAPaddressestheshortcomingsofMOLAPandROLAPbycombiningthecapabilitieso fboth approaches.
   - HOLAPtoolscanutilizebothpre-calculated cubesandrelationaldatasources.

1.11 DataPreprocessing:

1.11.1 DataIntegration:

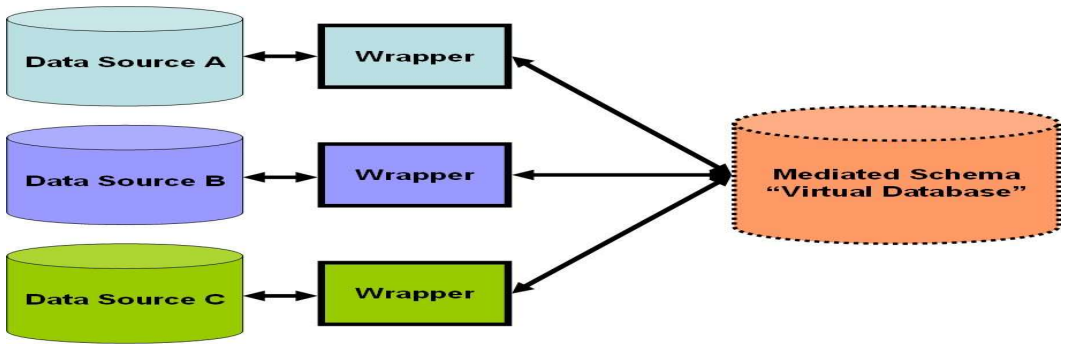
Itcombinesdata frommultiplesourcesintoacoherentdatastore,asindatawarehousing.Thesesources mayincludemultipledatabases, data cubes, or flat files.

Thedataintegrationsystemsareformallydefined

astriple<G,S,M>WhereG: The global schema

S:Heterogeneoussource ofschemas

M:Mappingbetweenthe queriesofsource andglobalschema



1.11.2 IssuesinDataintegration:

1. **Schemaintegrationandobjectmatching:**

   How can the data analyst or the computer be sure that customer id

inonedatabaseandcustomernumberin another referenceto thesame attribute.

2. **Redundancy:**
   An attribute (such as annual revenue, for instance) may be redundant if it can be derivedfrom another attribute or set of attributes. Inconsistencies in attribute or dimension namingcanalso cause redundancies in theresultingdataset.
3. **Detectionandresolutionof data valueconflicts:**
   Forthesamereal-world entity,attributevaluesfrom differentsourcesmaydiffer.

1.11.3 DataTransformation:

Indatatransformation,thedataaretransformedorconsolidatedintoformsappropriate formining.
Datatransformationcaninvolvethefollowing:

- **Smoothing**, which works to remove noise from the data. Such techniques include binning,regression,andclustering.
- **Aggregation**, where summary or aggregation operations are applied to the data. Forexample,thedailysalesdatamaybeaggregatedsoastocomputemonthlyandannual total amounts. This step is typically used in constructing a data cube for analysis ofthedata at multiplegranularities.
- **Generalization ofthedata**, where low-level or ―primitive‖(raw) data are replaced by higher-levelconceptsthroughtheuseofconcepthierarchies.Forexample,categorical attributes, like street, can be generalized to higher-level concepts, like city orcountry.
- **Normalization**,wheretheattributedataarescaledsoastofallwithinasmallspecified range,suchas 1:0 to 1:0, or 0:0 to 1:0.
- **Attribute construction** (or feature construction),wherenewattributes are constructed andaddedfrom thegiven setofattributes to help theminingprocess.

1.11.4 DataReduction:

Data reduction techniques can be applied to obtain a reduced representation of thedata set thatismuch smaller in volume, yet closely maintains the integrity of the originaldata. That is, miningon the reduced data set should be more efficient yet produce thesame (or almost the same)analyticalresults.
Strategies fordatareductionincludethefollowing:

- **Datacubeaggregation**,whereaggregationoperationsareappliedtothedatainthe constructionofadatacube.
- **Attribute subset selection**, where irrelevant, weakly relevant, or redundant attributes ordimensionsmaybedetected and removed.
- **Dimensionality reduction**, where encoding mechanisms are used to reduce the datasetsize.
- **Numerosityreduction**,wherethedataarereplacedorestimatedbyalternative,smallerdatarepr esentationssuchasparametricmodels(whichneedstoreonlythemodelparametersinsteadofthe actualdata)ornonparametricmethodssuchasclustering,sampling,andthe useof histograms.
- **Discretizationandconcepthierarchygeneration**,whererawdatavaluesforattributesare replaced by ranges or higher conceptual levels. Data discretization is a formof numerosityreductionthatisveryusefulfortheautomaticgenerationofconcepthierarchies. Discretizationandconcepthierarchygenerationarepowerfultoolsfordatamining,in that theyallow theminingof dataat multiplelevels of abstraction.

Chapter-2

2.1 Association RuleMining:

- Associationruleminingisa popularand wellresearched methodfordiscoveringinterestingrelationsbetween variables in largedatabases. Itisintendedtoidentifystrongrules discoveredindatabasesusingdifferentmeasuresofinterestingness.
- Basedontheconceptof strongrules, Rakesh Agrawaletal.introducedassociationrules.

ProblemDefinition:

Theproblem ofassociationruleminingis definedas:

Let $I = \{i_1, i_2, \ldots, i_n\}$ beaset of n binaryattributes called *items*.

Let $D = \{t_1, t_2, \ldots, t_m\}$ beasetoftransactionscalled the *database*.

Eachtransactionin D has a uniquetransactionID andcontainsasubsetofthe itemsin I.

A*rule*isdefinedasanimplicationoftheform $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$.

Thesetsofitems(forshort*itemset*) X and Y arecalled*antecedent*(left-hand-sideorLHS)and *consequent*(right-hand-sideorRHS)oftherule respectively.

Example:

Toillustratetheconcepts,weuseasmallexamplefromthesupermarketdomain.Thesetofitems is $I = \{\text{milk}, \text{bread}, \text{butter}, \text{beer}\}$ and a small database containing the items (1codespresenceand 0 absenceofanitem in atransaction) is shownin thetable.

An example rule for the supermarket could be $\{\text{butter}, \text{bread}\} \Rightarrow \{\text{milk}\}$ meaning that ifbutterand bread arebought, customers also buymilk.

Exampledatabasewith4itemsand5transactions

TransactionID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

2.1.1 ImportantconceptsofAssociationRuleMining:

- The**support** $\text{supp}(X)$ ofanitemset X isdefinedastheproportionoftransactionsinthedata set which contain the itemset. In the example database, the itemset $\{\text{milk}, \text{bread}, \text{butter}\}$ hasasupportof $1/5 = 0.2$ sinceitoccursin20%ofalltransactions(1 out of5 transactions).

- The**confidence** ofarule isdefined

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X).$$

For example, the rule $\{\text{butter}, \text{bread}\} \Rightarrow \{\text{milk}\}$hasaconfidenceof $0.2/0.2 = 1.0$inthedatabase,whichmeansthatfor100%ofthetransactionscontaining butter and bread the rule is correct (100% of the times a customer buys butterandbread,milkisboughtaswell).Confidencecanbeinterpretedasanestimateofthe probability$P(Y|X)$,theprobability offinding theRHSoftherule intransactionsunderthecondition that thesetransactionsalso contain theLHS.

- The***lift***ofarule is definedas

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

ortheratiooftheobservedsupporttothatexpectedifXandYwereindependent.Therule $\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$ hasa liftof$\frac{0.2}{0.4 \times 0.4} = 1.25$.

- The**conviction** ofarule isdefinedas

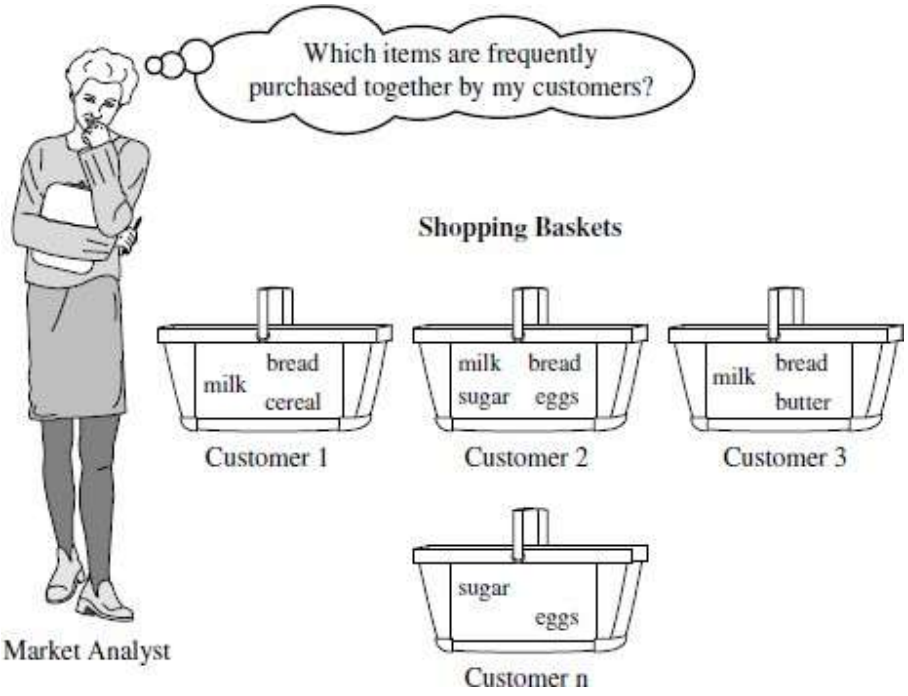
$$\text{conv}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)}$$

Therule$\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$hasa convictionof $\frac{1 - 0.4}{1 - .5} = 1.2$,

andcanbeinterpretedastheratiooftheexpectedfrequencythatXoccurswithoutY(that is to say, the frequency that the rule makes an incorrect prediction) if X and Y wereindependentdivided bytheobservedfrequencyofincorrect predictions.

2.2 Marketbasketanalysis:

This process analyzes customer buying habits by finding associations between the different itemsthat customersplaceintheirshoppingbaskets.Thediscoveryofsuchassociations canhelpretailersdevelopmarketingstrategiesbygaininginsightintowhichitems arefrequentlypurchased together by customers. For instance, if customers are buying milk, how likely are theyto also buy bread (and what kind of bread) on the same trip to the supermarket. Such informationcanlead toincreased salesbyhelpingretailersdo selectivemarketingand plan theirshelfspace.



Example:
If customers who purchase computers also tend to buy antivirus software at the same time, thenplacing the hardware display close to the software displaymay help increase the sales of bothitems. In an alternative strategy, placing hardware and software at opposite ends of the store mayentice customers who purchase such items to pick up other items along the way. For instance,after deciding on an expensive computer, a customer may observe security systems for sale whileheading toward the software display to purchase antivirus software and may decide to purchase ahome security system as well. Market basket analysis can also help retailers plan which items toput on sale at reduced prices. If customers tend to purchase computers and printers together,then havingasaleon printers mayencouragethesale ofprinters*as well as*computers.

2.3 FrequentPattern Mining:

Frequentpattern miningcanbeclassifiedinvarious ways,basedon the followingcriteria:

1. **Basedonthecompletenessof patternstobemined:**
   - Wecanminethecompletesetoffrequentitemsets,theclosedfrequentitemsets,andthemaximalfrequent item sets, given aminimum support threshold.
   - Wecanalsomineconstrainedfrequentitemsets,approximatefrequentitemsets,near-matchfrequent itemsets,top-k frequent itemsetsand soon.
2. **Basedonthelevelsofabstractioninvolvedin theruleset:**
   Somemethodsforassociation ruleminingcanfindrulesatdifferinglevelsofabstraction. Forexample,suppose thatasetofassociationrulesminedincludesthefollowingruleswhereX is avariable representingacustomer:
   buys(X,–computer‖))=>buys(X,–HPprinter‖) (1)
   buys(*X*, *–laptop computer*‖))=>*buys*(*X*,*–HPprinter*‖) (2)

In rule(1)and(2),theitemsboughtarereferenced atdifferentlevelsof abstraction(e.g., –*computer*‖isahigher-level abstraction of–*laptop computer*‖).

3. **Basedonthenumberof datadimensionsinvolvedin therule:**

- If the itemsor attributes inan associationrulereferenceonlyonedimension, thenit isasingle-dimensionalassociation rule.
  buys(X,–computer‖))=>buys(X,–antivirus software‖)

- If a rule references two or more dimensions, such as the dimensions age, income, and buys,then it is a multidimensional association rule. The following rule is an example of amultidimensionalrule:
  age(X,–30,31…39‖) ^ income(X,–42K,…48K‖))=>buys(X,–highresolution TV‖)

4. **Basedonthetypesof valueshandledintherule:**

- Ifaruleinvolvesassociationsbetweenthepresenceorabsenceofitems,itisaBooleanassociationrule.
- Ifaruledescribesassociationsbetweenquantitativeitemsorattributes,thenitisaquantitativeassociation rule.

