Data analytics for Cyber security

-Introduction to Data Mining-

Vandana P. Janeja

Outline

Knowledge Discovery and Data Mining Process Models

Data Preprocessing

- Data Cleaning
- Data Transformation and Integration
- Data Reduction

Data Mining

- Measures of Similarity
- Measures of Evaluation
- Clustering Algorithms
- Classification
- Pattern Mining: Association Rule Mining

Data Mining

Discovery of hidden and nontrivial patterns

In very large datasets

But could also be smaller datasets with large number of features

The discovered patterns are potentially actionable and beneficial to study the problem at hand

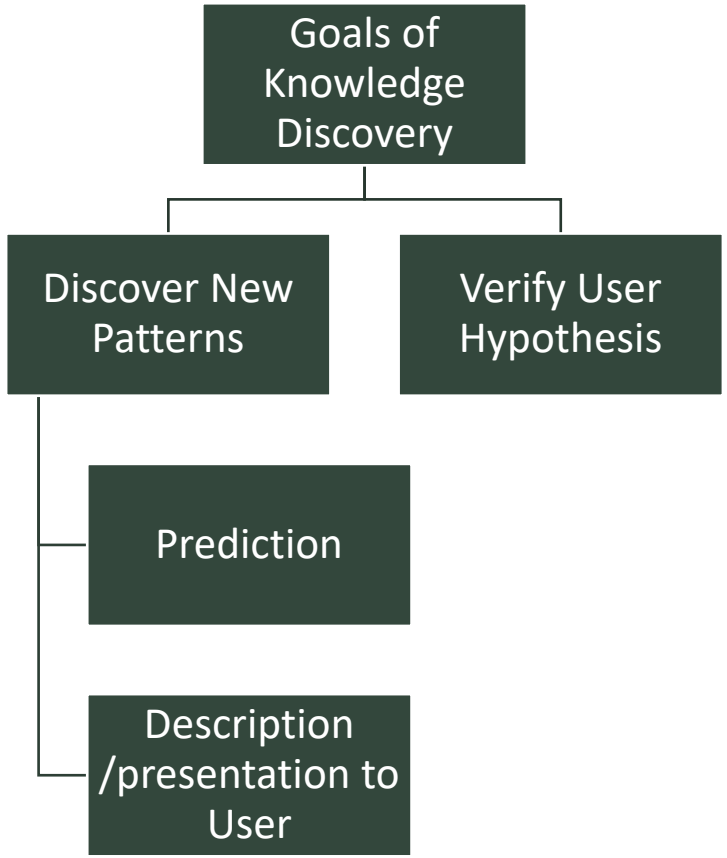
Data Mining

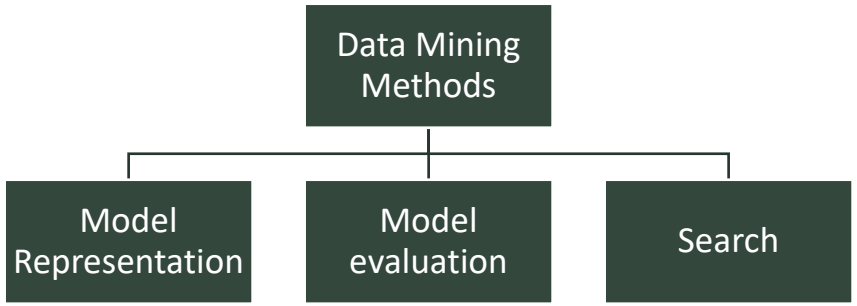
Goals of Knowledge Discovery include the discovery of new patterns or verifying a hypothesis that the users generally accept or are interested in evaluating, such as prediction of future events or explanation to the user in the form if understandable and intuitive knowledge discovered

Increasingly, as data are becoming very large, data mining fits into the core analytics of any big data solution

The core functionality of the mining algorithms does not change drastically; however, the infrastructure managing the parallelization is a key distinguishing factor in a big data environment.

