Data Mining Summary

Don't expect everything here

Andy Chong Chin Shin

Contents

Chapter 1

Introduction

1.1 Knowledge Discovery in Database, KDD

The overall process of non-trivial extraction of implicit, previously unknown and potentially useful knowledge from large amounts of data

$$
\begin{aligned}
Database &\xrightarrow{Data\ Cleaning} DataWarehouse \\
&\xrightarrow{Selection\ Transformation} Taskrelevant\ Data \\
&\xrightarrow{Data\ Mining} Pattern \\
&\xrightarrow{Pattern\ Evaluation\ and\ Visualisation} Knowledge
\end{aligned}
$$

1.2 Data Mining

Prediction Methods Classification, Outlier Detection, Regression

Description Methods Clustering, Association Rule, Sequence Pattern

Chapter 2

Data

Data is collection of data object and their attributes.

Type of Attributes Nominal, Ordinal, Interval, Ratio.

Attribute Values Property Distinctness, Order, Addition, Multiplication

One idea to determine if an interval attribute is ratio is to check if the attribute has a true or natural absolute zero point. The ratio scales are very common in physical scenario.

I do not focus on Types of Data Sets and Characteristics of Structured Data

2.1 Discrete and Continuous

Discrete Attribute Has only a finite or countably infinite set of values

Continuous Attribute Has real numbers as attribute values

2.2 Data Quality

Noise modification of original values

Outlier considerably different than most of the other data objects

2.3 Discretization and Binarization

Binarization transform either a continuous attribute or a categorical attribute into one or more binary attributes

Discretisation transform a continuous attribute into a categorical attribute

2.4 Normalization

Min-Max Normalization

$$(min_A, max_A) \rightarrow (new\_min_A, new\_max_A)$$

$$v^{'} = \frac{v - min_A}{max_A - min_A} \times (new\_max_A - new\_min_A) + new\_min_A$$

Z-score Normalization

$$v^{'} = \frac{v - \mu_A}{\sigma_A}$$

Normalisation by Decimal Scaling

$$v^{'} = \frac{v}{10^j}$$

where j = smallest integer such that $Max(\mid v^{'} \mid) < 1$

2.5 Proximity

Similarity Numerical measure of how alike two data objects are

Dissimilarity Numerical measure of how different two data objects are

Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Minkowski Distance

$$dist = \sum_{k=1}^{n} (\mid p_k - q_k \mid^r)^{1/r}$$

$$\begin{aligned} r = 1 &: Hamming\ distance / Manhattan\ distance \\ r = 2 &: Euclidean\ distance \\ r = inf &: supremum\ /\ Chebyshev\ distance, max(x - y) \end{aligned}$$

Simple Matching Coefficient

$$SMC = \frac{M_{11} + M_{00}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

Jaccard Coefficient

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Cosine Similarity

$$cos(p, q) = \frac{p \bullet q}{\|p\| \|q\|}$$

Tanimoto Coefficient

$$T(p, q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

2.6 Correlation

$$corr(p, q) = \frac{cov(p, q)}{std(p) \times std(q)}$$

$$cov(p, q) = \frac{1}{n-1} \sum_{k=1}^{n} (p_k - \bar{p})(q_k - \bar{q})$$

$$std(p) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (p_k - \bar{p})^2}$$

Chapter 3

Data Analytics

3.1 Data Warehouse

Data Warehouse subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process

Data Warehousing The process of constructing and using data warehouses

Query-Driven When a query comes, wrappers translate the query for each DB. Integrators combine results from different DB.

Update-Drive Information from heterogenous sources is integrated in advance and stored in a DW for direct querying and analysis. High performance, but no most recent information.

3.2 On-line Analytical Processing

OLAP uses a multidimensional array representation.

Data cube is a multidimensional representation of data, together with all possible aggregates.

Aggregates mean the result by selecting a proper subset of the dimensions and summing over all the remaining dimensions.

Categories of Data Cube Measures

- Distributive: sum, count, min, max
- Algebraic: average, std, maxN, minN
- Holistic: median, mostFrequent, rank

Chapter 4

Classification

In classification task, we find a model for class attribute as a function of the values of other attributes. The goal is to assign previously unseen records a class as accurately as possible.

