

CSE5230/DMS/2004/4

Data Mining - CSE5230

Clustering Techniques Association Rule Discovery

CSE5230 - Data Mining, 2004 Lecture 4.1

Lecture Outline

- ◆ **Automatic Cluster Detection**
 - ❖ The *k*-means Technique
 - ❖ Similarity, Association, Distance
 - » Types of Variables, Measures of Similarity, Weighting and Scaling
 - ❖ Agglomerative Techniques
- ◆ **Association Rules**
 - ❖ Usefulness
 - ❖ Example
 - ❖ Choosing the right item set
 - ❖ What is a rule?
 - ❖ Is the Rule a Useful Predictor?
 - ❖ Discovering Large Itemsets
 - ❖ Strengths and Weaknesses

CSE5230 - Data Mining, 2004 Lecture 4.2

Lecture Objectives

- ◆ **By the end of this lecture, you should be able to:**
 - ❖ explain what is meant by cluster detection, and give an example of clusters in data
 - ❖ understand how the k -means clustering technique works, and use it to do a simple example by hand
 - ❖ explain the importance of similarity measures for clustering, and why the Euclidean distance between raw data values is often not good enough
 - ❖ describe the components of an association rule (AR)
 - ❖ indicate why some ARs are more useful than others
 - ❖ give an example of why classes and taxonomies are important for association rule discovery
 - ❖ explain the factors that determine whether an AR is a useful predictor
 - ❖ Understand the basic idea of the *a priori* “trick”

Automatic Cluster Detection

- ◆ **If there are many competing patterns, a data set can appear to contain just noise**
- ◆ **Subdividing a data set into clusters where patterns can be more easily discerned can overcome this**
- ◆ **When we have no idea how to define the clusters automatic cluster detection methods can be useful**
- ◆ **Finding clusters is an *unsupervised learning* task**

Types of Clustering Techniques

- ◆ There are two main “families” of clustering techniques

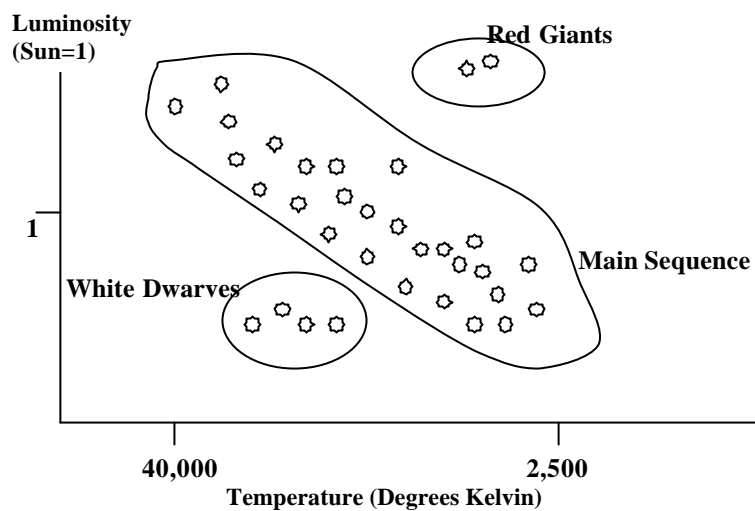
- ❖ Partitional – based on splitting up the data, e.g.

- » *k*-means
 - » Mixture models

- ❖ Agglomerative – based on merging data items or sub-clusters, e.g.

- » Ascendant Hierarchical Clustering
 - » Link-based methods

Example: The Hehrtzprung-Russell diagram



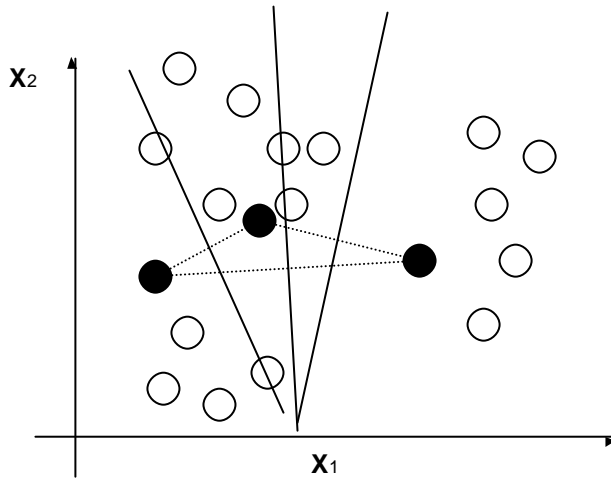
Automatic Cluster Detection - example

- ◆ **The Hehrtzprung-Russell diagram, which graphs stars' luminosities against their temperatures, reveals three clusters**
 - ❖ It is interesting to note that each of the clusters has a different relationship between luminosity and temperature.
- ◆ **In most data mining situations the variables to consider and the clusters that may be formed are not so easily determined**