Goals of Knowledge Discovery and components of Data Mining methods





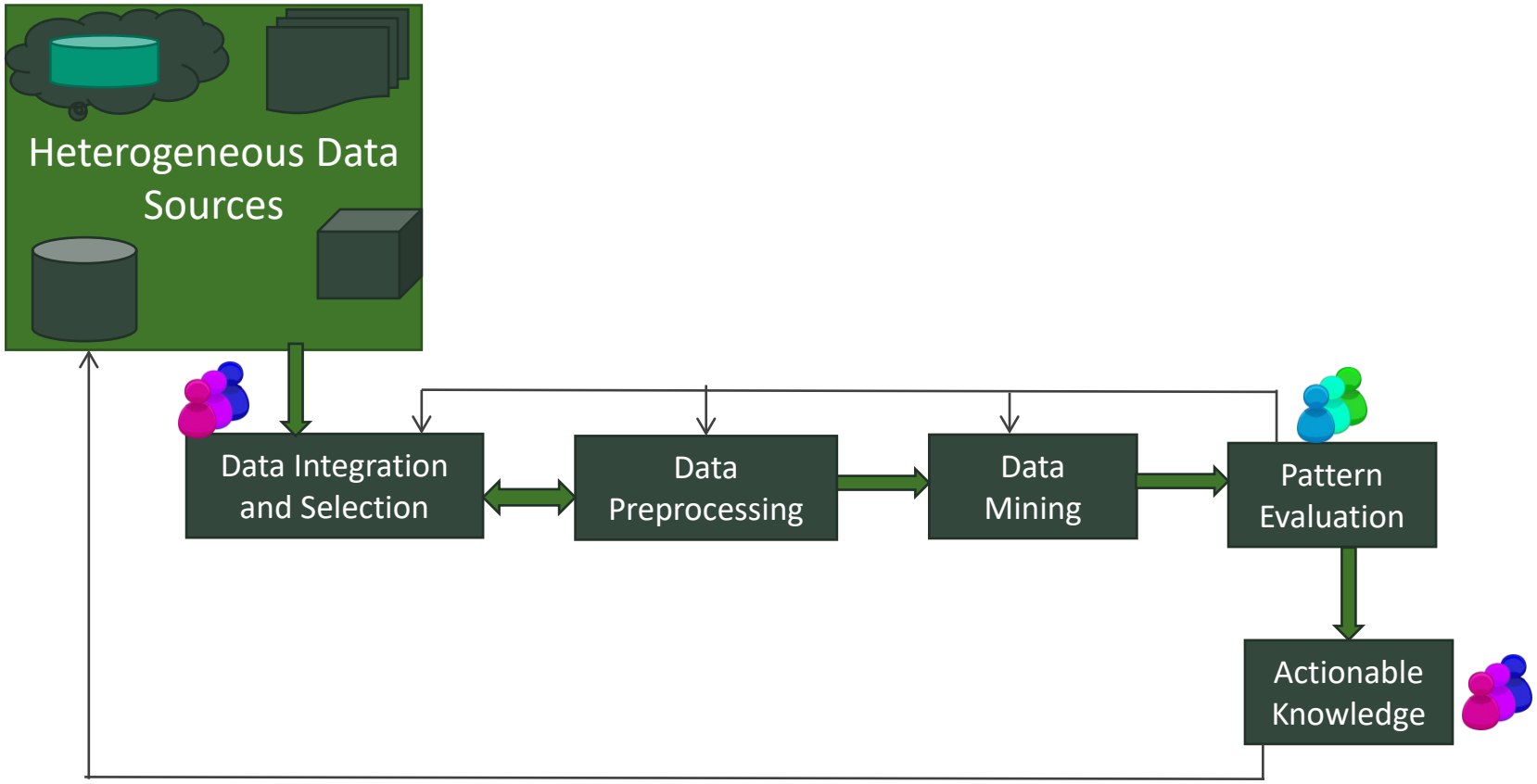
Data Mining method Components

The model being used and its representation	A method to evaluate the model	A search component
• This refers to the language used to describe the patterns along with any representational assumptions	• DM algorithms have an expected output, and quality of results based on this can be quantified through an evaluation metric (such as accuracy and error metrics)	• This involves a parameter search and model search. • This component allows for finding the optimal fit of both the model and parameter to use based on maximizing the evaluation criteria

DM method example evaluation criteria

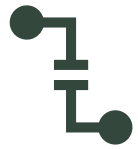
DM Method	Example Evaluation criteria
Classification	Accuracy, Precision, Recall, F-Measure
Clustering	Sum of Squared Error, Silhouette Coefficient
Association rule mining	Confidence, Support, Lift

Steps for a standard knowledge discovery process



Data Preprocessing

Data in the real world tend to be messy



Caused by device malfunction

Data entry errors

Or by data integration from heterogeneous sources.