4.1 Decision Tree based

4.1.1 Hunt's Algorithm

Hunt's algorithm grows a decision tree in a recursive fashion by partitioning the training records into successively purer subsets. Let D_t be the set of training records that reach a node t:

- If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
- If D_t is an empty set, then t is a leaf node labeled by the default class
- If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets.

4.1.2 Measure of Node Impurity

Gini Index

$$GINI(t) = 1 - \sum_j P(j \mid t)^2$$

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Entrophy

$$Entrophy(t) = -\sum P(j \mid t) log_2 P(j \mid t)$$

$$GAIN_{split} = Entrophy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entrophy(i)$$

Disadvantage: Tends to prefer splits that result in large number of partitions, each being small but pure.

Introduce Gain Ratio:

$$GainRATIO_{split} = \frac{GAIN_{split}}{SplitINFO}$$

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} log_2 \frac{n_i}{n}$$

Misclassification Error

$$Error(t) = 1 - max_i P(i \mid t)$$

Hint: To compute maximum value for Gini Index, Entrophy, or Misclassification Error, subtitute $P(j \mid t)$ with $1/n_c$. Minimum values are always 0.

4.2 Rule-based

Classify records by using a collection of "if... then..." rules.

4.2.1 Rule Coverage and Accuracy

Given a rule r in a dataset D:

$$r : A \rightarrow y$$

$$coverage = \frac{\mid A \mid}{\mid D \mid}$$

$$accuracy = \frac{\mid A \cap y \mid}{\mid A \mid}$$

4.2.2 Characteristics of Rule-Based Classifier

Mutually exclusive Every record is covered by at most one rule. No two rules are trigger by the same record.

Exhaustive Each record is covered by at least one rule

4.3 Model Evaluation

Focus on predictive capability of a model

4.3.1 Metrics for Performance Evaluation

Confusion Matrix:

Count	Predicted Class=YES	Predicted Class=NO
Actual Class = YES	a(TP)	b(FN)
Actual Class = NO	c(FP)	d (TN)

$$Accuracy = \frac{a+d}{a+b+c+d}$$

$$Weighted\ Accuracy = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

$$Precision, p = \frac{a}{a+c}$$

$$Recall, r = \frac{a}{a+b}$$

$$F - measure, f = \frac{2a}{2a+b+c}$$

4.3.2 Methods for Performance Evaluation

Holdout reserve 2/3 for training and 1/3 for testing

Random subsampling repeated holdout

Cross validation partition data into k disjoint subsets

Stratified sampling oversampling vs undersampling

Bootstrap sampling with replacement

4.3.3 Methods for Model Comparison

ROC curve is a graph of TP rate againts FP rate

$$TP\ rate, TPR = \frac{a}{a+b}$$

$$FP\ rate, FPR = \frac{c}{c+d}$$

Chapter 5

Classification: Alternative

5.1 Underfitting and Overfitting

Re-substitution Errors, e: error on training
Generalization Errors, e': error on testing

I do not focus on Support Vector Machines and Artificial Neural Network

Optimistic approach to estimate e':

$$e' = e$$

Pessimistic approach to estimate e':

$$e' = e + \frac{\#leaf \times 0.5}{Total\ \#\ Instances}$$

Occam's Razor Given two models of similar generalization errors, one should prefer the simpler model over the more complex model.

5.2 k-Nearest Neighbour

Instance based classifier: Use training records directly to predict the class label of unseen cases

K-nearest neighbors of a record x are data points that have the k smallest distance to x.