5. **Basedonthekindsof rulestobemined:**

- Frequentpatternanalysiscangeneratevariouskindsofrulesandotherinterestingrelationships.
- Associationruleminingcan generatealargenumberofrules,manyofwhichareredundantor do not indicate acorrelationrelationship among item sets.
- Thediscoveredassociationscanbefurtheranalyzedtouncoverstatisticalcorrelations,leadingto correlation rules.

6. **Basedonthekindsofpatternstobemined:**

- Manykindsof frequent patternscanbemined fromdifferentkindsof data sets.
- Sequential pattern mining searches for frequent subsequences in a sequence data set,whereasequencerecords an orderingofevents.
- For example, with sequential pattern mining, we can study the order in which items arefrequently purchased. For instance, customers may tend to first buy a PC, followed by adigitalcamera,andthen a memorycard.
- Structuredpatternminingsearchesforfrequentsubstructuresinastructureddataset. •

Singleitems arethesimplest form ofstructure.

- Eachelementof anitemset maycontainasubsequence,asubtree,and soon.
- Therefore,structured patternminingcanbeconsideredasthemostgeneralformoffrequentpattern mining.

2.4 EfficientFrequent ItemsetMiningMethods:

2.4.1 FindingFrequentItem setsUsingCandidateGeneration: The AprioriAlgorithm

- Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for miningfrequentitemsets for Boolean association rules.
- The name of the algorithm is based on the fact that the algorithm uses *prior knowledge*

offrequentitemset properties.

- Apriori employs an iterative approach known as a *level-wise* search, where *k*-itemsets areusedto explore(*k*+1)-itemsets.
- First, the set of frequent 1-itemsets is found by scanning the database to accumulate thecountforeachitem,andcollectingthoseitemsthatsatisfyminimumsupport.Theresulting set is denoted *L*1.Next, *L*1 is used to find *L*2, the set of frequent 2-itemsets,whichis usedto find *L*3, and soon, untilno morefrequent*k*-itemsets canbe found.
- Thefindingof each $L_k$requires onefull scanof thedatabase.
- Atwo-step process is followedin Aprioriconsistingof joinandpruneaction.

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D, a database of transactions;
- min_sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

```
(1)   L1 = find_frequent_1-itemsets(D);
(2)   for (k = 2; Lk−1 ≠ ϕ; k++) {
(3)       Ck = apriori_gen(Lk−1);
(4)       for each transaction t ∈ D { // scan D for counts
(5)           Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)           for each candidate c ∈ Ct
(7)               c.count++;
(8)       }
(9)       Lk = {c ∈ Ck | c.count ≥ min_sup}
(10)  }
(11)  return L = ∪k Lk;
```

```
procedure apriori_gen(Lk−1:frequent (k − 1)-itemsets)
(1)   for each itemset l1 ∈ Lk−1
(2)       for each itemset l2 ∈ Lk−1
(3)           if (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧ ... ∧ (l1[k − 2] = l2[k − 2]) ∧ (l1[k − 1] < l2[k − 1]) then {
(4)               c = l1 ⋈ l2; // join step: generate candidates
(5)               if has_infrequent_subset(c, Lk−1) then
(6)                   delete c; // prune step: remove unfruitful candidate
(7)               else add c to Ck;
(8)           }
(9)   return Ck;
```

```
procedure has_infrequent_subset(c: candidate k-itemset;
          Lk−1: frequent (k − 1)-itemsets); // use prior knowledge
(1)   for each (k − 1)-subset s of c
(2)       if s ∉ Lk−1 then
(3)           return TRUE;
(4)   return FALSE;
```

Example:

TID	Listof itemIDs
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Thereareninetransactionsinthis database,thatis,|D| =9.

Steps:

1. In the first iteration of the algorithm, each item is a member of the set of candidate1-itemsets, C1. The algorithm simply scans all of the transactions in order to countthe number ofoccurrencesof each item.

2. Suppose that the minimum support count required is 2, that is, min sup = 2. The set

offrequent1-itemsets,L1,canthenbedetermined.Itconsistsofthecandidate1-itemsetssatisfying minimum support. In our example, all of the candidates in C1 satisfy minimumsupport.

3. Todiscoverthesetoffrequent2-itemsets,L2,thealgorithmusesthejoinL1onL1to generate a candidate set of 2-itemsets, C2.No candidates are removed fromC2 during theprunestep becauseeachsubsetofthe candidates is also frequent.

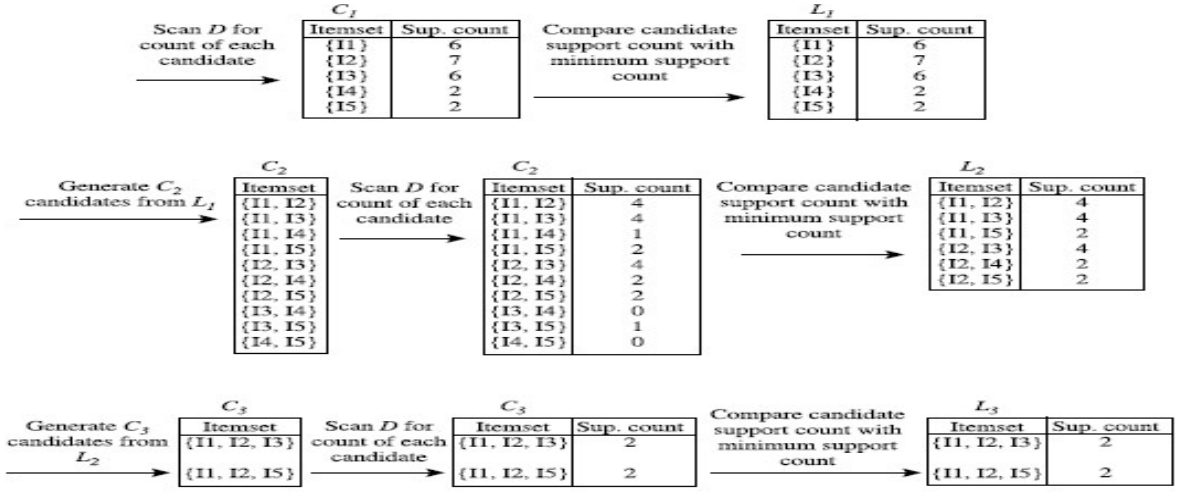
4.Next, the transactions inDare scanned and the support count of each candidate itemsetInC2 isaccumulated.

5. The set of frequent2-itemsets, L2, is then determined, consisting of those candidate2-itemsetsin C2 havingminimum support.

6. The generation of the set of candidate 3-itemsets,C3, Fromthejoin step, we first getC3 =L2xL2 = ({I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4},{I2, I3, I5}, {I2, I4, I5}. Based on theApriori property that all subsets of a frequentitemsetmust also be frequent, we can determinethatthe four lattercandidates cannotpossiblybefrequent.

7.Thetransactionsin Darescannedinorder todetermineL3,consistingof thosecandidate3-itemsetsin C3 having minimum support.

8.ThealgorithmusesL3xL3togenerateacandidatesetof4-itemsets,C4.



Generation of candidate itemsets and frequent itemsets, where the minimum support count is 2.

(a) Join: $C_3 = L_2 \bowtie L_2 = \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\} \bowtie \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$
$= \{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\}, \{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}.$

(b) Prune using the Apriori property: All nonempty subsets of a frequent itemset must also be frequent. Do any of the candidates have a subset that is not frequent?

- The 2-item subsets of {I1, I2, I3} are {I1, I2}, {I1, I3}, and {I2, I3}. All 2-item subsets of {I1, I2, I3} are members of L_2. Therefore, keep {I1, I2, I3} in C_3.
- The 2-item subsets of {I1, I2, I5} are {I1, I2}, {I1, I5}, and {I2, I5}. All 2-item subsets of {I1, I2, I5} are members of L_2. Therefore, keep {I1, I2, I5} in C_3.
- The 2-item subsets of {I1, I3, I5} are {I1, I3}, {I1, I5}, and {I3, I5}. {I3, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I1, I3, I5} from C_3.
- The 2-item subsets of {I2, I3, I4} are {I2, I3}, {I2, I4}, and {I3, I4}. {I3, I4} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I3, I4} from C_3.
- The 2-item subsets of {I2, I3, I5} are {I2, I3}, {I2, I5}, and {I3, I5}. {I3, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I3, I5} from C_3.
- The 2-item subsets of {I2, I4, I5} are {I2, I4}, {I2, I5}, and {I4, I5}. {I4, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I4, I5} from C_3.

(c) Therefore, $C_3 = \{\{I1, I2, I3\}, \{I1, I2, I5\}\}$ after pruning.

Generation and pruning of candidate 3-itemsets, C_3, from L_2 using the Apriori property.

2.4.2 GeneratingAssociationRules fromFrequentItemsets:

Once the frequent itemsets from transactions in a database D have been found, it isstraightforwardto generate strongassociation rules fromthem.

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

The conditional probability is expressed in terms of itemset support count, where $support\_count(A \cup B)$ is the number of transactions containing the itemsets $A \cup B$, and $support\_count(A)$ is the number of transactions containing the itemset A. Based on this equation, association rules can be generated as follows:

- For each frequent itemset l, generate all nonempty subsets of l.
- For every nonempty subset s of l, output the rule "$s \Rightarrow (l - s)$" if $\frac{support\_count(l)}{support\_count(s)} \geq$ min_conf, where min_conf is the minimum confidence threshold.

Example:

Generating association rules. Let's try an example based on the transactional data for *AllElectronics* shown in Table 5.1. Suppose the data contain the frequent itemset $l = \{I1, I2, I5\}$. What are the association rules that can be generated from l? The nonempty subsets of l are $\{I1, I2\}$, $\{I1, I5\}$, $\{I2, I5\}$, $\{I1\}$, $\{I2\}$, and $\{I5\}$. The resulting association rules are as shown below, each listed with its confidence:

$$\begin{aligned}
&I1 \wedge I2 \Rightarrow I5, && confidence = 2/4 = 50\% \\
&I1 \wedge I5 \Rightarrow I2, && confidence = 2/2 = 100\% \\
&I2 \wedge I5 \Rightarrow I1, && confidence = 2/2 = 100\% \\
&I1 \Rightarrow I2 \wedge I5, && confidence = 2/6 = 33\% \\
&I2 \Rightarrow I1 \wedge I5, && confidence = 2/7 = 29\% \\
&I5 \Rightarrow I1 \wedge I2, && confidence = 2/2 = 100\%
\end{aligned}$$

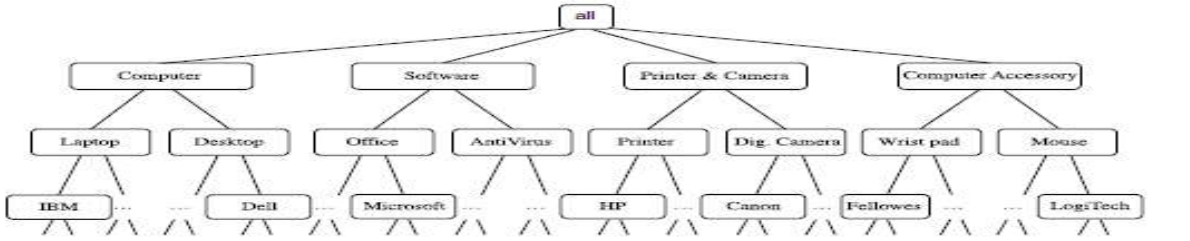
2.5 MiningMultilevelAssociationRules:

- For many applications, it is difficult to find strong associations among data items at loworprimitivelevels of abstraction dueto thesparsityofdata at thoselevels.
- Strongassociationsdiscovered athigh levelsof abstractionmayrepresentcommonsenseknowledge.
- Therefore,dataminingsystemsshouldprovide capabilitiesforminingassociationrulesat multiple levels of abstraction, with sufficient flexibility for easy traversalamongdifferentabstractionspaces.
- Association rules generated from mining data at multiple levels of abstraction arecalledmultiple-levelor multilevel association rules.
- Multilevel association rules can be mined efficiently using concept hierarchies under asupport-confidenceframework.
- Ingeneral,atop-downstrategy isemployed,wherecountsareaccumulatedforthecalculation of

frequent itemsets at each concept level, starting at the concept level 1 andworkingdownward inthe hierarchy towardthemore specificconceptlevels,untilnomorefrequent itemsets can befound.

Aconcept hierarchydefinesasequenceofmappingsfromasetoflow-levelconceptstohigherlevel,moregeneralconcepts.Datacanbegeneralizedbyreplacinglow-levelconceptswithinthedatabytheirhigher-levelconcepts,orancestors,from a concepthierarchy.

TID	Items Purchased
T100	IBM-ThinkPad-T40/2373, HP-Photosmart-7660
T200	Microsoft-Office-Professional-2003, Microsoft-Plus!-Digital-Media
T300	Logitech-MX700-Cordless-Mouse, Fellowes-Wrist-Rest
T400	Dell-Dimension-XPS, Canon-PowerShot-S400
T500	IBM-ThinkPad-R40/P4M, Symantec-Norton-Antivirus-2003
...	...