The *k*-means Technique

- ◆ ***k*, the number of clusters that are to be formed, must be decided before beginning**
 - ❖ **Step 1**
 - » Select *k* data points to act as the seeds (or initial cluster centroids)
 - ❖ **Step 2**
 - » Each record is assigned to the centroid which is nearest, thus forming a cluster
 - ❖ **Step 3**
 - » The centroids of the new clusters are then calculated. Go back to Step 2
 - ❖ This is continued until the clusters stop changing

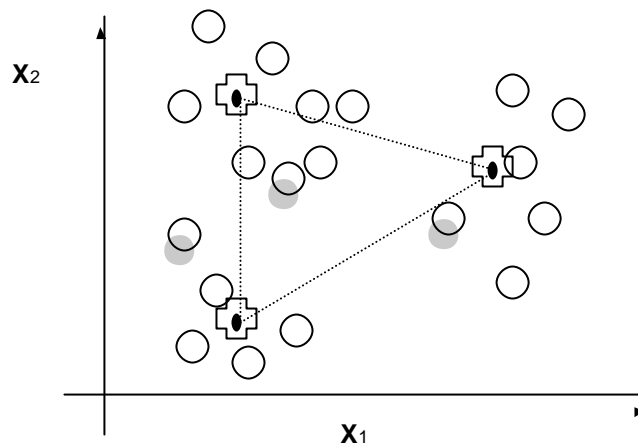
Assign Each Record to the Nearest Centroid



CSE5230 - Data Mining, 2004

Lecture 4.9

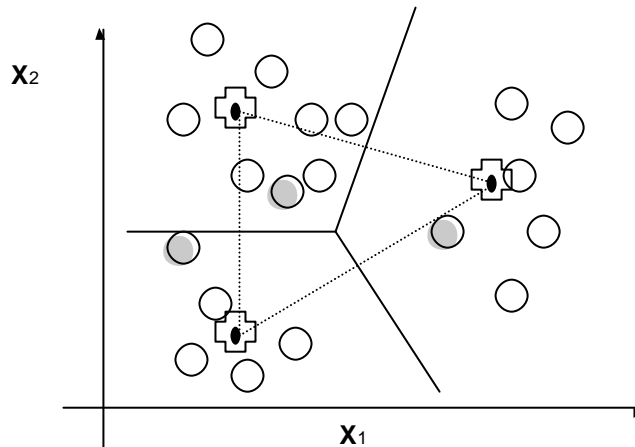
Calculate the New Centroids



CSE5230 - Data Mining, 2004

Lecture 4.10

Determine the New Cluster Boundaries



CSE5230 - Data Mining, 2004

Lecture 4.11

Similarity, Association and Distance

- ◆ The method just described assumes that each record can be described as a point in a metric space
 - ❖ This is not easily done for many data sets (e.g. categorical and some numeric variables)
 - » Preprocessing is often necessary
- ◆ Records in a cluster should have a natural association. A measure of similarity is required.
 - ❖ Euclidean distance is often used, but it is not always suitable
 - ❖ Euclidean distance treats changes in each dimension equally, but changes in one field may be more important than changes in another
 - » and changes of the same “size” in different fields can have very different significances
 - e.g. 1 metre difference in height vs. \$1 difference in annual income

CSE5230 - Data Mining, 2004

Lecture 4.12

Types of Variables

◆ Nominal (categories)

❖ e.g. Food Group: Grain, Dairy, Meat, etc.

◆ Ordinal (ranks)

❖ e.g. Food Quality: Premium, High Grade, Medium, Low

◆ Intervals

❖ e.g. The distance between temperatures

◆ True Measures

❖ The measures have a meaningful zero point so ratios have meaning as well as distances

Measures of Similarity

◆ Euclidean distance between vectors X and Y:

$$d(X, Y) = \sqrt{\sum_{i=1}^D (x_i - y_i)^2}$$

❖ In matrix notation, we write:

$$d(X, Y) = \sqrt{(X - Y)^T (X - Y)}$$

◆ Angle between two vectors (from origin to data point)

◆ The number of features in common (typically for nominal data)

◆ and many more...