Data Preprocessing

To get high-quality mining results that are actionable and beneficial, it is important to clean up the messy data

Data can be very large in terms of size and number of attributes, meaning that important patterns get hidden in the subspaces

It is important to focus on the right portions of the data both in terms of the volume and the attributes

Data preprocessing such as data cleaning, data transformation, and data reduction, as well as careful data integration, prepares the data for mining

These tasks will vary based on the data types and the complexity of the data

Data Types

Text

Numeric (discrete or continuous)

Ordinal (order or ranked data, such as severity scores, which have an inherent order to them)

Binary data (with only two possible outcomes)

Categorical data (with multiple categories).

Complex datasets may consists of multiple types of attributes in them

Examples Data Types

Attribute Data Type	Example
Numeric	Number of bytes sent
Categorical	Type of attack
Binary	Presence or absence of virus
Ordinal	Severity of alert(Low medium high)
Text	Intrusion alarm text
Complex Types of Datasets	**Example**
Unstructured Text	Spam email Intrusion log data
Spatial Data	Geographical Server locations Locations of Cyber physical sensors Locations of spam origination
Graph data	IP communication graph Routing paths for communication

Data Preprocessing

- The preprocessing tasks to apply are selected based on the complexity of the data and the attribute types
- For example:
  - Similarity in the data is defined differently for numeric attributes versus binary valued attributes
  - For complexity of the data, for example, spatial data, would require additional steps such as discovering spatial proximity, such that the mining task can be separated by spatial neighborhoods

Data Cleaning

Missing data values

Fluctuations in the data

Inconsistencies and redundancies in the data

Data Cleaning

Missing Data

- Replace by mean of the attribute if numerical or mode if categorical
- If the dataset is large or has multiple classes/groups - use the mean of the group to replace the missing values in that particular group
- Predict the missing values by training based on the records that have the complete data.
- Replace if roughly less than 5% of the data are missing; Consult with a domain expert or utilize domain heuristics to decide if data should be discarded or filled in to impute the missing data.

Noisy data

- Clean the noise by smoothing out the data to provide simpler and more refined representations
- Divide data into small groups, the mean behavior of the groups can be used to represent all the data in the group
- Small groups can be created by creating small bins or buckets in the data, and the mean or mode of the data can replace all the values in each bin.
- Binning - an unsupervised splitting-based discretization method of the data
- Equal frequency binning roughly deposits an equal number of points are across a fixed number of bins
- Equal distance binning distributes an equal distance across a fixed number of bins
- Clustering where similar data points are grouped together and the mean or centroid of the cluster can be used to replace the other values can also create a smooth dataset.

Outliers

- Decide beforehand through discussions with end users whether outliers need to be accommodated or eliminated
- Also determined by the types of patterns being discovered (for example: to find anomalies, outliers might be kept in the data)
- If the mining task is finding the normal behavior, such as the baseline network signatures, then consider removing outliers
- For univariate data, several statistical methods such as interquartile range (IQR)–based outlier detection, discordancy tests, and standard deviation–based tests can be used.
- For multivariate data, distance-based outlier detection methods can be utilized.

Data Transformation and Integration

- When data are merged from multiple sources, the resolution of the data may be different
  - Years rather than days.
  - Attribute ranges may vary, and these ranges may be very large
  - In these and other scenarios, the data will need to be transformed such that the traditional data mining algorithms can be uniformly applied.
  - Other data integration steps such as mapping the structure and content also need to be performed
- Normalization : One common data transformation method of normalization is the min–-max normalization or feature scaling

$$d' = \frac{d - minD}{maxD - minD}\left(new_{maxD} - new_{minD}\right) + new_{minD}$$

- In this method, a new range, for example 0 to –1, is created for the data such that the old min becomes 0 and the old max becomes 1
- All the data are transformed within the new range
- Allows for a rescaling of the data to a new well-defined range
- Other measures such as z-score normalization work well when the data is are normally distributed, and the mean is well defined.