5.3 Naïve Bayes Classifiers

Bayes Theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Naïve Bayes Classifiers: Compute the posterior probability for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C)P(C)}{P(A_1 A_2 \ldots A_n)}$$

Assume independence among attributes A_i

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 \mid C_j)P(A_2 \mid C_j) \ldots P(A_n \mid C_n)P(C)}{P(A_1 A_2 \ldots A_n)}$$

For discrete attribute:

$$P(A_i \mid C_j) = \frac{\mid A_{ik} \mid}{N_c}$$

For continuous attribute, can use probability density estimation:

$$P(A_i \mid C_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp^{-\frac{A_i - \mu_j^2}{2\sigma_j^2}}$$

5.4 Support Vector Machines

Find a linear hyperplane (decision boundary) that will separate the data

$$f(\vec{x}) = \begin{cases} 1 & if\ \vec{w} \bullet \vec{x} + b \geq 1 \\ 0 & if\ \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

5.5 Ensemble Classification

Predict class label of previously unseen records by aggregating predictions made by multiple classifiers

Assumption: Individual classifiers could be lousy, but the aggregate can usually classify correctly.

5.5.1 Bagging

Simplified steps:

1. Sampling with replacement to get k set of data
2. Train multiple k models on k different samples
3. For each test example, predict by using simple majority voting

5.5.2 Boosting

An iterative procedure to adaptively change distribution of training data by focusing more on previously misclassified records.

Records that are wrongly classified will have their weights increased. Records that are correctly classified will have their weights decreased

Chapter 6

Cluster Analysis

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

6.1 Type of Clustering

Partitional Clustering A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

Hierarchical clustering A set of nested clusters organized as a hierarchical tree

6.2 Type of Clusters

Well-Separated Clusters any point in a cluster is closer to every points in the cluster than to any *point* not in the cluster

Center-Based Clusters object in a cluster is closer to the center of a cluster, than to the *center* of any other cluster

Contiguous Clusters Neighbourhood relationship, *each point is close to another point* in the cluster, immediate neighbour

Density-Based Clusters A cluster is a *dense region* of points, which is separated by low-density regions, from other regions of high density

Property or Conceptual Clusters that share some common property or represent a particular concept

Described by Objective Function Find clusters that minimise or maximise an objective function

6.3 K-mean Clustering

1: Select k points as initial centroids
2: Repeat
3: Form k clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster
5: Until the centroids don't change

Evaluating k-means clusters using Sum of Squared Error (SSE):

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

6.4 Hierarchical Clustering

6.4.1 Agglomerative Clustering Algorithm

1: Compute the proximity matrix
2: Let each data point be a cluster
3: Repeat
4: Merge the two closest clusters
5: Update the proximity matrix
5: Until only a single cluster remains

Cluster Similarity: Group Average

$$proximity(Cluster_i, Cluster_j) = \frac{\sum proximity(p_i, p_j)}{\mid Cluster_i \mid\mid Cluster_j \mid}$$

6.4.2 Divisive Clustering Algorithm: Minimum Spanning Tree

1:	Compute a minimum spanning tree for the proximity graph
2:	Repeat
3:	Create a new cluster by breaking the link corresponding to the largest distance
4:	Until only singleton clusters remain

Same as single link agglomerative clustering

Chapter 7

Cluster Analysis: Alternative

7.1 DBSCAN

Two core points within *Radius* are put into the same cluster

Core point has more than *MinPoint* within *Radius*

Border point has fewer than *MinPoint* within *Radius*, but is in the neighborhood of a core point.

Noise point other points

7.2 CURE

Hierarchical Approach. Representative points are found by selecting constant number of points from a cluster and then "shrinking" them toward the center of the cluster.

7.3 Graph-Based Clustering

Builds a graph using proximity matrix, then breaks the graph using sparsification. *Sparsification* keeps the connections to the most similar neighbors while breaking the connections to less similar points. Clusters are connected components in the graph.

7.3.1 Chameleon

Use a dynamic model to measure the similarity between clusters

1: Build a k-nearest neighbor graph
2: partition the graph using a multilevel graph partitioning algorithm
3: **repeat** merge the clusters that best preserve the cluster self-similarity with respect to relative inter-connectivity and relative closeness
4: **until** no more cluster can be merged

7.3.2 Shared Near Neighbor Approach

SNN Graph The weight of an edge is the number of shared neighbours between vertices

Computing shared nearest neighbor similarity:

1: Find the k-nearest neighbors of all points
2: **if** two points, x and y are not among the k-nearest neighbors of eat other **then**
3: $similarity(x, y) = 0$
4: **else**
5: $similarity(x, y) =$ number of shared neighbors