A concept hierarchy for *AllElectronics* computer items.

Theconcepthierarchyhasfivelevels,respectivelyreferredtoaslevels0to4,startingwithlevel0at theroot node forall.

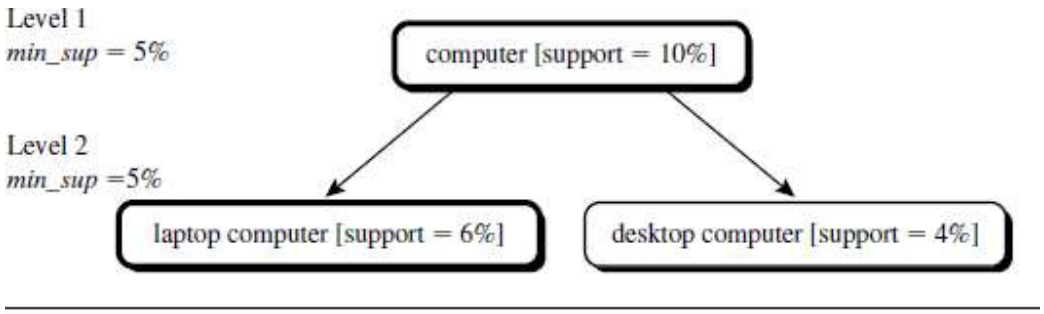
- Here,Level1includescomputer,software,printer&camera, andcomputeraccessory. • Level2includeslaptopcomputer,desktopcomputer,officesoftware,antivirussoftware• Level3 includesIBM desktopcomputer, .. ., Microsoftofficesoftware,andso on.
  Level4isthemostspecificabstractionlevelof thishierarchy.

•

2.5.1 ApproachesForMiningMultilevelAssociation Rules:

1.Uniform MinimumSupport:

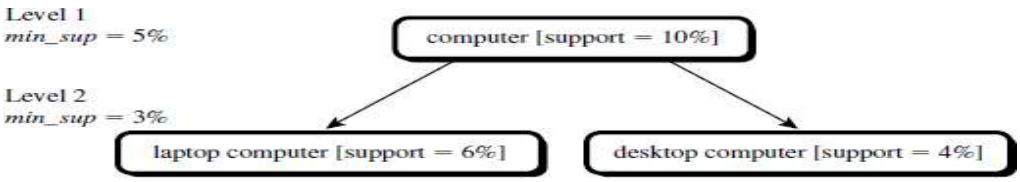
- The same minimum support threshold is used when mining at each level of abstraction.When a uniform minimum support threshold is used, the search procedure is simplified.The method is also simple in that users are required to specify only one minimum supportthreshold.
- The uniform support approach, however, has some difficulties. It is unlikely that items atlowerlevelsof abstractionwilloccurasfrequentlyasthoseathigherlevels ofabstraction.
- If the minimum support threshold is set too high, it could miss some meaningful associationsoccurring at low abstraction levels. If the threshold is set too low, it may generate manyuninterestingassociations occurringat high abstraction levels.



Multilevel mining with uniform support.

2. ReducedMinimumSupport:

- Eachlevelofabstractionhasitsownminimumsupportthreshold.
- Thedeeper thelevel of abstraction, the smallerthecorrespondingthreshold is.
- Forexample,theminimumsupportthresholdsforlevels1and2are5%and3%,respectively.In this way,-computer,‖-laptop computer,‖and -desktop computer‖areall considered frequent.



3. Group-BasedMinimumSupport:

- Because users or experts often have insight as to which groups are more important thanothers, it is sometimes more desirable to set up user-specific, item, or group based minimalsupportthresholds whenminingmultilevel rules.
- For example, a user could set up the minimum support thresholds based on product price, oronitemsofinterest,suchasbysettingparticularlylowsupportthresholdsforlaptopcomputers and flash drives in order to pay particular attention to the association patternscontainingitems in these categories.

2.6 MiningMultidimensionalAssociationRules fromRelationalDatabasesandDataWarehouses:

- Singledimensionalorinterdimensionalassociationrulecontainsasingledistinctpredicate (e.g., buys)with multiple occurrencesi.e., the predicate occurs more than oncewithinthe rule.
  buys(*X*, -*digital camera*‖)=>*buys*(*X*, -*HP printer*‖)
- Association rules that involve two or more dimensions or predicates can be referred toasmultidimensional association rules.
  age(X,"20…29")^occupation(X,"student")=>buys(X,"laptop")
- Above Rule contains three predicates (age, occupation, and buys), each of which occursonlyoncein therule. Hence, wesaythat it has no repeated predicates.
- Multidimensionalassociationruleswithnorepeatedpredicatesare calledinterdimensionalassociation rules.
- We can also mine multidimensional association rules with repeated predicates, whichcontainmultipleoccurrencesofsomepredicates. Theserulesarecalledhybrid-dimensional association rules. An example of sucha rule is the following, where thepredicatebuys is repeated:
  age(X,-20…29‖)^buys(X,-laptop‖)=>buys(X,-HPprinter‖)

2.7 MiningQuantitativeAssociationRules:

- Quantitative association rules are multidimensional association rules in which the numericattributes are *dynamically* discretized during the mining process so as to satisfy some miningcriteria,suchas maximizingthe confidenceorcompactness ofthe rulesmined.
- Inthissection,wefocusspecificallyonhowtominequantitativeassociationruleshavingtwoquantitativeattributesontheleft-handsideoftheruleandonecategoricalattributeontheright-hand sideof the rule. That is

*Aquan*1^*Aquan*2=>*Acat*

where*Aquan*1and*Aquan*2aretestsonquantitative attributeinterval *Acat*testsacategoricalattributefromthetask-relevantdata.

- Such rules have been referred to as two-dimensional quantitative association rules,becausetheycontain twoquantitative dimensions.
- Forinstance,supposeyouarecuriousabouttheassociationrelationshipbetweenpairsofquantitative attributes, like customer age and income, and the type of television (such as*high-definitionTV,* i.e.,*HDTV*) that customers like tobuy.

Anexampleofsucha2-D quantitativeassociation ruleis

age(*X*, ‒30…39‖)^*income*(*X*, ‒42*K*…48*K*‖)=>*buys*(*X*, ‒*HDTV*‖)

2.8 FromAssociationMiningtoCorrelationAnalysis:

- A correlation measure can be used to augment the support-confidence framework forassociationrules. This leads to*correlationrules*ofthe form *A*=>*B*[*support*,*confidence*,*correlation*]
- That is, a correlation rule is measured not only by its support and confidence but also bythecorrelationbetweenitemsetand*B*.Therearemanydifferentcorrelationmeasures from which to choose. In this section, we study various correlation measuresto determinewhich wouldbegoodforminglargedata sets.

- Liftisasimple correlationmeasurethatisgivenasfollows. Theoccurrenceofitemset*A* is independent of the occurrence of itemset*B*if $P(A \cup B) = P(A)P(B)$; otherwise,itemsets*A*and *B* are dependent and correlated as events. This definition can easily beextendedto morethan two itemsets.

Thelift betweentheoccurrenceof *A*and*B*can be measuredbycomputing

- $lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$.
- If the lift(A,B) is less than 1, then the occurrence of A is negatively correlated with theoccurrenceofB.
- If the resulting value is greater than 1, then A and B are positively correlated, meaning thatheoccurrenceof oneimplies the occurrenceof theother.
- If theresultingvalueisequal to1,thenAandB areindependentand there isno correlationbetweenthem.

Chapter-3

3.1 ClassificationandPrediction:

- Classification and prediction are two forms of data analysis that can be used to extract modelsdescribingimportant dataclasses or to predict futuredata trends.
- Classification predictscategorical(discrete,unordered)labels,*prediction*modelscontinuousvaluedfunctions.
- Forexample, we can build a classification model to categorize bank loan applicationsaseithersafeorrisky,orapredictionmodeltopredicttheexpenditures ofpotentialcustomersoncomputer equipmentgiven their incomeand occupation.
- A predictor is constructed that predicts a continuous-valued function, or ordered value, asopposedto a categorical label.
- Regressionanalysisisastatistical methodologythatismostoftenusedfornumericprediction.
- Many classification and prediction methods have been proposed by researchers in machinelearning,patternrecognition, and statistics.
- Most algorithms are memory resident, typically assuming a small data size. Recent
- datamining research has built on such work, developing scalable classification and predictiontechniquescapable ofhandlinglargedisk-residentdata.

3.1.1 IssuesRegardingClassificationandPrediction:

1.PreparingtheDataforClassificationandPrediction:

Thefollowingpreprocessingstepsmaybeappliedto thedata tohelp improvetheaccuracy,efficiency,and scalabilityoftheclassificationorprediction process.

(i) Datacleaning:

- This refers to the preprocessing of data in order to remove or reduce *noise* (by applyingsmoothing techniques) and the treatment of *missing values*(e.g., by replacing a missingvalue with the most commonly occurring value for that attribute, or with the most probablevaluebased on statistics).
- Althoughmostclassificationalgorithmshave somemechanismsfor handling noisy ormissingdata, this step can help reduceconfusionduringlearning.

(ii) Relevanceanalysis:

- Manyoftheattributesinthedata maybe*redundant*.
- Correlationanalysiscanbeusedtoidentifywhetheranytwogivenattributesarestatistically
- related.
- For example, a strong correlation between attributes $A1$ and $A2$ would suggest that one ofthetwo could be removed from further analysis.
- Adatabasemayalsocontain*irrelevant*attributes.Attributesubsetselectioncanbeusedintheseca
- ses tofindareducedsetofattributessuchthattheresultingprobabilitydistribution of the data classes is as close as possible to the original distribution obtainedusingall attributes.
- Hence,relevanceanalysis,intheformofcorrelationanalysisandattributesubsetselection, can be used to detect attributes that do not contribute to the classification orpredictiontask.
- Suchanalysiscanhelpimproveclassificationefficiencyandscalability.

(iii) DataTransformationAndReduction

- Thedatamaybetransformedbynormalization,particularlywhenneuralnetworksormethodsi
-

nvolvingdistancemeasurements areused in thelearningstep.
Normalizationinvolvesscalingallvaluesforagivenattributesothattheyfallwithinasmallspecifiedrange, such as-1 to +1 or 0 to 1.
The data can also be transformed by *generalizing* it to higher-level concepts.

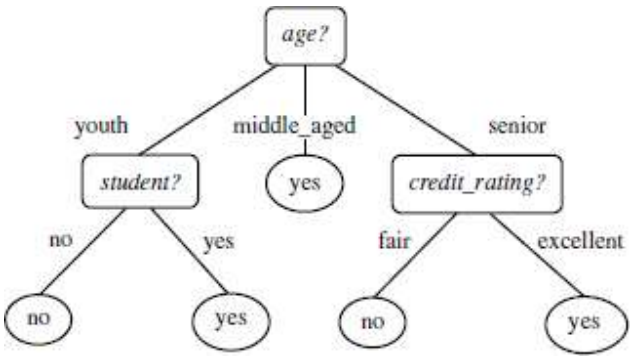
- Concepthierarchiesmaybeusedforthispurpose.Thisisparticularlyusefulforcontinuousvalued attributes.
- For example, numeric values for the attribute *income* can be generalized to discreteranges, such as *low, medium*, and *high*. Similarly, categorical attributes, like *street*, canbegeneralized to higher-level concepts, like *city.*
  Data can also be reduced by applying many other methods, ranging from wavelettransformationandprinciplecomponentsanalysis
- todiscretizationtechniques,suchasbinning, histogram analysis, andclustering.

3.1.2 ComparingClassificationandPredictionMethods:

- **Accuracy:**
  - The accuracy of a classifier refers to the ability of a given classifier to correctly predicttheclasslabelofneworpreviouslyunseendata(i.e.,tupleswithoutclasslabelinformation).
  - The accuracy of a predictor refers to how well a given predictor can guess the value ofthepredictedattributefornew orpreviouslyunseen data.
- **Speed:**
  Thisreferstothe computationalcosts involvedingeneratingandusingthegiven classifier orpredictor.
- **Robustness:**
  Thisis the abilityof theclassifier orpredictorto makecorrect predictionsgivennoisydata or datawith missingvalues.
- **Scalability:**
  This refers to the ability to construct the classifier or predictor efficientlygivenlargeamounts of data.
- **Interpretability:**
  - This refers to the level of understanding and insight that is provided by the classifier orpredictor.
  - Interpretabilityis subjectiveand thereforemoredifficult to assess.

3.2 Classificationby DecisionTreeInduction:

- Decisiontreeinductionisthelearningofdecisiontreesfromclass-labeled trainingtuples.
- Adecision treeis aflowchart-liketreestructure, where
  - Eachinternalnode denotesatestonanattribute.
  - Eachbranchrepresents anoutcomeofthetest.
  - Eachleafnodeholdsa classlabel.
  - Thetopmost nodeina treeistheroot node.