Weighting and Scaling

- ◆ **Weighting allows some variables to assume greater importance than others, e.g.**

$$d_w(X, Y) = \sqrt{\sum_{i=1}^D w_i (x_i - y_i)^2}$$

- ❖ The domain expert must decide if certain variables deserve a greater weighting
- ❖ Statistical weighting techniques also exist
- ◆ **Scaling is a kind of weighting that attempts to apply a common range to variables, so that differences are comparable between variables**
 - ❖ This can also be statistically based

Mahalanobis Distance

- ◆ **The Mahalanobis distance between two data vectors X and Y with respect to a cluster j:**

$$d_{M_j}(X, Y) = \sqrt{(X - Y)^T \Sigma_j^{-1} (X - Y)}$$

where Σ_j is the covariance matrix for cluster j

- ◆ **The Mahalanobis takes into account the variances and covariances of each dimension of each cluster**
 - ❖ This handles variables with different scales, AND
 - ❖ importantly, this also for the differing shapes of clusters to be taken into account
- ◆ **The use of Euclidean distance tacitly assumes that clusters are spherical, and of the same size**
- ◆ **The Mahalanobis distance allows for ellipsoidal, of varying sizes and orientations.**

Variants of the k -means Technique

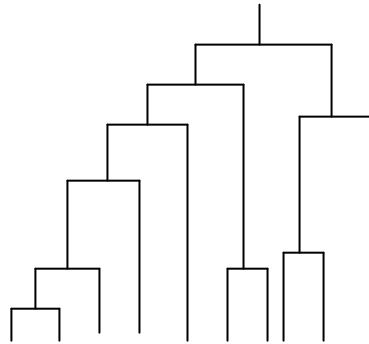
- ◆ There are problems with simple k -means method:
 - ❖ It does not deal well with overlapping clusters.
 - ❖ The clusters can be pulled off-centre by outliers.
 - ❖ Records are either in or out of the cluster, so there is no notion of likelihood of being in a particular cluster or not
- ◆ A *Gaussian Mixture Model* varies the approach already outlined by attaching a weighting based on a probability distribution to records which are close to or distant from the centroids initially chosen. There is then less chance that outliers will distort the situation. Each record contributes to some degree to each of the centroids.

Agglomerative Techniques - 1

- ◆ A completely unsupervised technique would not pre-determine the number of clusters
- ◆ A hierarchical clustering offers offer a hierarchy of clusters from large to small. This can be achieved in a number of ways, e.g.
 - ❖ An agglomerative technique starts out by considering each record as a cluster and gradually building larger clusters by merging the records which are near each other
 - ❖ The alternative is to start with one cluster for the whole data set, and then split it recursively

Agglomerative Techniques - 2

- ◆ An example of an agglomerative cluster tree:



CSE5230 - Data Mining, 2004

Lecture 4.19

Evaluating Clusters

- ◆ We desire clusters to have members which are close to each other and we also want the clusters to be widely spaced
- ◆ Variance measures are often used. Ideally, we want to minimize within-cluster variance and maximize between-cluster variance
- ◆ But variance is not the only important factor, for example it will favor not merging clusters in an hierarchical technique

CSE5230 - Data Mining, 2004

Lecture 4.20

Strengths of Automatic Cluster Detection

◆ Strengths

- ❖ is an undirected knowledge discovery technique
- ❖ works well with many types of data
- ❖ is relatively simple to carry out

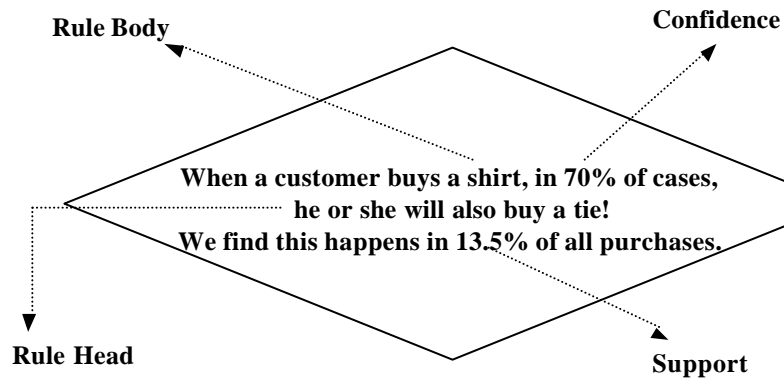
◆ Weaknesses

- ❖ can be difficult to choose the distance measures and weightings
- ❖ can be sensitive to initial parameter choices
- ❖ the clusters found can be difficult to interpret

Association Rules (1)

- ◆ Association Rule (AR) Discovery is often referred to as Market Basket Analysis (MBA), and is also referred to as Affinity Grouping
- ◆ A common example is the discovery of which items are frequently sold together at a supermarket. If this is known, decisions can be made about:
 - ❖ arranging items on shelves
 - ❖ which items should be promoted together
 - ❖ which items should not simultaneously be discounted

Association Rules (2)



CSE5230 - Data Mining, 2004

Lecture 4.23

Usefulness of ARs

- ◆ **Some rules are useful:**
 - ❖ unknown, unexpected and indicative of some action to take.
- ◆ **Some rules are trivial:**
 - ❖ known by anyone familiar with the business.
- ◆ **Some rules are inexplicable:**
 - ❖ seem to have no explanation and do not suggest a course of action.