7.3.3 Jarvis-Patrick Clustering

A threshold is used to sparsify SNN similarity matrix

A pair of points is put in the same cluster if:

1. share more than T neighbours
2. in each others k nearest neighbour list

1: Compute the SNN similarity graph
2: Sparsify the SNN similarity graph by applying a similarity threshold
3: Find the connected components of the sparsified SNN similarity graph

7.4 Cluster Validity

External Index measure the extent to which cluster labels match externally supplied class labels (Entrophy)

Internal Index measure the goodness of clustering structure without respect to external information (SSE)

Relative Index compare two different clustering or clusters

7.4.1 Measuring Cluster Validity via Correlation

Compute the correlation between proximity matrix and incidence matrix. High correlation indicates that points that belong to the same cluster are close to each other.

7.4.2 Internal Measures: Cohesion and Separation

Cohesion, within cluster SSE:

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

Separation, between cluster SSE:

$$BSS = \sum_i \mid C_i \mid (m - m_i)^2$$

where $\mid C_i \mid$ is the size of cluster i

and m is true mean of all point

7.4.3 Internal Measures: Silhouette Coefficient

$$s = 1 - \frac{a}{b}$$

i = individual point a = average distance of i to the points in its cluster
b = min (average distance of i to points in another cluster)
Closer to 1 is better

Chapter 8

Association Rule Mining

Itemset A collection of one or more items

Frequent itemset An itemset whose support is greater than or equal to a *minsup* threshold

$$let\ X \Rightarrow Y$$

$$\begin{aligned} Support\ Count, \sigma &= \#Trans(X \cup Y) \\ Support, s &= \frac{\#Trans(X \cup Y)}{Total\ \#\ Trans} \\ Confidence = P(Y|X) &= \frac{\#Trans(X \cup Y)}{\#Trans\ contain\ X} \end{aligned}$$

Goal of Association Rule Mining:

$$support \geq min\_sup$$
$$confident \geq min\_conf$$

8.1 Frequent Itemset Generation

Apriori Principle: If an itemset is frequent, then all of its subsets must also be frequent. This principle reduce the number of frequent itemset candidates

$$\forall X, Y : X \subseteq Y \Rightarrow s(X) \geq s(Y)$$

Store each candidate in a hash tree structure to count the support efficiently

Maximal Frequent itemset none of its immediate supersets is frequent

Closed itemset none of its immediate supersets has the same support as the itemset

8.1.1 Alternative: FP-Growth Algorithm

Construct FP-tree from a transactional DB:

1:	Scan DB once, find frequent 1-itemset
2:	Order frequent items in frequency descending order (L-order)
3:	Process DB based on L-order

Mining frequent patterns using FP-tree:

1:	Construct conditional pattern base for each item in header table
2:	Construct conditional FP-tree from each conditional pattern-base
3:	Recursively mine conditional FP-trees and grow frequent patterns obtained so far

8.2 Rule Generation

Confidence of rules generated from the same itemset has an anti-monotone property:

$$c(ABC \to D) \geq c(AB \to CD) \geq c(A \to BCD)$$

Chapter 9

Anomaly Detection

9.1 Type of Anomaly

1. Point Anomalies (Our Focus)
2. Contextual Anomalies
3. Collective Anomalies

9.2 Anomaly Detection Techniques

General steps

1:	Build profile of normal behavior
2:	Use the normal profile to detect anomalies

Graphical Approaches Boxplot, Scatter plot, Spin plot

Statistical Approaches points are determined to be outliers depending on their relationship with this model

Nearest Neighbor Based Approaches Distance based methods or density based methods

Classification Based Approaches Supervised learning techniques

Evaluation of Anomaly Detection: Use F-measure and ROC curve

Chapter 10

Appendix

Classifier	**Advantages**	**Disadvantages**
Decision Tree Based	Inexpensive to construct Extremely fast at classifying unknown records Easy to interpret for small-sized trees Accuracy is comparable to other classification techniques for many simple data sets	
k-Nearest Neighbour	Easy to implement Incremental addition of training data trivial	Lazy learner, expensive classification Sensitive to noise because it only uses local information