- The construction of decision tree classifiers does not require any domain knowledge orparametersetting,and thereforeIappropriate for exploratoryknowledgediscovery.
- Decisiontrees canhandlehighdimensional data.
- Their representation of acquired knowledge in tree forms intuitive and generally easy toassimilatebyhumans.
- Thelearningandclassificationstepsofdecisiontree inductionaresimpleandfast.● Ingeneral, decision tree classifiers havegoodaccuracy.
- Decision tree induction algorithms have been used for classification in many applicationareas, such as medicine, manufacturing and production, financial analysis, astronomy, andmolecularbiology.

3.2.1 AlgorithmForDecisionTreeInduction:

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split point* or *splitting subset*.

Output: A decision tree.

Method:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

Thealgorithmiscalledwiththreeparameters:

- Datapartition
- Attributelist
- Attributeselectionmethod

- Theparameterattributelistisa listof attributesdescribingthetuples.
- Attributeselectionmethodspecifiesaheuristic procedure forselectingthe attributethat ―best‖discriminatesthegiven tuples accordingtoclass.
- Thetreestartsas asingle node, *N*, representingthetrainingtuplesin*D*.

- If the tuples in *D* are all of the same class, then node *N* becomes a leaf and is labeledwiththat class.
- Alloftheterminatingconditionsareexplainedattheendofthe algorithm.
- Otherwise, the algorithm calls Attribute selection method to determine the splittingcriterion.
- Thesplittingcriterion tellsus which attributeto test at nodeNbydeterminingthe–best‖ wayto separate orpartition the tuples in D into individual classes.

Therearethreepossiblescenarios.Let *A*bethesplittingattribute.*A*has*v*distinctvalues, {a_1,a_2,…,a_v},basedonthetrainingdata.

1 Aisdiscrete-valued:

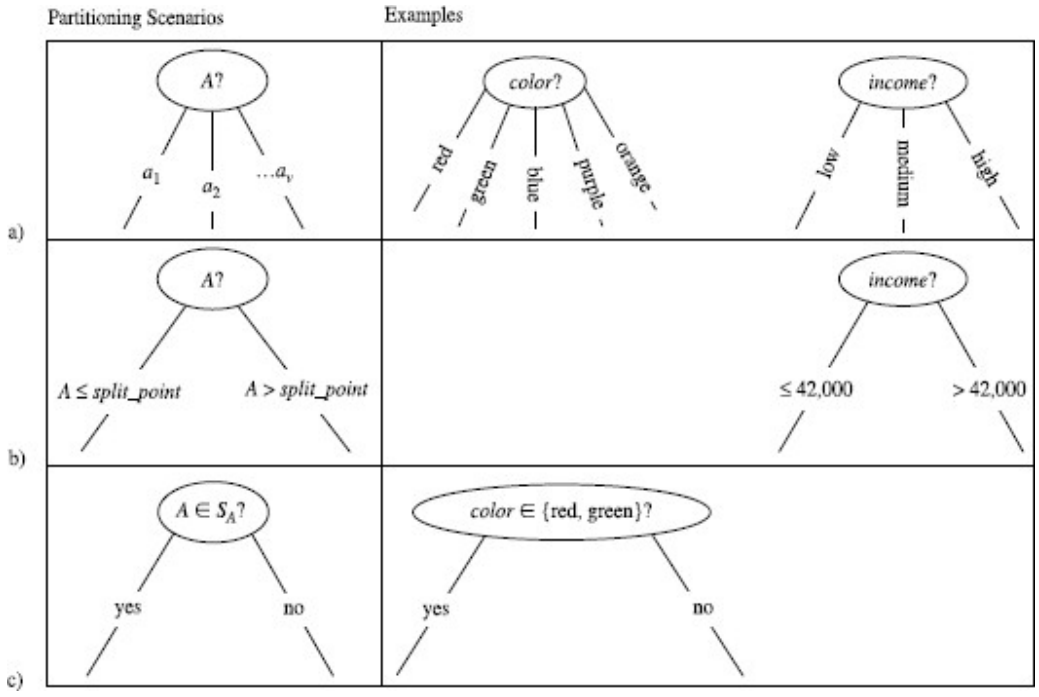
- Inthis case, the outcomes of thetest at node Ncorresponddirectlytotheknownvalues ofA.
- A branch is created for each known value, aj, of A and labeled with that
- value.A neednot beconsideredin anyfuturepartitioning of thetuples.

2 Ais continuous-valued:

In this case, the test at node N has two possible outcomes, corresponding to the conditionsA<=split pointand A>split point, respectively
Where splitpointisthesplit-pointreturnedbyAttributeselectionmethodaspartofthesplittingcriterion.

3 Aisdiscrete-valuedandabinarytreemustbeproduced:

Thetestat nodeNis of the form–A€SA?‖.
SAisthesplittingsubsetforA,returnedbyAttributeselectionmethod aspartofthesplittingcriterion.It is a subset oftheknown values of A.



(a) If A is Discrete valued (b)If A is continuous valued (c) IfA is discrete-valued and a binarytreemust beproduced:

3.3 BayesianClassification:

- Bayesianclassifiersarestatisticalclassifiers.
- Theycan predictclassmembership probabilities,such as theprobabilitythata giventuplebelongstoa particularclass.
- Bayesianclassificationisbasedon Bayes' theorem.

3.3.1 Bayes' Theorem:

- LetXbeadatatuple.InBayesianterms,Xisconsidered‐evidence.‖anditisdescribedby measurementsmadeon aset ofn attributes.
- LetHbesomehypothesis,such asthatthedatatupleXbelongstoa specifiedclassC.
- Forclassificationproblems,wewanttodetermineP(H|X),theprobabilitythatthehypothesisHholds given the‐evidence‖or observeddata tuple X.
  P(H|X)istheposteriorprobability,oraposterioriprobability,ofHconditionedonX.
- Bayes' theoremisusefulinthatitprovidesawayofcalculatingtheposteriorprobability, $P(H|\boldsymbol{X})$,from$P(H)$,$P(\boldsymbol{X}|H)$,and$P(\boldsymbol{X})$.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}.$$

3.3.2 NaïveBayesianClassification:

Thenaïve Bayesianclassifier,orsimpleBayesian classifier,worksasfollows:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple isrepresentedbyann-dimensionalattributevector,X=(x1,x2,…,xn),depictingnmeasurementsmadeon thetuplefrom nattributes,respectively, A1, A2,…,An.

2. Suppose that there are m classes, C1, C2, …, Cm. Given a tuple, X, the classifier willpredictthatXbelongstotheclasshavingthehighestposteriorprobability,conditionedonX. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and onlyif

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \le j \le m, j \ne i.$$

ThuswemaximizeP(Cij**X**).TheclassCiforwhichP(Cij**X**)ismaximizediscalledthemaxim umposteriori hypothesis. ByBayes' theorem

3. AsP(X)isconstant forallclasses, onlyP(X|Ci)P(Ci)need bemaximized.Iftheclass

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}.$$

priorprobabilitiesarenotknown,thenitiscommonlyassumedthattheclassesareequallylikely,that is,P(C1)=P(C2)=…=P(Cm),and wewouldthereforemaximizeP(X|Ci). Otherwise,wemaximizeP(X|Ci)P(Ci).

4. Given data sets with many attributes, it would be extremely computationally expensivetocompute P(X|Ci). In order to reduce computation in evaluating P(X|Ci), the naive assumptionof class conditional independence is made. This presumes that the values of the attributesareconditionallyindependentofoneanother,giventheclasslabelofthetuple.Thus,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i).$$

Wecaneasilyestimatetheprobabilities$P(x1|Ci)$,$P(x2|Ci)$,:::,$P(xn|Ci)$fromthetrainingtuples.Foreachattribute,welookatwhethertheattributeiscategoricalorcontinuous-valued.Forinstance, to compute$P(\boldsymbol{X}|Ci)$, weconsiderthefollowing:

- If$A_k$iscategorical,then$P(x_k|Ci)$isthenumberoftuplesofclassCiin$D$havingthevalue $x_k$forA_k, divided by$|C_{i,D}|$ the number oftuples ofclass$C_i$in$D$.
- If $A_k$is continuous-valued, thenweneedtodoabitmorework,but the calculationis prettystraightforward.

Acontinuous-valuedattributeistypicallyassumedtohaveaGaussiandistributionwithamean$\mu$and standard deviation, defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}).$$

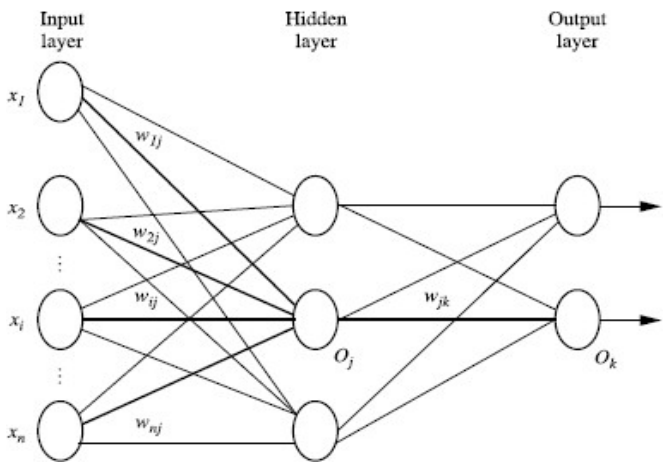
5.Inordertopredictttheclasslabelof X,$P(X|Ci)P(Ci)$isevaluatedforeachclass Ci.Theclassifier predicts thattheclass labelof tuple$X$is theclassCiif and onlyif

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad \text{for } 1 \le j \le m, j \ne i.$$

3.4 AMultilayer Feed-ForwardNeuralNetwork:

- Thebackpropagationalgorithmperformslearningonamultilayerfeed-forward neuralnetwork.
  Ititerativelylearnsasetofweightsforpredictionoftheclasslabeloftuples.
- Amultilayerfeed-forwardneuralnetworkconsistsofaninputlayer,oneormorehiddenlayers,andanoutput layer.

Example:



- The inputs to the network correspond to the attributes measured for each training tuple. Theinputs are fed simultaneously into the units making up the input layer. These inputs passthrough the input layer and are then weighted and fed simultaneously to a second layerknownas a hidden layer.
  The outputs of the hidden layer units can be input to another hidden layer, and so on. Thenumberof hidden layersis arbitrary.
- The weighted outputs of the last hidden layer are input to units making up the output
- layer,whichemits thenetwork'sprediction forgiven tuples

3.4.1 ClassificationbyBackpropagation:

Backpropagationisaneural networklearningalgorithm.

- Aneuralnetworkisasetofconnectedinput/outputunitsin whicheachconnectionhasaweight
- associatedwith it.
  Duringthelearningphase,thenetworklearnsbyadjustingtheweightssoastobeabletopredictthe correct class label of theinput tuples.
- Neuralnetworklearningisalsoreferredtoasconnectionistlearningduetotheconnectionsbetween
- units.
- Neuralnetworksinvolvelongtrainingtimesandarethereforemoresuitableforapplications wherethis is feasible.
- Backpropagation learns by iteratively processing a data set of training tuples, comparingthenetwork's predictionforeach tuplewiththeactual knowntarget value.
- The target value may be the known class label of the training tuple (for classificationproblems)oracontinuous value (forprediction).
- For each training tuple, the weights are modified so as to minimize the mean squarederror betweenthe network's prediction andthe actual target value. Thesemodificationsare made inthe–backwards‖ direction,thatis, fromtheoutput layer,througheachhidden layerdown to thefirst hiddenlayer hencethenameisbackpropagation.
- Although it is not guaranteed, in general the weights will eventually converge, and thelearningprocess stops.

Advantages:

- Itinclude their high tolerance of noisy data as well as their ability to classify patterns onwhich theyhavenot beentrained.
- They can be used when you may have little knowledge of the relationships betweenattributes andclasses.
- They are well-suited for continuous-valued inputs and outputs, unlike most decision treealgorithms.
- They have been successful on a wide array of real-world data, including handwrittencharacter recognition, pathology and laboratory medicine, and training a computer topronounceEnglish text.
- Neural network algorithms are inherently parallel; parallelization techniques can be usedtospeed up the computation process.

Process:

Initializetheweights:

Theweightsinthenetworkareinitializedtosmallrandomnumbers
ranging from-1.0 to 1.0, or -0.5 to 0.5. Each unit has a *bias* associated with it. The biases aresimilarlyinitialized to small random numbers.

Eachtrainingtuple, ***X***, is processed bythefollowingsteps.

Propagatetheinputsforward:

First, the training tuple is fed to the input layer of thenetwork. The inputs pass through the inputunits, unchanged. That is, for an input unit*j*, its output, *Oj*, is equal to its input value, *Ij*. Next, thenet input and output of eachunit in the hidden and output layers are computed. The net input to aunitin thehiddenoroutputlayers is computed asalinear combination ofitsinputs.