"The key to success in business is to know something that nobody else knows"

Aristotle Onassis

CSE5230 - Data Mining, 2004

Lecture 4.24

AR Example: Co-Occurrence Table

Customer	Items
1	orange juice (OJ), cola
2	milk, orange juice, window cleaner
3	orange juice, detergent
4	orange juice, detergent, cola
5	window cleaner, cola

	OJ	Cleaner	Milk	Cola	Detergent
OJ	4	1	1	2	2
Cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Cola	2	1	0	3	1
Detergent	2	0	0	1	2

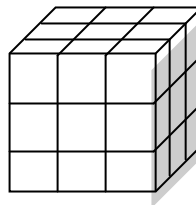
CSE5230 - Data Mining, 2004

Lecture 4.25

Association Rule Discovery Process

◆ A co-occurrence cube would show associations in three dimensions

- ❖ it is hard to visualize more dimensions than that
- ❖ Worse, the number of cells in a co-occurrence hypercube grows exponentially with the number of items: It rapidly becomes impossible to store the required number of cells
- ❖ Smart algorithms are thus needed for finding frequent large itemsets



◆ We must:

- ❖ Choose the right set of items
- ❖ Generate rules by deciphering the counts in the co-occurrence matrix (for two-item rules)
- ❖ Overcome the practical limits imposed by many items in large numbers of transactions

CSE5230 - Data Mining, 2004

Lecture 4.26

ARs: Choosing the Right Item Set

- ◆ **Choosing the right level of detail (the creation of classes and a taxonomy)**
 - ❖ For example, we might look for associations between product categories, rather than at the finest-grain level of product detail, e.g.
 - » “Corn Chips” and “Salsa”, rather than
 - » “Doritos Nacho Cheese Corn Chips (250g)” and “Masterfoods Mild Salsa (300g)”
 - Important associations can be missed if we look at the wrong level of detail
- ◆ **Virtual items may be added to take advantage of information that goes beyond the taxonomy**
- ◆ **Anonymous versus signed transactions**

ARs: What is a Rule?

if condition then result

Note:

if (nappies **and** Thursday) **then** beer

is usually better than (in the sense that it is more actionable)

if Thursday **then** nappies **and** beer

because it has just one item in the result. If a 3 way combination is the most common, then consider rules with just 1 item in the result, e.g.

if (A **and** B) **then** C

if (A **and** C) **then** B

AR: Is the Rule a Useful Predictor? (1)

- ◆ **Confidence** is the ratio of the number of transactions with all the items in the rule to the number of transactions with just the items in the condition. Consider:

if B and C then A

- ◆ If this rule has a confidence of 0.33, it means that when B and C occur in a transaction, there is a 33% chance that A also occurs.

AR: Is the Rule a Useful Predictor? (2)

- ◆ Consider the following table of probabilities of items and their combinations:

Combination	Probability
A	0.45
B	0.42
C	0.40
A and B	0.25
A and C	0.20
B and C	0.15
A and B and C	0.05

AR: Is the Rule a Useful Predictor? (3)

- ◆ Now consider the following rules:

Rule	p(condition)	p(condition and result)	confidence
if A and B then C	0.25	0.05	0.20
if A and C then B	0.20	0.05	0.25
if B and C then A	0.15	0.05	0.33

- ◆ It is tempting to choose “If B and C then A”, because it is the most confident (33%) - but there is a problem

AR: Is the Rule a Useful Predictor? (4)

- ◆ This rule is actually worse than just saying that A randomly occurs in the transaction - which happens 45% of the time
- ◆ A measure called *improvement* indicates whether the rule predicts the result better than just assuming the result in the first place

$$\text{improvement} = \frac{p(\text{condition and result})}{p(\text{condition})p(\text{result})}$$

AR: Is the Rule a Useful Predictor? (5)

- ◆ When *improvement* > 1, the rule is better at predicting the result than random chance
- ◆ The *improvement* measure is based on whether or not the probability $p(\text{condition and result})$ is higher than it would be if *condition* and *result* were statistically independent
- ◆ If there is no statistical dependence between *condition* and *result*, *improvement* = 1.

