Chapter-3

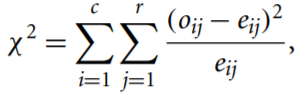
Data Preprocessing

* 1. Overview:
     1. Makes mining efficient.
     2. Cleaning + Integration + Reduction + Transformation.
     3. Data Quality = Accuracy + Completeness + Consistency.
     4. Disguised missing data like bday = jan1.
     5. Timeliness(eg. Sales record submitted at month end) + Believability(by user) + Interpretability(understood by user).
  2. Major Tasks in Data Preprocessing:
     1. Data cleaning:
        1. Fill missing values, remove noise, outliers, reduce inconsistency.
     2. Data integration:
        1. Integrating might lead to inconsistency bcz different attribute names. Or same value with different value used like first name, last name in different sets
     3. Data reduction:
        1. Dimensionality reduction:
           1. Data encoding schema to get reduced, compressed data
           2. Data compression(Wavelet transform, PCA)
           3. Attribute subset selection
           4. Attribute construction
        2. Numerosity reduction
           1. Data replaced by smaller representations:
           2. Parametric(regression, log linear)
           3. Non parametric models(histogram, cluster, sampling, data agression).
     4. Data transformation:
        1. Normalization, discretization, concept hierarchy generalisation.
  3. Data cleaning:
     1. Missing values
        1. Ignoring the tuple
        2. Full missing Val manually.
        3. Use global constant.
           1. Might have problems like all tuples have multiple similar values for similarity.
        4. Measure of Central tendency
           1. Mean for symmetric
           2. Median for skewed.
        5. Measure of Central tendency for a class.
        6. Most probable value
        7. Methods 3-6 bias the data
        8. Each attribute should have its own method.
        9. Null values can be kept
     2. Noisy data(Outliers)
        1. Random error or variance.
        2. Binning:
           1. Smooth sorted data by consulting its neighbourhood
           2. Eg. first partition in equal bins
           3. Then smoothing by bin means replacing all values with mean.
           4. OR smoothing by bin boundaries means replacing values close to boundaries.
           5. Also used as discretization technique.
        3. Regression
           1. Using best line to fit 2 attributes
        4. Outlier analysis
           1. Clustering
        5. Some NN has built in smoothing
     3. Data cleaning as a process:
        1. Discrepancy detection
        2. Metadata used (data collected while getting to know your data chapter 2).
        3. Field overloading
        4. Unique rule, consecutive rule, null rule.
        5. Tools:
           1. Data scrubbing
           2. Data auditing tools
           3. Data migration tools
           4. ETL(extraction, transformation, loading).
        6. Data discrepancy, data transformation iterates
        7. Eg. Potter’s Wheel
        8. SQL used for increased interactivity with data cleaning.
  4. Data integration
     + 1. Merging from multiple sources→ reduces redundancy, inconsistency
       2. Entity identification problem
          1. Finding out same attributes.
          2. Metadata like name,type,null etc can be used.
          3. Functional dependency & referential constraints must be taken care of.
       3. Redundancy & correlation analysis
          1. Derived attributes and attribute naming → can cause Redundancy.
          2. Some redundancy can be detected by Correlation analysis.

Saying how strongly 1 attribute implies other.

X^2(chi-square) for NOMINAL attribute.

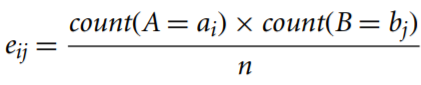
Contingency table created having r rows, c columns where c is no of values of 1 attribute and r is no of values for other attribute.



i.e. over all values of the table created

O\_ij is no of tuples having both a\_i,b\_j

E\_ij is no of tuples having a\_i or b\_j.

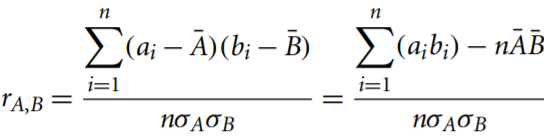


Chi^2 finds out if they are independent or not. (r-1)\*(c-1) degree of freedom.

There is a critical chi^2 value for different degree of freedom, if the X^2 is above it then the hypothesis that they are independent is rejected. Which means they are correlated.