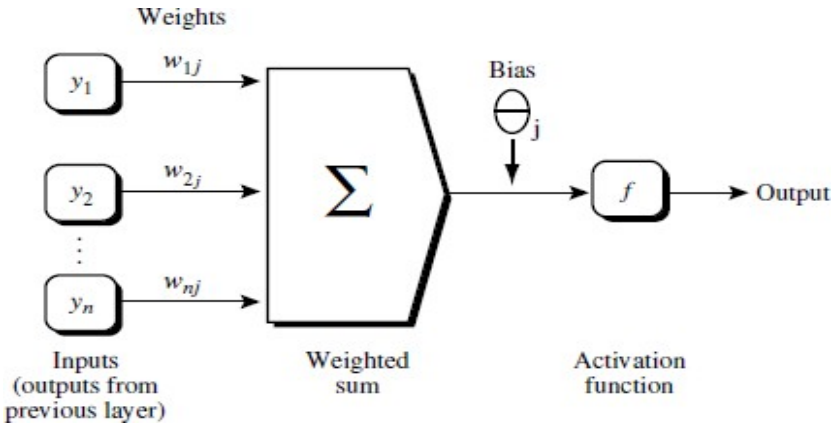
Eachsuch unithasanumber of inputstoitthatare,infact,the outputs of the unitsconnected toitin theprevious layer. Each connection has a weight. To compute the net input to the unit, eachinputconnected to theunit is multiplied byits corresponding weight, and thisis summed.

$$I_j = \sum_i w_{ij} O_i + \theta_j,$$

where$w_{i,j}$is the weight of the connection from unit iin the previous layer to unit j;$O_i$istheoutput of unitifrom theprevious layer

$\theta_j$is thebias ofthe unit &itactsas athreshold in that it servesto varytheactivityof theunit.

Eachunitinthehiddenandoutputlayers takesitsnetinputandthenappliesan activationfunctionto it.



Backpropagatetheerror:

Theerrorispropagatedbackwardbyupdatingtheweightsandbiasestoreflecttheerrorofthenetwork's prediction.Foraunit*j* intheoutput layer, theerror$Err_j$is computedby

$$Err_j = O_j(1 - O_j)(T_j - O_j)$$

where$O_j$istheactualoutputofunitj,and$T_j$istheknowntargetvalueofthegiventrainingtuple.

Theerror ofahidden layerunit*j* is

$$Err_j = O_j(1 - O_j) \sum_k Err_k w_{jk}$$

where$w_{jk}$is the weight of the connection from unit j to a unit k in the next higher layer,andErr_kistheerror ofunitk.

Weights areupdatedbythefollowing equations,where D*wij*isthechangeinweightw_{ij}:

$$\Delta w_{ij} = (l) Err_j O_i$$

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

Biasesareupdated bythefollowingequations below

$$\Delta \theta_j = (l) Err_j$$

$$\theta_j = \theta_j + \Delta \theta_j$$

Algorithm:

Input:

- D, a data set consisting of the training tuples and their associated target values;
- l, the learning rate;
- *network*, a multilayer feed-forward network.

Output: A trained neural network.

Method:

(1) Initialize all weights and biases in *network*;
(2) **while** terminating condition is not satisfied {
(3) **for** each training tuple $\boldsymbol{X}$ in D {
(4) // Propagate the inputs forward:
(5) **for** each input layer unit j {
(6) $O_j = I_j$; // output of an input unit is its actual input value
(7) **for** each hidden or output layer unit j {
(8) $I_j = \sum_i w_{ij} O_i + \theta_j$; //compute the net input of unit j with respect to the previous layer, i
(9) $O_j = \frac{1}{1+e^{-I_j}}$; } // compute the output of each unit j
(10) // Backpropagate the errors:
(11) **for** each unit j in the output layer
(12) $Err_j = O_j(1 - O_j)(T_j - O_j)$; // compute the error
(13) **for** each unit j in the hidden layers, from the last to the first hidden layer
(14) $Err_j = O_j(1 - O_j) \sum_k Err_k w_{jk}$; // compute the error with respect to the next higher layer, k
(15) **for** each weight w_{ij} in *network* {
(16) $\Delta w_{ij} = (l) Err_j O_i$; // weight increment
(17) $w_{ij} = w_{ij} + \Delta w_{ij}$; } // weight update
(18) **for** each bias θ_j in *network* {
(19) $\Delta \theta_j = (l) Err_j$; // bias increment
(20) $\theta_j = \theta_j + \Delta \theta_j$; } // bias update
(21) } }

k-Nearest-NeighborClassifier:

- Nearest-neighbor classifiersarebasedon learningbyanalogy, thatis, bycomparingagiventest tuplewith trainingtuples that aresimilar to it.
- The training tuples are describedby*n* attributes. Each tuple represents a point in an *n*-dimensional space. In this way,all of the training tuples are stored in an *n*-dimensionalpattern space. When given anunknown tuple, a ***k***-nearest-neighbor classifier searches thepattern space for the *k* trainingtuples that are closest to the unknown tuple. These *k* trainingtuplesarethe*k*nearestneighborsof theunknowntuple.
- Closenessisdefined interms ofadistancemetric,suchas Euclideandistance.
- TheEuclidean distancebetween two points ortuples, say,$\boldsymbol{X}_1=(x_{11}, x_{12}, \ldots, x_{1n})$ and $\boldsymbol{X}_2=(x_{21}, x_{22}, \ldots, x_{2n})$, is

$$dist(\boldsymbol{X_1}, \boldsymbol{X_2}) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

Inotherwords,foreach numericattribute,wetakethedifferencebetweenthecorrespondingvaluesofthat attributeintuple$\boldsymbol{X_1}$and intuple$\boldsymbol{X_2}$,squarethisdifference,and accumulateit. Thesquarerootistakenofthetotalaccumulateddistancecount.
Min-Maxnormalizationcanbeusedtotransformavalue$v$ofanumericattribute$A$to$v_0$intherange[0, 1]bycomputing

wheremin_Aandmax_Aaretheminimum andmaximumvalues ofattribute$A$

$$v' = \frac{v - min_A}{max_A - min_A},$$

- For *k*-nearest-neighborclassification,theunknowntupleisassignedthemostcommonclassamongits *k*nearestneighbors.
- When*k*=1,theunknowntupleisassignedtheclassofthetrainingtuplethatisclosesttoitin pattern space.
- Nearestneighborclassifiers can also be used for prediction, that is, to return a real-valuedpredictionforagiven unknown tuple.
- Inthiscase,theclassifierreturnstheaveragevalueofthereal-valuedlabelsassociatedwiththe*k*nearest neighbors of theunknowntuple.

3.5 OtherClassificationMethods:

3.5.1 GeneticAlgorithms:

Geneticalgorithmsattempttoincorporateideasofnaturalevolution.Ingeneral,geneticlearningstartsas follows.

- Aninitialpopulationiscreatedconsistingofrandomlygeneratedrules.Eachrulecanberepresentedbyastringofbits.Asasimpleexample,supposethatsamplesinagiven training set are described by two Boolean attributes, A1 and A2, and that there are twoclasses,C_1andC_2.
- Therule–IFA_1ANDNOTA_2THENC_2‖canbeencodedasthebitstring–100,‖where the two leftmost bits represent attributes A_1 and A_2, respectively, and the rightmost bitrepresentstheclass.
- Similarly,therule–IFNOTA_1ANDNOTA_2 THENC_1‖ can beencodedas–001.‖
- Ifanattributehaskvalues,wherek>2,thenkbitsmaybeusedtoencodetheattribute'svalues.

Classescanbe encodedinasimilarfashion.

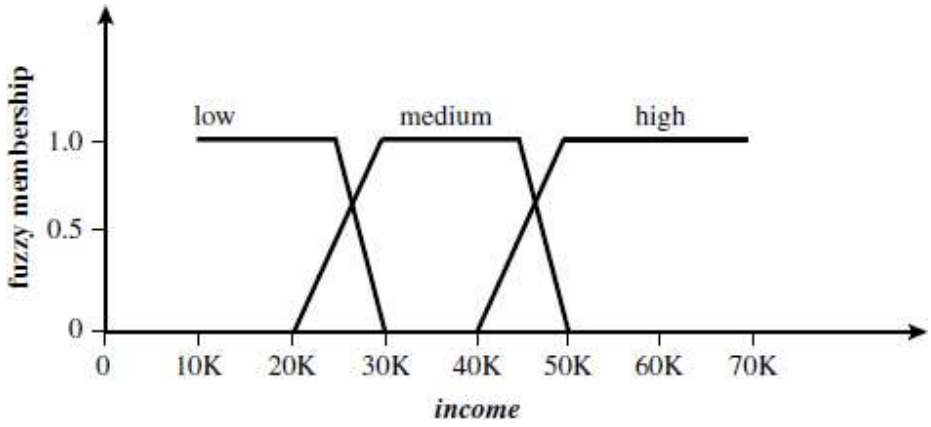
- Basedon thenotionofsurvival ofthefittest, anew population isformed to consistofthefittest rulesin thecurrent population, aswell asoffspringoftheserules.
- Typically,thefitnessofaruleisassessedbyitsclassificationaccuracyonasetoftrainingsamples.
- Offspringarecreated byapplyinggeneticoperatorssuchascrossoverandmutation.
- In crossover,substringsfrompairsofrules areswappedtoformnewpairsofrules.
- Inmutation,randomlyselected bits ina rule'sstringareinverted.
- The process of generating new populations based on prior populations of rules continuesuntilapopulation, P,evolveswhereeachrulein Psatisfies aprespecifiedfitnessthreshold.
- Genetic algorithms are easily parallelizable and have been used for classification aswellas other optimization problems.In data mining, theymaybeusedto evaluatethefitnessofotheralgorithms.

3.5.2 Fuzzy SetApproaches:

- Fuzzylogicusestruthvaluesbetween0.0and1.0torepresentthedegreeofmembershipthatacertain valuehas inagivencategory.Each categorythen represents afuzzyset.
- Fuzzylogicsystemstypicallyprovidegraphicaltoolstoassistusersinconvertingattributevalues to fuzzytruthvalues.
- Fuzzyset theoryis alsoknown as possibilitytheory.

- ItwasproposedbyLotfiZadehin1965asanalternativetotraditionaltwo-valuelogicandprobabilitytheory.
- Itletsusworkatahighlevelofabstractionandoffersameansfordealingwithimprecisemeasurement ofdata.
- Most important, fuzzysettheoryallows usto dealwith vague or inexact facts.
- Unlikethenotionoftraditional–crisp‖setswhereanelementeitherbelongstoaset*S*orits complement,in fuzzyset theory,elements canbelongtomorethan onefuzzyset.
- Fuzzyset theoryis usefulfordata miningsystems performingrule-basedclassification.● Itprovides operations for combiningfuzzymeasurements.
- Severalproceduresexistfortranslatingtheresultingfuzzyoutputintoa*defuzzified*orcrispvaluethat is returned bythesystem.
- Fuzzylogicsystemshavebeenusedinnumerousareasforclassification,includingmarket research,finance,healthcare,andenvironmental engineering.

Example:



3.6 RegressionAnalysis:

- Regressionanalysiscanbe usedtomodelthe relationshipbetweenoneormoreindependentorpredictorvariables and a dependentorresponsevariablewhich iscontinuous-valued.
- Inthecontextofdatamining,thepredictorvariablesaretheattributesofinterestdescribingthetuple (i.e., makingupthe attributevector).
- In general,thevaluesofthepredictorvariables areknown.
- Theresponsevariableiswhat wewanttopredict.

3.6.1 LinearRegression:

- Straight-lineregressionanalysisinvolvesaresponsevariable,*y*,and asinglepredictorvariablex.
- Itisthesimplestformof regression,andmodels*y* asalinear functionof*x*.

Thatis,y=b+wx
wherethe varianceof*y*isassumedtobe constant
*b*and*w* areregression coefficientsspecifyingthe Y-intercept and slopeoftheline.

- Theregressioncoefficients,*w*and*b*, canalsobethoughtof as weights,sothatwecanequivalentlywrite,$y=w_0+w_1x$
- Thesecoefficientscan be solved forbythemethod ofleast squares,whichestimatesthebest-fitting straight line as the one that minimizes the error between the actual data andtheestimateof theline.
- Let *D* be a training set consisting of values of predictor variable, *x*, for some population andtheir associated values for response variable, *y*. The training set contains |*D*| data points oftheform(x_1, y_1), (x_2, y_2),… , $(x_{|D|}, y_{|D|})$.

Theregressioncoefficientscanbeestimatedusingthismethodwiththefollowingequations:

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{|D|}(x_i-\bar{x})^2}$$

$$w_0 = \bar{y} - w_1\bar{x}$$

where*x* isthemeanvalueof $x_1,x_2,\ldots,x_{|D|}$,and*y*isthemeanvalueof$y_1,y_2,\ldots,y_{|D|}$.Thecoefficients $w_0$andw_1 oftenprovidegoodapproximationstootherwisecomplicatedregressionequations.