AR: Is the Rule a Useful Predictor? (6)

- ◆ Consider the improvement for our rules:

Rule	support	confidence	improvement
if A and B then C	0.05	0.20	0.50
if A and C then B	0.05	0.25	0.59
if B and C then A	0.05	0.33	0.74
if A then B	0.25	0.59	1.31

- ◆ None of the rules with three items shows any improvement - the best rule in the data actually has only two items: "if A then B". A predicts the occurrence of B 1.31 times better than chance.

AR: Is the Rule a Useful Predictor? (7)

- ◆ When improvement < 1 , *negating* the result produces a better rule. For example

if B and C then not A

has a confidence of 0.67 and thus an improvement of $0.67/0.55 = 1.22$

- ◆ Negated rules may not be as useful as the original association rules when it comes to acting on the results

AR: Discovering Large Item Sets

- ◆ The term “frequent itemset” means “a set S that appears in at least fraction s of the baskets,” where s is some chosen constant, typically 0.01 (i.e. 1%).
- ◆ DM datasets are usually too large to fit in main memory. When evaluating the running time of AR discovery algorithms we:
 - ❖ count the number of passes through the data
 - » Since the principal cost is often the time it takes to read data from disk, the number of times we need to read each datum is often the best measure of running time of the algorithm.

AR: Discovering Large Item Sets (2)

- ◆ There is a key principle, called *monotonicity* or the *a-priori trick* that helps us find frequent itemsets [AgS1994]:

- ❖ If a set of items S is frequent (i.e., appears in at least fraction s of the baskets), then every subset of S is also frequent.

- ◆ To find frequent itemsets, we can:

1. Proceed level-wise, finding first the frequent items (sets of size 1), then the frequent pairs, the frequent triples, etc.

- ❖ Level-wise algorithms use one pass per level.

2. Find all maximal frequent itemsets (i.e., sets S such that no proper superset of S is frequent) in one (or few) passes

AR: The A-priori Algorithm (1)

- ◆ The A-priori algorithm proceeds level-wise.

1. Given support threshold s , in the first pass we find the items that appear in at least fraction s of the baskets. This set is called L_1 , the frequent items

(Presumably there is enough main memory to count occurrences of each item, since a typical store sells no more than 100,000 different items.)

2. Pairs of items in L_1 become the candidate pairs C_2 for the second pass. We hope that the size of C_2 is not so large that there is not room for an integer count per candidate pair. The pairs in C_2 whose count reaches s are the frequent pairs, L_2 .

AR: The A-priori Algorithm (2)

3. The candidate triples, C_3 are those sets $\{A, B, C\}$ such that all of $\{A, B\}$, $\{A, C\}$ and $\{B, C\}$ are in L_2 . On the third pass, count the occurrences of triples in C_3 ; those with a count of at least s are the frequent triples, L_3 .
 4. Proceed as far as you like (or until the sets become empty). L_i is the frequent sets of size i ; C_{i+1} is the set of sets of size $i + 1$ such that each subset of size i is in L_i .
- ◆ The A-priori algorithm helps because the number tuples which must be considered at each level is much smaller than it otherwise would be.

AR: Strengths and Weaknesses

- ◆ **Strengths**
 - ❖ Clear understandable results
 - ❖ Supports undirected data mining
 - ❖ Works on variable length data
 - ❖ Is simple to understand
- ◆ **Weaknesses**
 - ❖ Requires exponentially more computational effort as the problem size grows
 - ❖ Suits items in transactions but not all problems fit this description
 - ❖ It can be difficult to determine the right set of items to analysis
 - ❖ It does not handle rare items well; simply considering the level of support will exclude these items

References

- ◆ [JMF1999] A. K. Jain, M. N. Murty and P. J. Flynn, Data clustering: a review, *ACM Computing Surveys*, Volume 31 , Issue 3, pp. 264-323, 1999.
- ◆ [BeL1997] Michael J. A. Berry and Gordon Linoff, *Automatic Cluster Detection*, Ch. 10 in *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, 1997.
- ◆ [BeL1997a] Michael J. A. Berry and Gordon Linoff, *Market Basket Analysis*, Ch. 8 in *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, 1997.
- ◆ [AgS1994] Rakesh Agrawal and Ramakrishnan Srikant, *Fast Algorithms for Mining Association Rules*, In Jorge B. Bocca, Matthias Jarke and Carlo Zaniolo eds., VLDB'94, Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, pp. 487-499, September 12-15 1994