Correlation coefficient & Covariance for NUMERIC attribute.

Correlation coefficient==Pearson’s product moment coefficient.

 where a\_i is value and A\_bar is mean of all values of that attribute a similarly for b, n is no of tuples, sigma is Standard deviation.

-1<=r(A,B)<=1

-1 means negatively correlated, +1 means positively, 0 means not correlated.

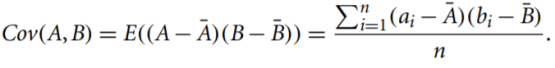
Higher val implies A(or B) can be removed.

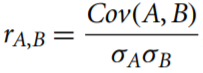
Correlation can also be cal from scatter plot.

Correlation doesn’t infer casualty like A causes B.

Covariance of NUMERIC attribute.

Correlation, covariance both says how much 2 attributes change together.

where E is the expectation.



 if for independent attributes E(A.B) = E(A).E(B) therefor cov will be 0. Vice versa is not true i.e. may have cov 0 but are not independent.

Cov positive→ if A larger than A\_bar then B is also likely to be larger than B\_bar. Otherwise negatively

Variance is a special case of covariance, where the two attributes are identical.

* + - 1. Tuple duplication
         1. Use of denormalized table → also source of redundancy.
         2. Inconsistencies often arise between various duplicates, due to inaccurate data entry or updating some but not all data occurrences.
      2. Data value conflict detection & resolution
         1. Data integration also involves the detection and resolution of data value conflicts.
         2. Eg. attribute values from different sources may differ like weight in kg in one and pounds in other.
         3. Attributes may also differ on the abstraction level, eg. one attribute is for one branch while in one the same attribute is for one school.
         4. Data discrepancy further discussed in Data cleaning as a process.
  1. Data reduction:
     1. Reduced data should produce same result s of huge data.
     2. Overview of Data Reduction Strategies :
        1. Dimensionality reduction, Numerosity reduction, data compression.
        2. Dimensionality reduction:
           1. Reduce no of Random variables or attributes.
           2. Wavelet transform, PCA

Transforms original data into smaller space.

* + - * 1. Attribute subset selection:

Removes irrelevant, weekly relevant or redundant attributes.

* + - 1. Numerosity reduction:
         1. replace the original data volume by alternative, smaller forms of data representation.
         2. Parametric:

Regression, Log-linear models.

* + - * 1. Non Parametric:

Histogram, clustering, sampling, data cube aggregation.

* + - 1. Data compression:
         1. Lossless, lossy. Above two reduction also comes under compression.
      2. The computational time spent on data reduction should not outweigh or “erase” the time saved by mining on a reduced data set size.
    1. Wavelet Transform:
       1. Discrete WT transforms a vector to another with same dimensions with different values so that new one can be truncated
       2. Compressed info of data can be retained by storing only a small fraction of the strongest of the wavelet coefficients.
       3. Some coefficients above user-specified threshold are retained rest are made 0. Hence data→ sparse.
       4. DWT is better lossy than DFT(fourier). I.e. with same no of coefficients.
       5. Only 1 DFT while there is families of DWT like Haar-2 etc.
       6. Iterative:(hierarchical pyramid algo)
          1. L length of data vector must be int power of 2, if not then pad.
          2. Each transform→ data smoothing (weighted avg or sum) + weighted difference(brings detailed features).
          3. Above functions are done at pairs x(2i), x(2i+1).
          4. Lengths after function applying → L/2.
          5. The two functions are recursively applied to the data sets obtained in the previous loop, until the resulting data sets obtained are of length 2
          6. Selected values from the data sets obtained in the previous iterations are designated the wavelet coefficients of the transformed data.
       7. Other method: Matrix multiplication::[will not cover].
    2. Principal Component Analysis(PCA):
       1. Normalize input data.
       2. Computes k orthonormal vectors that provide a basis for the normalized input data→ they are unit vectors(pointing in a direction perpendicular to the others PCs→input data are linear combination of the PCs.
       3. PCs→sorted in order of decreasing “significance” or strength→serve as a new set of axes for the data(sorted like 1st has highest variance in data)
       4. Weeker elements should be avoided (lowest variances).
       5. Adv:
          1. PCs can be used for regression and cluster analysis.
          2. PCA better for sparse data while WT better for high dimensionality data.
    3. Attribute Subset Selection:
       1. Ignoring irrelevant and redundant attributes (but which for which type of pattern mining requires domain expertise).
       2. Find out minimal set of attributes s.t. resulting probability distribution of the data classes is as close as possible to the original distribution.
       3. 2^n subsets are possible. We wll have reduced search space then apply greedy .
       4. Attribute eval measures like Information Gain in DT etc are used.
       5. Types:
          1. Stepwise forward selection: empty set of attributes, best attributes → determined and added.
          2. Stepwise backward elimination full set of attributes, removes the worst attribute.
          3. Combination of forward selection and backward elimination: selects the best attribute and removes the worst
          4. Decision tree induction:

Non-leaf node→ test on attribute

Leaf node→ class predicted

branch→ outcome of a stage

For attribute subset selection→ consider a data and construct DT whichever attributes dont come are irrelevant.

* + - 1. Attribute construction like area from height, width.
    1. Regression and Log-Linear Models: Parametric Data Reduction:
       1. Linear Regression:
          1. y=w\*x+b (y is response variable, x is predictor, w & b are regression coefficients).
          2. Y’s variance→ assumed to be constant.
          3. To find out w,b→ method of least squares(minimizes error between actual and predicted line).
          4. Multiple linear regression→ having 2 or more predictors.
       2. Log-linear model.
          1. approximate discrete multidimensional probability distributions.
          2. Finds out probability of a data point represented by n-attributes in a smaller attribute space.
          3. Also dimensionality reduction.
          4. Also does data smoothing since aggregate estimates are less subjective to sampling variations in smaller dimension.
          5. Can be for sparse data
       3. Regression good for skewed data.
       4. Log linear good scalability for high dimension like 10 or so
    2. .Histogram
       1. Use binning to approximate data distribution. Good data reduction
       2. Partitions data into disjoint subsets→ named as buckets/bins.
       3. If bucket represent single attribute’s (value/freq) pair→ singleton bucket.
       4. Values are combined in some range→ equal width histogram.
       5. Partitioning rules:
          1. Equal-width.
          2. Equal-freq: buckets are created so that, roughly, the frequency of each bucket is constant.
       6. Good with sparse/skewed uniform data.
       7. Multidimensional histogram: multiple attributes.
    3. Clustering:
       1. Clusters of objects: with similarity with each other and dissimilar with other cluster’s objects.
       2. similarity→ how close objects are(distance).
       3. Quality of cluster→ diameter of cluster, centroid of cluster(avg dist of objects from centroid cluster).
       4. Data reduction uses these clusters instead of data.
    4. Sampling:
       1. Large dataset to small dataset
       2. Ways to sample for m samples
          1. Simple random sample without replacement (SRSWOR) of size s:

Every tuple has 1/m probability of getting picked.

* + - * 1. Simple random sample with replacement (SRSWR) of size s:

Same as SRSWOR except the drawn sample is placed back and hence can be picked up again.

* + - * 1. Cluster sample:

Sampling of clusters.

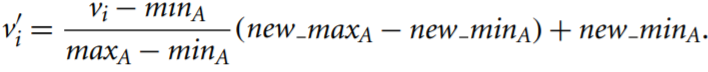
SRSWOR etc can be applied on clusters.

* + - * 1. Stratified sample:

D data divided into mutually disjoint parts called strata. Eg. age groups

* + - 1. Advantage of Sampling:
         1. Cost of obtaining a sample→ proportional to sample size instead of data size.
         2. Hence, sampling complexity is sublinear to data size.
         3. Other strategies require atleast 1 pass through whole D.
         4. While for a fixed sample size, sampling complexity increases with dimensions.
         5. It is possible (using the central limit theorem) to determine a sufficient sample size for estimating a given function within a specified degree of error.
    1. Data cube aggregation:
       1. Combining different data to smaller dataset eg sales per quarter to per year as per analysis task.
       2. Data aggregated without info loss.
       3. Data cube stores multidimensional aggregated info.
       4. Concept hierarchy may exist for each attribute.
       5. Base cuboid:
          1. Cube created at lowest abstraction leve.
       6. Apex cuboid:
          1. Cube at highest level of abstraction.
  1. Data Transformation and Data Discretization:
     1. transformed/consolidated→ so that resulting mining process may be efficient, and the patterns found may be easier to understand.
     2. Data Transformation Strategies Overview:
        1. Smoothing:
           1. Remove noise
           2. By binning, regression and clustering.
        2. Attribute construction:
           1. New attributes added.
           2. To improve mining.
        3. Aggregation:
           1. Data aggregated.
           2. Used to create data cube at different levels of abstractions.
        4. Normalization:
           1. Attribute data are scaled to fall under same range
        5. Discretization:
           1. Raw values of numeric data are replaced by intervals/or conceptual labels.
           2. Concept hierarchy can be done in here also.
           3. supervised/unsupervised
        6. Concept hierarchy generation for nominal data:
           1. Level hierarchy.
     3. Data Transformation by Normalization:
        1. normalized/standardized(like kg and grams etc).
        2. attempts to give all attributes an equal weight.
        3. Normalization→ for classification algorithms involving NN or distance measurements such as nearest-neighbor classification and clustering.
        4. Types:
           1. Min-Max Normalisation:

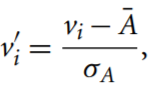
Maps to new range[min\_new\_A, max\_new\_A]



Preserves relation among attribute’s other values.

Gives “out of bounds” error if original range of attribute changes.

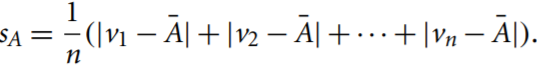
* + - * 1. Z-score Normalization(zero mean Normalization):

where A\_bar is mean of A and sigma\_A is its standard deviation.

Useful when we don't know min,max values and also when outliers dominate.

Variation to z-norm is :

Replace SD with mena absolute SD:



* + - * 1. Normalization by decimal scaling:

By moving decimal points

where j is the smallest int s.t. max(|v\_j|)<1.

* + 1. Discretization by binning:
       1. Top-down splitting technique based on a specified number of bins.
       2. equal-width/equal fre binning→ then smoothing by bin mean/median
       3. Unsupervised discretization.
    2. Discretization by histogram analysis:
       1. Unsupervised algo.
       2. Same as described under histograms section.
       3. Can be applied recursively to create multi level concept hierarchies.
       4. Can also be partitioned based on cluster analysis.
    3. Discretization by Cluster, Decision Tree, and Correlation Analyses:
       1. Cluster analysis:
          1. Of numeric attribute:
          2. Can be used to create concept hierarchy either top-down or bottom-up splitting.
       2. Decision tree:
          1. Top-down splitting
          2. Supervised.
          3. Entropy used to have points of splitting

Selects points having minimum entropy and recursively go on.

* + - 1. Correlation:
         1. ChiMerge method:

Bottom up approach

Finding best neighboring intervals and merging them.

* + - * 1. Chimerge steps:

Each distinct val of numeric attribute is considered a interval.

Chi^2 tests are done to merge adjacent intervals least ch^2 values

Least chi^2 val implies similar class distribution

Recursively done until a predefined stopping condition met.

* + 1. Concept Hierarchy Generation for Nominal Data:
       1. Nominal attributes have finite no of distinct val with no ordering.
       2. Methods to generate concept hierarchies for nominal data:
          1. Specification of a partial ordering of attributes explicitly at the schema level by users or experts

Like street < city < province or state < country.

* + - * 1. Specification of a portion of a hierarchy by explicit data grouping:

At intermediate levels

* + - * 1. Specification of a set of attributes, but not of their partial ordering:

User specified which attributes but not their ordering

Take max distinct val at the last then go on till least max distinct val attribute. This heuristic is good many times.

Does not always work like 20 yrs 12 months etc.

* + - * 1. Specification of only a partial set of attributes:

Like user being careless take only subset of attributes to put together.