Example Data preprocessing

Data
29
33
7
28
?
11
38
47
41
8
28
10
25
27
5
25
39
29
47
38
48
24

(a)

sorted values
5
7
8
10
11
24
25
25
27
27.95
28
28
29
29
33
38
38
39
41
47
47
48

Sorted Data with mean replacement

Equal Frequency binning

	Equal Freq Binning
	5
	7
Bin 1	8
	10
	11
	24
	25
	25
	27
	27.95
Bin2	28
	28
	29
	29
	33
	38
	38
Bin 3	39
	41
	47
	47
	48

For num bins B=3,
num obs=22
Bin Width=22/3=7
Last BW=7+1

Equal Distance binning

	Equal distance bin	Distance in Bin
	5	0
	7	2
Bin 1	8	3
	10	5
	11	6
	24	0
	25	1
	25	1
	27	3
	27.95	3.95
Bin2	28	4
	28	4
	29	5
	29	5
	33	9
	38	14
	38	14
	39	0
	41	2
Bin 3	47	8
	47	8
	48	9

(b)

For num bins B=3,
max=48, min=5
Distance Threshold=
(48-5)/3=14.33

Min-Max Normalization

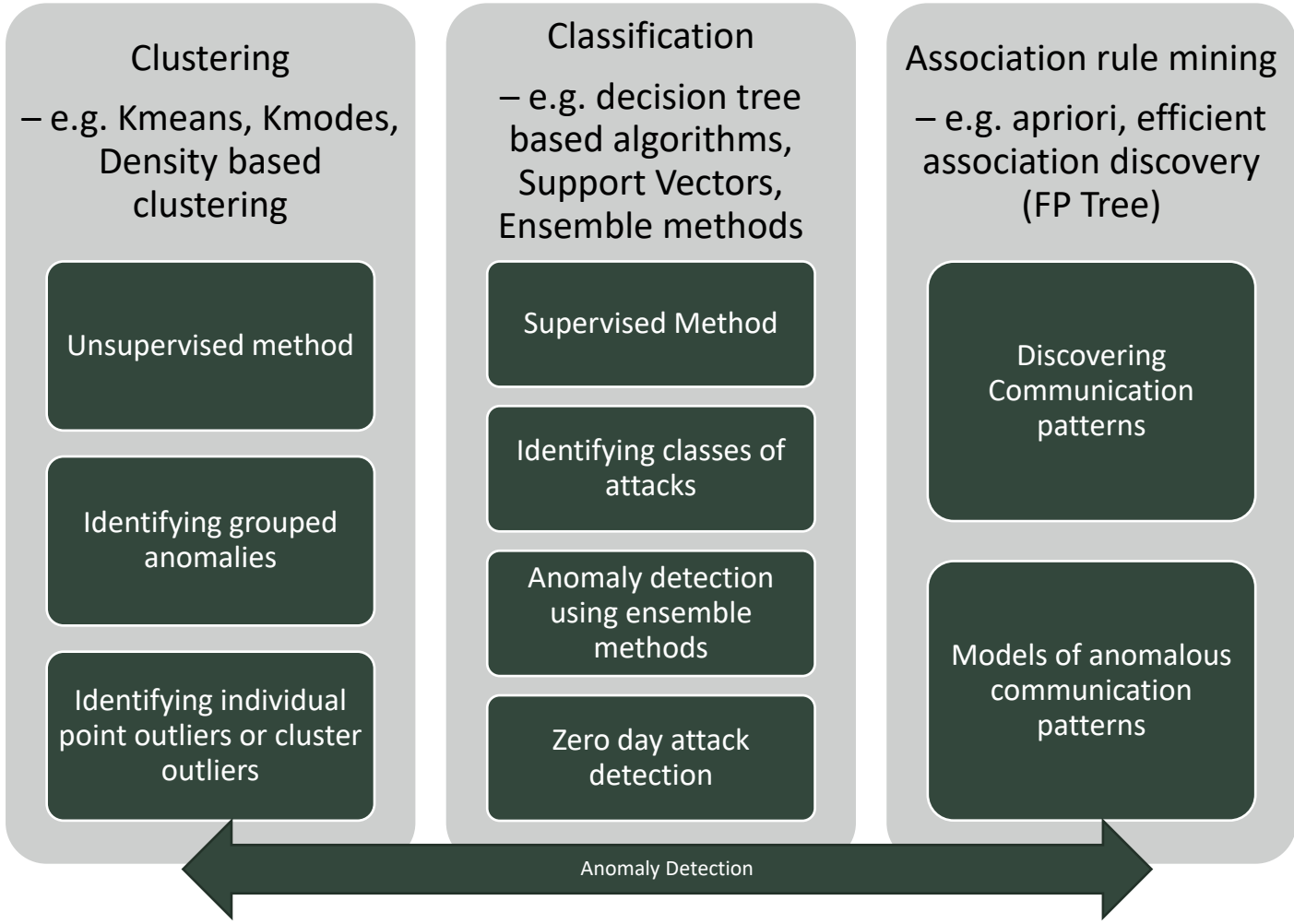
(c)