3.6.2 MultipleLinearRegression:

- Itisanextensionofstraight-lineregressionsoastoinvolvemorethanonepredictorvariable.
- Itallowsresponsevariableytobemodeledasalinearfunctionof,say,npredictorvariablesorattributes,A_1, A_2, …, An, describingatuple,X.
- Anexampleofamultiplelinearregressionmodelbasedontwopredictorattributesorvariables,$A_1$and A_2, is$y= w_0+w_1x_1+w_2x_2$

wherex_1 and x_2 arethevaluesofattributes A_1andA_2,respectively, in X.

- Multipleregressionproblemsareinsteadcommonlysolvedwiththeuseofstatisticalsoftwarepackages, suchas SAS,SPSS, and S-Plus.

3.6.3 NonlinearRegression:

- Itcan bemodeledbyaddingpolynomial termsto thebasic linearmodel.
- By applying transformations to the variables, we can convert the nonlinear model into alinearonethat can then besolved bythe method of least squares.
- Polynomial Regression is a special case of multiple regression. That is, the addition ofhigh-order terms like x^2, x^3, and so on, which are simple functions of the single variable, x,canbeconsideredequivalent to addingnewindependent variables.

Transformationofapolynomialregressionmodeltoalinearregressionmodel:
Consideracubicpolynomial relationshipgivenby

$$y=w_0+w_1x+w_2x_2+w_3x_3$$

Toconvertthisequationtolinearform,wedefinenewvariables:
$x_1 =x, x_2 =x^2 , x_3 =x^3$

It canthenbeconvertedtolinear formby applying theaboveassignments,resulting intheequation$y=w_0+w_1x+w_2x_2+w_3x_3$
whichis easilysolved bythemethod of least squaresusingsoftware for regressionanalysis.

3.7 ClassifierAccuracy:

- The accuracy of a classifier on a given test set is the percentage of test set tuples that arecorrectlyclassified bytheclassifier.
- In the pattern recognition literature, this is also referred to as the overall recognition rate oftheclassifier, that is, it reflects how well the classifier recognizes tuples of the variousclasses.
- The error rate or misclassification rate of a classifier,M, which is simply 1-Acc(M),whereAcc(M)is theaccuracyof M.
- The confusion matrix is a useful tool for analyzing how well your classifier can recognizetuplesof different classes.
- Truepositives refer to the positivetuples that werecorrectlylabeled bytheclassifier.● Truenegatives arethe negativetuplesthat werecorrectlylabeled bythe classifier.
- Falsepositivesarethenegativetuplesthatwereincorrectlylabeled.
- Howwelltheclassifiercanrecognize,forthissensitivityandspecificitymeasurescanbeused.

Accuracyis a function ofsensitivityand specificity.

$$accuracy = sensitivity\frac{pos}{(pos+neg)} + specificity\frac{neg}{(pos+neg)}.$$

$$sensitivity = \frac{t\_pos}{pos}$$

$$specificity = \frac{t\_neg}{neg}$$

$$precision = \frac{t\_pos}{(t\_pos + f\_pos)}$$

wheret_posisthenumberoftruepositivespos isthenumberof positive tuples
t _negis the number of true negativesnegis the number of negativetuples,f_posisthenumberoffalsepositives

Chapter-4

4.1 ClusterAnalysis:

- The process of grouping a set of physical or abstract objects into classes of similar objectsiscalled clustering.
- Aclusterisacollectionofdataobjectsthataresimilartooneanotherwithinthesameclusterand aredissimilarto theobjects in otherclusters.

- Aclusterofdataobjectscanbetreatedcollectivelyasonegroupandsomaybeconsideredasa form of data compression.
- Cluster analysis tools based on k-means, k-medoids, and severalmethods have also beenbuilt into many statisticalanalysis software packages or systems, such as S-Plus, SPSS, andSAS.

4.1.1 Applications:

- Cluster analysis has been widely used in numerous applications, including market research,patternrecognition, dataanalysis,andimageprocessing.
- In business, clustering can help marketers discover distinct groups in their customer basesandcharacterizecustomer groups based on purchasingpatterns.
- In biology, it can be used to derive plant and animal taxonomies, categorize genes withsimilarfunctionality, andgain insightinto structuresinherent inpopulations.
- Clustering may also help in the identification of areas of similar land use in an earthobservation database and in the identification of groups of houses in a city according tohousetype,value,andgeographiclocation,aswellastheidentificationofgroupsofautomobilein surancepolicyholders with ahigh averageclaim cost.
- Clusteringisalsocalled datasegmentationinsomeapplicationsbecause clusteringpartitionslargedata sets into groups accordingto their*similarity*.
- Clustering can also be used for outlier detection,Applications of outlier detection includethe detection of credit card fraud and the monitoring of criminal activities in electroniccommerce.

4.1.2 TypicalRequirementsOfClusteringInDataMining:

- **Scalability:**
  Many clustering algorithms work well on small data sets containing fewer than severalhundred data objects; however, a large database may contain millions of objects. Clusteringonasampleofagiven largedata set maylead to biased results.
  Highlyscalableclusteringalgorithmsareneeded.
- **Abilitytodealwithdifferenttypesof attributes:**
  Manyalgorithmsaredesignedtoclusterinterval-based(numerical)data.However,applicationsmayrequireclusteringothertypesofdata,suchasbi nary,categorical(nominal),and ordinal data, ormixtures of thesedata types.
- **Discoveryof clusterswitharbitraryshape:**
  Many clustering algorithms determine clusters based on Euclidean or Manhattan distancemeasures. Algorithms based on such distance measures tend to find spherical clusters withsimilarsizeand density.
  However, a cluster could be of any shape. It is important to develop algorithms thatcandetectclusters of arbitraryshape.
- **Minimalrequirementsfordomainknowledgetodetermineinputparameters:**
  Many clustering algorithms require users to input certain parameters in cluster analysis(suchasthenumberofdesiredclusters).Theclusteringresultscanbequitesensitivetoinput parameters.Parametersareoftendifficulttodetermine,especiallyfordatasetscontaininghigh-dimensionalobjects.Thisnotonlyburdensusers,butitalsomakesthequalityof clusteringdifficult to control.
- **Abilitytodealwithnoisydata:**

Mostreal-worlddatabasescontainoutliersormissing,unknown,orerroneousdata. Someclusteringalgorithmsaresensitivetosuchdataandmayleadtoclustersofpoorquality.

- **Incrementalclusteringandinsensitivitytotheorderofinputrecords:**
  Some clustering algorithms cannot incorporate newly inserted data (i.e., database updates)intoexistingclusteringstructuresand,instead,mustdetermineanew clusteringfromscratch.Someclusteringalgorithms aresensitiveto the order ofinputdata. Thatis,givenasetofdataobjects,suchanalgorithmmayreturndramaticallydifferentclusteringsde pendingonthe order ofpresentationof theinput objects. Itisimportanttodevelopincrementalclusteringalgorithmsandalgorithmsthatareinsensitiveto the order ofinput.
- **Highdimensionality:**
  Adatabaseoradatawarehousecancontainseveraldimensionsorattributes.Manyclustering algorithmsaregoodathandling low-dimensionaldata,involving only twotothree dimensions. Human eyes are good at judging the qualityof clustering for up to threedimensions.Findingclustersofdataobjectsinhighdimensionalspaceischallenging,especia llyconsideringthatsuch data can besparseand highlyskewed.
- **Constraint-basedclustering:**
  Real-world applications may need to perform clustering under various kinds of constraints.Suppose that your job is to choose the locations for a given number of new automaticbankingmachines(ATMs)inacity.Todecideuponthis,youmayclusterhouseholdswhi le considering constraints such as the city's rivers and highway networks, and the typeand number of customers per cluster. A challenging task is to find groups of data with goodclusteringbehavior that satisfyspecifiedconstraints.
- **Interpretabilityandusability:**
  Users expect clustering results to be interpretable, comprehensible, and usable. That is,clustering may need to be tied to specific semantic interpretations and applications. It isimportanttostudy howanapplicationgoalmayinfluencetheselectionofclusteringfeaturesandmethods.

4.2 MajorClusteringMethods:

- PartitioningMethods
- HierarchicalMethods
- Density-BasedMethods
- Grid-BasedMethods
- Model-BasedMethods

4.2.1 PartitioningMethods:

A partitioning method constructs *k* partitions of the data, where each partition represents acluster and *k* <= *n*. That is, it classifies the data into *k* groups, which together satisfy thefollowingrequirements:

- Eachgroupmustcontainatleastoneobject,and
- Eachobject mustbelongtoexactlyonegroup.

A partitioning method creates an initial partitioning. It then uses an iterative relocationtechnique that attempts to improve the partitioning by moving objects from one group toanother.

The general criterion of a good partitioning is that objects in the same cluster are close orrelatedtoeach other,whereas objectsof differentclustersarefarapart orverydifferent.

4.2.2 HierarchicalMethods:

A hierarchical method creates a hierarchical decomposition ofthe given set of data objects. Ahierarchical method can be classified as being eitheragglomerative or divisive, based onhowthehierarchical decomposition is formed.

- Theagglomerativeapproach,alsocalledthebottom-upapproach,startswitheachobjectforming a separate group. It successively merges the objects or groups that arecloseto one another, until all of the groups are merged into one or until a terminationcondition holds.
- The divisive approach, also calledthe top-down approach, starts with all of the objects inthe same cluster. In each successiveiteration, a cluster is split up into smaller clusters,untileventuallyeach objectis in onecluster,oruntil a termination conditionholds.

Hierarchical methods suffer fromthe fact that once a step (merge or split) is done,it can neverbe undone. This rigidity is useful in that it leads to smaller computationcosts by not havingtoworryabout acombinatorial number ofdifferentchoices.

Therearetwo approachestoimprovingthequalityofhierarchical clustering:

- Performcarefulanalysisofobject−linkages‖ateachhierarchicalpartitioning,suchasin Chameleon,or
- Integratehierarchicalagglomerationandotherapproachesbyfirstusingahierarchicalagglomerativealgorithmtogroupobjectsintomicroclusters,andthenperformingmacroclustering on the microclusters using another clustering method such asiterativerelocation.

4.2.3 Density-basedmethods:

- Most partitioning methods cluster objects based on the distance between objects. Suchmethods can find only spherical-shaped clusters and encounter difficulty at discoveringclusters ofarbitraryshapes.
- Other clustering methods have been developed based on the notion of density. Theirgeneral idea is to continue growing the given cluster as long as the density in theneighborhoodexceedssomethreshold;thatis,foreachdatapointwithinagivencluster, the neighborhood of a given radius has to contain at least a minimum number ofpoints. Such a method can be used to filter out noise (outliers)and discover clusters ofarbitraryshape.
- DBSCANanditsextension,OPTICS,aretypicaldensity-basedmethodsthatgrowclustersaccordingtoadensity-basedconnectivityanalysis.DENCLUEisamethodthat clusters objects based on the analysis of the value distributions of densityfunctions.

4.2.4 Grid-BasedMethods:

- Grid-based methods quantize the object space into a finite number of cells that form agridstructure.
- All of the clustering operations are performed on the grid structure i.e., on the quantizedspace.Themainadvantageofthisapproachisitsfastprocessingtime,whichistypically independent of the number of data objects and dependent only on the numberofcells in each dimension in thequantized space.
- STING is a typical example of a grid-based method. Wave Cluster applies wavelettransformationfor clusteringanalysisand isbothgrid-based anddensity-based.

4.2.5 Model-BasedMethods:

- ❖ Model-based methods hypothesize a model for each of the clusters and find the best fitofthe data to thegiven model.
- ❖ A model-based algorithm may locate clusters by constructing a density function thatreflectsthespatial distribution ofthe data points.
- ❖ It also leads to a way of automatically determining the number of clusters based onstandard statistics, taking ―noise‖or outliers into account and thusyielding robust clusteringmethods.

4.3 TasksinDataMining:

- ➢ ClusteringHigh-DimensionalData
- ➢ Constraint-BasedClustering

4.3.1 ClusteringHigh-DimensionalData:

- It is a particularly important task in cluster analysis because many applicationsrequiretheanalysisof objectscontaining alargenumberoffeatures ordimensions.
- Forexample,textdocumentsmaycontainthousandsoftermsorkeywordsasfeatures, and DNA micro array data may provide information on the expressionlevelsof thousands ofgenes under hundredsof conditions.
- Clusteringhigh-dimensionaldataischallengingduetothecurseofdimensionality.
- Manydimensionsmaynotberelevant.Asthenumberofdimensionsincreases, thedata become increasingly sparse so that the distance measurement between pairsofpointsbecomemeaninglessandtheaverage density of pointsanywhere inthedata islikely to be low. Therefore, a different clustering methodology needs to bedevelopedforhigh-dimensional data.
- CLIQUE and PROCLUS are two influential subspace clustering methods, whichsearchforclusters insubspaces ofthedata,ratherthanover theentiredataspace.
- Frequent pattern–based clustering,another clustering methodology, extractsdistinctfrequent patterns among subsets ofdimensions that occur frequently. It uses suchpatternsto group objectsandgeneratemeaningfulclusters.