Naive Bayes Classifier	Robust to isolated noise points Handle missing values by ignoring the instance during probability estimate calculations Robust to irrelevant attributes	Independence assumption may not hold for some attributes Use other techniques such as Bayesian Belief Networks (BBN)
Ensemble Classifier: Bagging	Decrease variance, improve stability (tolerance to noise) Can be parallelized	Reduces accuracy for stable classifiers because sample size reduced by 36%

Table 10.1: Advantages and disadvantages of various classifiers

Clustering	**Advantages**	**Disadvantages**
K-mean Clustering		K-means has problems when clusters are of different size, density and shapes K-means has problems when the data contains outliers
Hierarchical Clustering	Do not have to assume any particular number of clusters They may correspond to meaningful taxonomies	Once a decision is made to combine two clusters, it cannot be undone No objective function is directly minimized

Agglomerative Clustering using MIN or Single Link	Can handle non-elliptical shapes	Sensitive to noise and outliers
Agglomerative Clustering using MAX or Complete Link	Less susceptible to noise and outliers	Tends to break large clusters Biased towards globular clusters
Agglomerative Clustering using Group Average	Less susceptible to noise and outliers	Biased towards globular clusters
Agglomerative Clustering using Ward's Method	Less susceptible to noise and outliers	Biased towards globular clusters
DBSCAN	Resistant to Noise Can handle clusters of different shapes and sizes	Cannot handle different densities Cannot handle high-dimensional data
CURE	Shrinking representative points toward the center helps avoid problems with noise and outliers Handle clusters of arbitrary shapes and sizes	Cannot handle differing densities
Jarvis-Patrick Clustering	Can handle different density	Cannot handle different shapes
SNN Density Based Clustering		Does not cluster all the points Complexity of SNN Clustering is high

Table 10.2: Advantages and disadvantages of various clustering techniques

Association Rule Mining algorithms	**Advantages**	**Disadvantages**
Apriori Algorithm		Multiple database scans are costly Mining long patterns needs many passes of scanning and generates lots of candidates Bottleneck: candidate generation and test
FP-Tree	Highly condensed, but complete for frequent pattern mining Avoid costly database scans Develop an efficient, FP-tree-based frequent pattern mining method Avoid candidate generation	Support dependent; cannot accommodate dynamic support threshold Cannot accommodate incremental DB update Mining requires recursive operations

Table 10.3: Advantages and disadvantages of various Association Rule Mining algorithms

Anomaly Detection Technique	**Advantages**	**Disadvantages**
Graphical Approaches		Time consuming Subjective
Statistical Approaches	Utilize existing statistical modeling techniques to model various type of distributions	With high dimensions, difficult to estimate distributions Parametric assumptions often do not hold for real data sets Most of the tests are for a single attribute
Nearest Neighbour Based Approaches	Can be used in unsupervised or semi-supervised setting	If normal points do not have sufficient number of neighbors the techniques may fail Computationally expensive
Classification Based Approaches	Models that can be easily understood High accuracy in detecting many kinds of know anomalies	Require both labels from both normal and anomaly class Cannot detect unknown and emerging anomalies

Table 10.4: Advantages and disadvantages of various Anomaly Detection Techniques

Summary Statistic	**Advantages**	**Disadvantages**
Mean	Mathematical	Sensitive to outlier
Median	Not sensitive to outlier	Not mathematical
Mode	Not affected by outlier	Not mathematical
Inter-quartile range	Capture majority of the data very cheaply	Does not say how data is distributed within the range
Scatter plot	Show all points	If there are too many points it would be unclear
Discretisation	Helps to apply algorithms that cannot handle continuous data	Information loss
Assuming Normal Distribution	If data is truly normally distributed, it approximates well	Otherwise it makes mistakes
Standard Deviation	Mathematical	Sensitive to outlier
Histogram	Easy to inspect and analyse	Information loss
Raw Data	No information loss	Difficult to process and comprehend

Table 10.5: Advantages and disadvantages of various summary statistics