	Data	Min Max Normalization
	5	0
	7	0.046512
	8	0.069767
	10	0.116279
	11	0.139535
	24	0.44186
	25	0.465116
	25	0.465116
	27	0.511628
	27.95	0.533721
	28	0.534884
	28	0.534884
	29	0.55814
	29	0.55814
	33	0.651163
	38	0.767442
	38	0.767442
	39	0.790698
	41	0.837209
	47	0.976744
	47	0.976744
	48	1
min	5	0
max	48	1

Data Reduction

- Patterns in very large amounts of data- For example, network traffic data from Wireshark can become terabytes of data even within a few seconds; Similarly, source and destination-based communication graph can be a massive dataset. In such cases data reduction is an important strategy to help reduce the volume and attribute set of the data.
- Reducing dimensions - For example, in the Wireshark data capture, we might mainly be interested in looking at just the source and destination to generate a communication graph and annotate the graph with packet size.
- Sparse datasets - data reduction helps identify the key subspaces of the data
- Volume reduction- For example, smoothing techniques, central measures of the smoothed data can be used instead of all the data points.
- Feature selection - to identify the relevant features for performing data mining
  - Ranking-based methods that use entropy measures to quantify how much data are useful in each of the attributes.
  - Singular-value decomposition and principal component analysis work on the common principal to transform the data such that a combination of the features can be used as transformed features.

Data Mining Methods



Data Mining

- Data in security applications are heterogeneous, high-dimensional, and often complex.
- Three key groups of methods, namely clustering, classification, and association rule mining.
- The supervised methods are also referred to as machine learning, where historic data help train the algorithm to make predictions.
- Each of these methods also feeds into the discovery of anomalies or unusual patterns, which is particularly relevant to cybersecurity.
- Anomaly detection deals with finding individual objects and rare combinations of objects that are unusual (anomalous) with respect to other “normal” data points.
- Often the normal is a cluster or a set of clusters and the anomalies are the outlier data points.

Preliminaries to Data Mining

Measures of similarity: useful to find what is similar enough for creating clusters

- Distance Measures
- Similarity in Binary valued variables

Measures of evaluation: Various metrics that are used to evaluate the multiple data mining approaches

- Accuracy
- Precision
- Recall
- F-measure
- Lift
- Confidence

Example distance metrics

- Primarily used in clustering
- A non-negative function d(x,y), measuring distance between objects x and y, is a metric if it satisfies:
  - the identity property, d(x,x) = 0,
  - the symmetry property, d(x,y) = d(y,x), and
  - the triangle inequality, g(x,y) + g(y,z)≥g(x,z)

Distance Metric	Formula
Minkowski	$\text{MnD(X,Y)}=\sqrt[p]{(x_1 - y_1)^p + (x_2 - y_2)^p + \cdots + (xn - yn)^p}$
Manhattan	$\text{MD(X,Y)} = (\lvert x_1 - y_1\rvert + \lvert x2 - y2\rvert + \cdots + \lvert xn - yn\rvert)$
Euclidean	$\text{ED(X,Y)} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots . + (xn - yn)^2}$

Similarity in Binary valued variables

- Similarity coefficient matrix for binary valued attributes as illustrated
- Start with the data matrix comprising of feature vectors for each object
- The data matrix is thus an n ×m matrix consisting of n objects each with a feature vector of m features.
- The first step is to determine the contingency matrix using this data matrix which essentially is the values of a, b, c, and d for each data object pair.
  - a equals the number of positive matches such that objects oi and oj both have a value of 1,
  - b equals the number of mismatches such that oi has a value of 1 and oj has a value of 0,
  - c equals the number of mismatches such that oi has a value of 0 and oj has a value of 1, and
  - d equals the number of negative matches such that oi and oj both have a value of 0.
- Once these values are computed, they can be used to compute the similarity using different types of similarity coefficients such as Jaccard Coefficient (a/a+b+c). The output is a n × n triangular matrix with the computed similarity between each object pair