4.3.2 Constraint-BasedClustering:

- It is a clustering approach that performs clustering by incorporation of user-specifiedorapplication-oriented constraints.
- Aconstraintexpressesauser'sexpectationordescribespropertiesofthedesiredclusteringresults,andprovidesaneffectivemeansforcommunicatingwiththeclusteringprocess.
- Various kinds of constraints can be specified, either by a user or as per applicationrequirements.
- Spatial clustering employs with the existence of obstacles and clustering under user-specifiedconstraints.Inaddition,semi-supervisedclusteringemploysforpairwiseconstraintsin order to improvethe qualityofthe resultingclustering.

4.4 ClassicalPartitioningMethods:

Themostwell-knownandcommonlyusedpartitioningmethodsare

- The*k*-MeansMethod
- k-MedoidsMethod

4.4.1 Centroid-BasedTechnique:The*K*-Means Method:

The k-means algorithm takes the input parameter, k, and partitions a set of n objects intokclusterssothattheresultingintraclustersimilarityishighbuttheinterclustersimilarityislow. Cluster similarity is measured in regard to the mean value of the objects in a cluster, whichcanbeviewed asthecluster'scentroid orcenter of gravity.

The*k*-meansalgorithmproceeds asfollows.

- First,itrandomlyselects*k*oftheobjects,eachofwhichinitiallyrepresentsaclustermeanor center.
- Foreachoftheremainingobjects,anobjectisassignedtotheclustertowhichitisthemostsimilar, based onthe distancebetween theobject and theclustermean.
- Itthencomputesthenewmeanforeach cluster.
- Thisprocessiteratesuntilthecriterionfunctionconverges.

Typically,thesquare-errorcriterionisused,definedas

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2,$$

whereEisthesumof thesquareerror forall objectsinthedata setpisthepoint in spacerepresentinga given object
$m_i$isthemean of clusterC_i

4.4.1 Thek-meanspartitioningalgorithm:

The*k*-meansalgorithmforpartitioning,whereeachcluster'scenterisrepresentedbythemeanvalueofthe objects in the cluster.

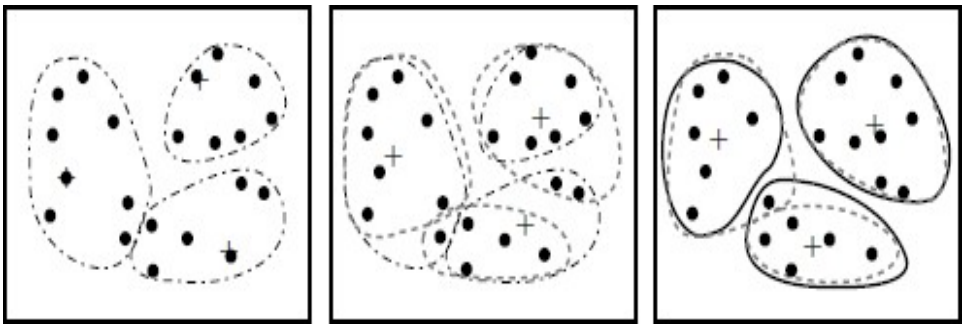
Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign 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, i.e., calculate the mean value of the objects for each cluster;
(5) **until** no change;



Clusteringof aset of objects based on the*k*-meansmethod

4.4.2 The*k*-MedoidsMethod:

- The k-means algorithm is sensitive to outliers because an object with an extremely largevaluemaysubstantiallydistortthedistributionofdata.Thiseffectisparticularlyexacerbatedd ueto theuseof the square-errorfunction.
- Instead of taking the mean value of the objects in a cluster as a reference point, we can pickactual objects to represent the clusters, using one representative object per cluster. Eachremainingobject isclustered withtherepresentativeobject towhich itis themost similar.
- Thepartitioning method is then performed based on the principle of minimizing the sum ofthe dissimilarities between each object and its corresponding reference point. That is, anabsolute-errorcriterion is used, definedas

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j|,$$

where*E*isthesumoftheabsoluteerror forallobjectsinthedata set
pisthepointinspacerepresentinga givenobjectinclusterC_j

$\mathbf{o_j}$istherepresentative objectof C_j

- The initial representative objects are chosen arbitrarily. The iterative process of replacingrepresentative objects by non representative objects continues as long as the quality of theresultingclusteringis improved.
- Thisqualityisestimatedusingacostfunctionthatmeasurestheaveragedissimilaritybetweenanobj ectand therepresentativeobject ofits cluster.
- To determine whether a non representative object, oj random, is a good replacement for acurrent representativeobject, oj, the following four cases are examined for each of thenonrepresentativeobjects.

Case1:
pcurrentlybelongstorepresentativeobject,o_j.If$o_j$isreplacedbyo_{random}asarepresentativeobjectandpisclo sest to oneoftheother representativeobjects,o_i,i≠j, thenpis reassignedtoo_i.
Case2:
pcurrently belongstorepresentativeobject,oj.Ifojisreplacedbyo_{random}asarepresentativeobjectand pis

closest to o_{random},then p is reassigned to o_{random}.

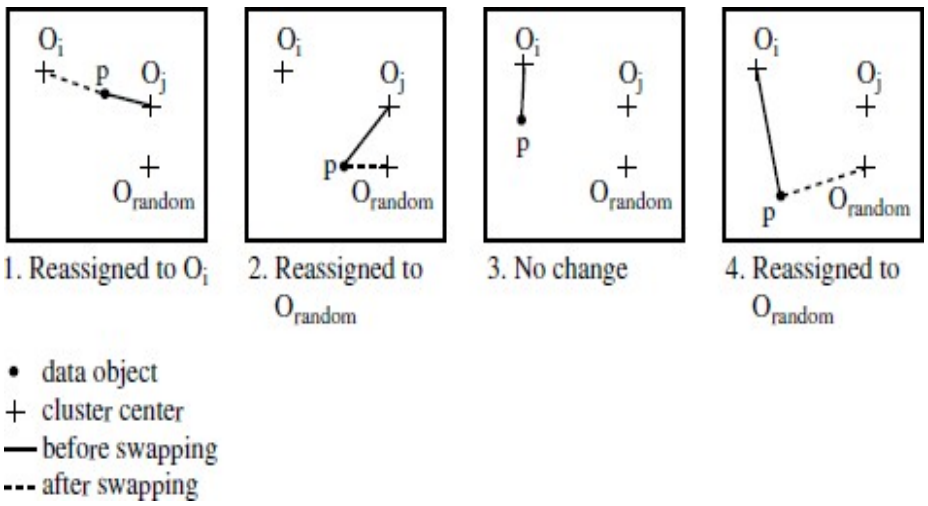
Case3:

pcurrentlybelongstorepresentativeobject,o_i,i≠j.Ifojisreplacedbyo_{random}asarepresentativeobjectandp is stillclosest too_i, then theassignment does notchange.

Case4:

pcurrentlybelongstorepresentativeobject,oi,i≠j.If$o_j$isreplacedbyo_{random}asarepresentativeobjectandp is closest to o_{random}, thenp is reassigned

too_{random}.



Fourcasesofthe costfunctionfor *k*-medoidsclustering

4.4.2The*k*-MedoidsAlgorithm:

Thek-medoidsalgorithmforpartitioningbasedonmedoidorcentralobjects.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, o_{random};
(5) compute the total cost, *S*, of swapping representative object, o_j, with o_{random};
(6) **if** $S < 0$ **then** swap o_j with o_{random} to form the new set of *k* representative objects;
(7) **until** no change;

The *k*-medoids method is more robust than *k*-means in the presence of noise and outliers,because a medoid is less influenced by outliers or other extreme values than a mean. However,itsprocessingis morecostlythan the*k*-means method.

4.5 HierarchicalClusteringMethods:

- Ahierarchicalclusteringmethod worksbygroupingdataobjectsintoa treeofclusters.
- Thequalityofapurehierarchicalclustering methodsuffersfrom itsinabilitytoperform adjustment once a merge or split decision has been executed. That is, if a particularmerge or split decision later turns out to have been a poor choice, the method cannotbacktrackand correct it.

Hierarchical clustering methods can be further classified as either agglomerative or divisive,depending on whether the hierarchical decomposition is formed in a bottom-up or top-downfashion.

4.5.1 Agglomerativehierarchicalclustering:

- This bottom-up strategy starts by placing each object in its own cluster and then mergesthese atomic clusters into larger and larger clusters, until all of the objects are in a singleclusteroruntil certain termination conditions aresatisfied.
- Most hierarchical clustering methods belong to this category. They differ only in theirdefinitionofinterclustersimilarity.

4.5.2 Divisivehierarchicalclustering:

- Thistop-downstrategydoesthereverseofagglomerative hierarchicalclusteringbystartingwith all objects in onecluster.
- It subdivides the cluster into smaller and smaller pieces, until each object forms a clusteron its own or until it satisfies certain termination conditions, such as a desired numberof clustersis obtained orthediameterofeach cluster is within acertain threshold.

4.6 Constraint-BasedClusterAnalysis:

Constraint-based clustering finds clusters that satisfy user-specified preferences or constraints.Depending on the nature of the constraints, constraint-based clustering may adopt rather differentapproaches.

Thereareafewcategoriesofconstraints.

- **Constraintsonindividualobjects:**
  We can specify constraints on the objects to be clustered. In a real estate application, forexample, one may like to spatially cluster only those luxury mansions worth over a milliondollars.Thisconstraintconfinesthesetofobjectstobeclustered.Itcaneasilybehandledbyp reprocessingafterwhichtheproblemreducestoaninstanceof unconstrainedclustering.

- **Constraintsontheselectionof clusteringparameters:**
  A user may like to set a desired range for each clustering parameter. Clustering parametersare usually quite specific to the given clustering algorithm. Examples of parameters includek, the desired number of clusters in a k-means algorithm; or e the radius andthe minimumnumber of points in the DBSCAN algorithm. Although such user-specified parameters maystrongly influence the clustering results, they are usually confined

to the algorithm itself.hus, their fine tuning and processing are usually not considered a form of constraint-basedclustering.

- **Constraintsondistanceorsimilarityfunctions:**

  We can specify different distance or similarity functions for specific attributes of the objectsto be clustered, or different distance measures for specific pairs of objects. When clusteringsportsmen, for example, we may use different weighting schemes for height, body weight,age, and skill level. Although this will likely change the mining results, it may not alter theclustering process per se. However, in some cases, such changes may make the evaluation ofthe distance function nontrivial, especially when it is tightly intertwined with the clusteringprocess.

- **User-specifiedconstraintsonthepropertiesofindividualclusters:**
  Auser may liketo specify desiredcharacteristicsoftheresulting clusters,whichmaystronglyinfluence the clusteringprocess.
- **Semi-supervisedclusteringbasedonpartialsupervision:**
  The quality of unsupervised clustering can be significantly improved using some weak formof supervision. This may be in the form of pairwise constraints (i.e., pairs of objects labeledas belonging to the same or different cluster). Such a constrained clustering process iscalled semi-supervisedclustering.

4.7 OutlierAnalysis:

- There exist data objects that do not comply with the general behavior or model of the data.Suchdataobjects,whicharegrosslydifferentfromorinconsistentwiththeremainingsetofdata, arecalled outliers.
- Many data mining algorithms try to minimize the influence of outliers or eliminate them alltogether. This, however, could result in the loss of important hidden information becauseone person's noise could be another person's signal. In other words, the outliers may be ofparticular interest, such as in the case of fraud detection, where outliers may indicatefraudulent activity. Thus, outlier detection and analysis is an interesting data mining task,referredtoas outlier mining.
- It can be used in fraud detection, for example, by detecting unusual usage of credit cards ortelecommunicationservices.Inaddition,itisusefulincustomizedmarketingforidentifying the spending behavior of customers with extremely low or extremely highincomes,orinmedical analysisfor findingunusualresponsestovariousmedicaltreatments.

Outlier mining can be described as follows: Given a set of *n* data points or objects and*k*, theexpectednumberofoutliers,findthetop*k*objectsthatareconsiderably dissimilar,exceptional, or inconsistent with respect to the remaining data. The outlier mining problemcanbeviewed as two subproblems:

- Definewhatdatacanbe considered asinconsistentinagivendataset,andFindan efficient methodto mine theoutliersso defined.
-

Typesofoutlierdetection:

- StatisticalDistribution-BasedOutlierDetection
- Distance-BasedOutlierDetection
- Density-BasedLocalOutlierDetection
- Deviation-BasedOutlierDetection

4.7.1 StatisticalDistribution-BasedOutlierDetection:

The statistical distribution-based approach to outlier detection assumes a distribution orprobability model for the given data set (e.g., a normal or Poisson distribution) and thenidentifies outliers with respect to the model using a discordancy test. Application of thetest requires knowledge of the data set parametersknowledge of distribution parameterssuchas the meanand varianceand the expected numberof outliers.