Data Matrix

O1	1	1	0	0
O2	1	1	1	0
O3	0	0	0	1

Contingency Matrix

	a (1-1)	b (1-0)	c (0-1)	d (0-0)
S(O1,O2)	2	0	1	1
S(O1,O3)	0	1	1	2
S(O2,O3)	0	3	1	0

Similarity Matrix (Jaccard Coefficient)

	O1	O2	O3
O1	1	0.5	0
O2		1	0
O3			1

Types of Similarity Coefficients

- The coefficients can be divided into Asymmetric and Symmetric.
- Asymmetric coefficients ignore negative matches (d), and should therefore be applied to data where absences (0s) are thought to carry no information.
  - An example is Jaccard's Coefficient.
  - For example if we would like to study attacks then a system which did not have a set of attacks is not a relevant question so features with a 0-0 match (i.e. not had an attack) is not relevant.
- Symmetric coefficients acknowledge negative matches, and should therefore be applied to data where absences are thought to carry information.
  - An example is the Simple Matching Coefficient.
  - For example if we are studying impact of a vulnerability patch on a software then if in two types of applications there was no impact of the patch then this is pertinent information we want to capture.
- There are some other types of coefficients called as hybrid which include the 0-0 match in either the numerator or denominator but not both.

Types of Similarity Coefficients

Asymmetric		
Coefficient	Expression	Prior Uses
Anderberg (range 0 to 1)	$\frac{a}{a+2(b+c)}$	Biochemistry & Molecular Biology; Genetics & Heredity
Jaccard/Tanimoto (range 0 to 1)	$\frac{a}{a+b+c}$	Plant Sciences; Agronomy; Horticulture; Genetics & Heredity; Ecology
Kulczynski (range 0 to 1)	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	Marine & Freshwater Biology; Ecology; Agriculture; Chemistry;Forestry
Ochiai/Cosine (range 0 to 1)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	Computer Science; Analytical Chemistry; Information Science & Library Science; Biochemical Research Methods; Chemistry
Sorensen-Dice (range 0 to 1)	$\frac{2a}{2a+b+c}$	Biochemistry & Molecular Biology; Genetics & Heredity; Agriculture; Biology; Plant Sciences
Symmetric		
Baulieu (range -1 to 1)	$\frac{4(ad-bc)}{(a+b+c+d)^2}$	Mathematics; Mathematical Psychology
Simple Matching (range 0 to 1)	$\frac{a+d}{a+b+c+d}$	Plant Sciences; Biotechnology & Applied Microbiology; Electrical & Electronic Engineering; Agronomy; Microbiology
Russel & Rao (Hybrid, range 0 to 1)	$\frac{a}{a+b+c+d}$	Chemistry; Information Systems; Computer Science; Ecology; Forestry

Evaluation Criteria

- Every data mining method needs to have a way to evaluate the results.
- The algorithms have a certain expected output,
- Evaluation measures help ensure the quality of results from an algorithmic perspective.
- Additional quality measures such as usefulness, ground truth evaluation and business evaluation also need to be done.

DM Method	Example Evaluation criteria
Classification	Accuracy = $\frac{TP+TN}{(TP+FP+FN+TN)}$ Precision=$\frac{TP}{(TP+FP)}$ Recall=$\frac{TP}{(TP+FN)}$ F-Measure=$\frac{2*Precision*Recall}{(Precision+Recall)}$
Clustering	Sum of Squared Error= $\sum_{i=1}^{K}\sum_{x\in C_i} dist^2(m_i,x)$ Silhouette Coefficient=$S(o)=\frac{b(o)-a(o)}{\max\{a(o),b(o)\}}$
Association rule mining	Confidence X$\rightarrow$Y=P(Y|X)=$\frac{P(XUY)}{P(X)}$ Support=X$\rightarrow$Y=$\frac{P(XUY)}{Total\ Population}$ Lift =$\frac{P(XUY)}{P(X)P(Y)}$