Astatisticaldiscordancytestexaminestwohypotheses:●

Aworkinghypothesis

- Analternativehypothesis

Aworkinghypothesis,H,isastatementthattheentiredatasetofnobjectscomesfromaninitial distribution model, F, that is,

$$H : o_i \in F, \quad \text{where } i = 1, 2, \ldots, n.$$

The hypothesis is retained if there is no statistically significant evidence supporting itsrejection. A discordancy testverifies whether an object,oi,issignificantly large (orsmall) in relation to the distribution F. Different test statistics have been proposed for useas a discordancy test, depending on the available knowledge of the data. Assuming thatsomestatistic,T,hasbeenchosenfordiscordancytesting,andthevalueofthestatisticfor object oi is vi, then the distribution of T is constructed. Significance probability,SP(vi)=Prob(T > vi), is evaluated. If SP(vi) is sufficiently small, then oi is discordant andtheworkinghypothesis is rejected.

An alternative hypothesis, H, which states that o_i comes from another distribution model,G, is adopted. The result is very much dependent on which model F is chosen because$o_i$maybeanoutlierunderonemodelandaperfectlyvalidvalueunderanother.The alternativedistributionisveryimportantindeterminingthepowerofthetest,thatis,theprobabilit ythat the workinghypothesis is rejectedwhen oi is reallyan outlier.

Therearedifferentkindsofalternativedistributions.

- **Inherentalternativedistribution:**
  In this case, the working hypothesis that all of the objects come from distribution F isrejected in favor of the alternative hypothesis that all of the objects arise from anotherdistribution, G:
  H:oi € G,wherei=1, 2,…, n
  FandGmaybedifferentdistributionsordifferonlyinparametersofthesamedistribution.
  There are constraints on the form of the *G* distribution in that it must have potential toproduce outliers. For example, it may have a different mean or dispersion, or a longertail.
- **Mixturealternativedistribution:**
  The mixture alternative states that discordant values are not outliers in the F population,butcontaminants from some other population,

G. Inthiscase,thealternativehypothesisis

$$\overline{H} : o_i \in (1-\lambda)F + \lambda G, \quad \text{where } i = 1, 2, \ldots, n.$$

Slippagealternativedistribution:

This alternative states that all of the objects (apart from some prescribed small number)arise independently from the initial model, F, with its given parameters, whereas theremaining objects are independent observations from a modified version of F in whichtheparameters havebeenshifted.

Therearetwobasictypesofproceduresfordetectingoutliers:

Blockprocedures:

Inthiscase,eitherallofthesuspectobjectsaretreatedasoutliers orallofthemareacceptedasconsistent.

Consecutiveprocedures:

Anexampleofsuchaprocedureistheinsideoutprocedure.Itsmainideaisthattheobjectthatisleastlikelytobeanoutlieristestedfirst.Ifitisfoundtobeanoutlier,thenallofthe more extreme values are also considered outliers; otherwise, the next most extreme object istested,and so on.This proceduretends to bemore effectivethan blockprocedures.

4.7.2 Distance-BasedOutlierDetection:

The notion of distance-based outliers was introduced to counter the main limitations imposedby statistical methods. An object, o, in a data set, D, is a distance-based (DB)outlier withparameters pct and dmin,that is, a DB(pct;dmin)-outlier, if at least a fraction,pct, of theobjects in D lie at a distance greater than dmin from o. In other words, rather that relying onstatistical tests, we can think of distance-based outliers as those objects that do not haveenough neighbors, where neighbors are defined based on distance from the given object. Incomparison with statistical-based methods, distance-based outlier detection generalizes theideas behind discordancy testing for various standard distributions. Distance-based outlierdetection avoids the excessive computation that can be associated with fitting the observeddistributioninto some standard distribution and in selectingdiscordancytests.

For many discordancy tests, it can be shown that if an object, o, is an outlier according to thegiventest, theno is alsoaDB(pct, dmin)-outlierforsomesuitablydefinedpct anddmin.

Forexample,if objectsthatlie threeormorestandarddeviations fromthe mean areconsideredtobe outliers,assuminganormaldistribution, thenthisdefinitioncanbegeneralized bya DB(0.9988, 0.13s) outlier.

Severalefficientalgorithmsforminingdistance-basedoutliershavebeendeveloped.

Index-basedalgorithm:

Given a data set, the index-based algorithm uses multidimensional indexing structures, suchas R-trees or k-d trees, to search for neighbors of each object***o*** within radius *dmin*around thatobject. Let *M*be the maximum number of objects within the *dmin*-neighborhood of an outlier.Therefore, once*M*+1 neighbors of object ***o*** are found, it is clear that ***o*** is not an outlier. Thisalgorithm has a worst-case complexity of $O(n2^k)$, where n is the number of objects in the datasetand$k$isthe dimensionality.Theindex-basedalgorithmscaleswellas$k$increases.However, this complexity evaluation takes only the search time into account, even though thetask ofbuildingan indexin itself can becomputationallyintensive.

Nested-loopalgorithm:

Thenested-loopalgorithmhasthesamecomputationalcomplexity astheindex-basedalgorithm but avoids index structure construction and tries to minimize the number of I/Os. Itdivides the memory buffer space into two halves and the data set into several logical blocks.By carefully choosing the order in which blocks are loaded into each half, I/O efficiency canbeachieved.

Cell-basedalgorithm:

To avoid$O(n^2)$ computational complexity, a cell-based algorithm was developed for memory-resident data sets. Its complexity is $O(c^k+n)$, where c is a constant depending on the numberofcells andk is thedimensionality.

Inthismethod,thedata spaceispartitioned intocellswithasidelengthequalto $\frac{dmin}{2\sqrt{k}}$ Each cellhastwolayerssurroundingit.Thefirstlayerisonecellthick,whilethesecondis $\lceil 2\sqrt{k}-1 \rceil$cells thick, rounded up to the closest integer. The algorithm counts outliers on acell-by-cellratherthananobject-by-objectbasis.Foragivencell,it accumulatesthreecounts—the number of objectsinthe cell, inthe celland the first layer together, and inthecell and both layers together. Let's refer to these counts as cell count, cell + 1 layer count, andcell+2 layers count, respectively.

Let Mbe the maximum number ofoutliers that can exist in the dmin-neighborhood of anoutlier.

- An object, **o**, in the current cell is considered an outlier only if cell + 1 layer countis lessthan or equal to M. If this condition does not hold, then all of the objectsin the cell can beremovedfromfurtherinvestigation as theycannotbeoutliers.
- If cell_+ 2_layers_count is less than or equal to M, then all of the objects in thecell areconsidered outliers. Otherwise, if this number is more than M, then itis possible that someoftheobjectsinthecellmaybeoutliers.Todetecttheseoutliers,object-by-objectprocessing is used where, for each object,**o**, inthe cell,objects in the second layer of**o**are examined. For objects in the cell, only thoseobjects having no more than M points intheirdmin-neighborhoodsareoutliers.Thedmin-neighborhoodofanobjectconsistsoftheobject's cell, all ofits firstlayer, and someofits second layer.

A variation to the algorithm is linear with respect to n and guarantees that no morethan threepasses over the data set are required. It can be used for large disk-residentdata sets, yet doesnotscale well forhigh dimensions.

4.7.3 Density-BasedLocalOutlierDetection:

Statisticalanddistance-basedoutlierdetectionbothdependontheoverallorglobal distributionofthegivensetofdatapoints,D.However,dataareusuallynotuniformlydistributed. These methods encounter difficulties when analyzing data with ratherdifferent densitydistributions.

To define the local outlier factor of an object, we need to introduce the concepts ofk-distance,k-distanceneighborhood, reachabilitydistance,13and localreachabilitydensity. Thesearedefinedas follows:

Thek-distanceofanobjectpisthemaximaldistancethatpgetsfromitsk-nearestneighbors.

Thisdistance isdenotedask-distance(p).Itisdefinedasthedistance,d(p,o),betweenpandanobjecto2D,suchthatforatleastkobjects,$o_0$2D,itholdsthatd(p,o·)_d(p,o).Thatis,there areatleastkobjectsinDthatareasclose

asorcloser topthano,and forat most k-1 objects, o002 D, it holds that d(p;o'')<d(p, o).

That is, there are at most k-1 objects that are closer to p than o. You may be wondering at thispoint how k is determined. The LOF method links to density-basedclustering in that it sets kto the parameter rMinPts,which specifies the minimumnumberof points for use in identifyingclustersbased on density.

Here,MinPts(as k)is usedto definethelocalneighborhood ofanobject, p.

The k-distance neighborhood of an object p is denoted $N_{kdistance(p)}(p)$, or $N_k(p)$for short. Bysetting k to MinPts, we get $N_{MinPts}(p)$. It contains the MinPts-nearestneighbors of p. That is, itcontainseveryobject whose distanceis not greater than theMinPts-distanceof p.

The reachability distance of an object p with respect to object o (where o is withintheMinPts-nearestneighbors ofp), is definedas reach distMinPts(p,o)=max{MinPtsdistance(o),d(p,o)}.

Intuitively, if an object p is far away , then the reachability distance between the two is simplytheiractualdistance.However,iftheyaresufficiently close(i.e.,wherepiswithintheMinPts-distanceneighborhoodofo),then theactualdistanceisreplacedby theMinPts-distance of o. This helps to significantly reduce the statistical fluctuations of d(p, o) for all ofthep close to o.

The higher the value of MinPts is, the more similar is the reachability distance for objectswithin thesame neighborhood.

Intuitively, the local reachability density of p is the inverse of the average reachabilitydensitybased on the MinPts-nearestneighbors ofp.It isdefined as

The local outlier factor (LOF) of ***p*** captures the degree to which we call ***p*** an

$$lrd_{MinPts}(\boldsymbol{p}) = \frac{|N_{MinPts}(\boldsymbol{p})|}{\sum_{\boldsymbol{o}\in N_{MinPts}(\boldsymbol{p})} reach\_dist_{MinPts}(\boldsymbol{p},\boldsymbol{o})}.$$

outlier.Itis defined as

$$LOF_{MinPts}(\boldsymbol{p}) = \frac{\sum_{\boldsymbol{o}\in N_{MinPts}(\boldsymbol{p})} \frac{lrd_{MinPts}(\boldsymbol{o})}{lrd_{MinPts}(\boldsymbol{p})}}{|N_{MinPts}(\boldsymbol{p})|}.$$

It is the average of the ratio of the local reachability density of ***p*** and those of ***p***'s*MinPts*-nearest neighbors. It is easy to see that the lower ***p***'s local reachability densityis, and the higher the local reachability density of ***p***'s *MinPts*-nearest neighbors are,thehigher*LOF*(***p***) is.

4.7.4 Deviation-BasedOutlierDetection:

Deviation-based outlier detection does not use statistical tests or distance-based measures toidentifyexceptionalobjects.Instead,itidentifiesoutliersbyexaminingthe maincharacteristicsofobjectsinagroup. Objectsthat–deviate‖fromthisdescriptionareconsidered outliers. Hence, in this approach the term deviations is typically used to referto outliers. Inthissection,westudytwotechniquesfordeviation-basedoutlierdetection.Thefirstsequentiallycomparesobjectsinaset,whilethesecondemploysanOLAPdatacubeapproach.

SequentialExceptionTechnique:

Thesequentialexceptiontechniquesimulatesthewayinwhichhumanscandistinguish unusual

objects from among a series of supposedly like objects. It uses implicitredundancy of the data. Given a data set, D, of n objects, it builds a sequence of subsets,{D1,D2,…,Dm}, of theseobjects with2<=m <=n such that

$$D_{j-1} \subset D_j, \quad \text{where } D_j \subseteq D.$$

Dissimilarities are assessed between subsets in the sequence. The technique introduces thefollowingkeyterms.

Exceptionset:

This is the set of deviations or outliers. It is defined as the smallest subset of objects whoseremovalresults in the greatest reduction ofdissimilarityinthe residual set.

Dissimilarityfunction:

This function does not require a metric distance between the objects. It is any function that, ifgiven a set of objects, returns a low value if the objects are similar to one another. The greaterthedissimilarityamongtheobjects,thehigher thevaluereturnedbythefunction.Thedissimilarity of a subset is incrementally computed based on the subset prior to it in thesequence. Given a subset of *n* numbers, $\{x_1, \ldots, x_n\}$, a possible dissimilarity function is thevarianceof thenumbers in theset, that is,

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where x is the mean of the n numbers in the set. For character strings, the dissimilarity functionmay be in the form of a pattern string (e.g., containing wildcard characters that is used to coverall of the patterns seen so far. The dissimilarity increases when the pattern covering all of thestrings in D_{j-1}does not cover anystringin D_j that isnot in D_{j-1}.

Cardinalityfunction:

Thisis typicallythecount of thenumberof objects inagiven set.

Smoothingfactor:

Thisfunctioniscomputedforeachsubsetinthesequence.It assesseshowmuchthedissimilaritycan bereduced byremovingthe subset from the original set ofobjects.