DiverseRank in Classification and Clustering

Mayank Gupta, Ankush Jain

Introduction - Problem

Exploring **Diverse-Frequent Patterns** in **Classification**

- **Diversity**
  - Measure of variety
  - {Bread, Butter, Milk} vs {Soap, Chocolate, Rice}
- **Classification**
  - Create a model using labelled training data
  - Predict the class label of test data
- **Clustering**
  - Divide test data into clusters based on some distance metric

Background: Frequent Patterns

- **Problem**: To efficiently find patterns that occur frequently in a large database
- Useful in marketing, web usage mining, intrusion detection, product warehousing etc.
- **Example**: Given a list of all the transactions made in a grocery store, figure out which items are frequently bought together

Background: Example

{Bread, Butter, Eggs, Orange}
{Bread, Butter, Eggs, Apple}
{Bread, Eggs, Battery, Milk, Tea
{Bread, Eggs, Battery, Cherry}
{Butter, Diapers, Hair Spray, Whiskey}
{Butter, Diapers, Hair Spray}

Patterns

{Bread, Eggs}
{Bread, Apple}
{Bread, Butter, Eggs}
{Bread, Eggs, Battery}

But {bread, eggs} is much more important!!

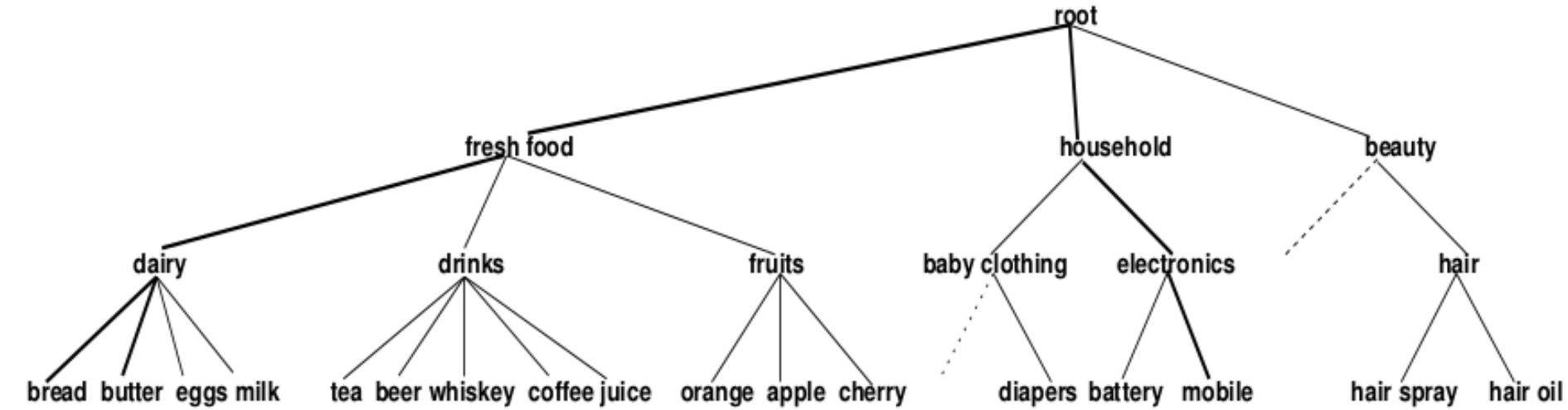
- Support (of a pattern)
  - (No. of transactions a rule appears in) / (Total number of transactions)
  - Patterns with support greater than minSupport are called frequent patterns

Background: Association Rules

- Every pattern can be converted to a rule
  - {bread, butter, eggs}
    - {bread} => {butter, eggs}
    - {butter, eggs} => {bread}
    - ...
- Lots of frequent patterns are generated, how to find more important ones?
  - Confidence
    - Confidence(A => B)
      - Support(A union B) / Support(A)
    - Filter rules by minConfidence

DiverseRank

- DiverseRank - measure of diversity
- A data structure called concept hierarchy is used
- Patterns are more diverse if their root is farther from leaf nodes



DiverseRank - Description

Depends on Level Factor (LF), Merging Factor (MF), and Adjustment Factor(AF)

- **MF -** Measure of the number of parents the children merged into
- **LF -** Gives higher weightage to merging at higher levels (since that implies more diversity)
- **AF -** Used to compensate the effect of dummy nodes added to an unbalanced concept

Related Work - I

- *Fast Algorithms For Mining Association Rules, Agarwal et al*
  - Introduces frequent patterns and a basic algorithm for mining them (Apriori algorithm)
- *Mining Frequent Patterns without Candidate Generation, Jiawei Han et al*
  - *Describes an efficient algorithm for mining frequent patterns (FP-Growth)*

Related Work - II

- *Discovering Diverse-Frequent Patterns In Transactional Databases, Somya Srivastava et al*
  - Introduces the notion of diversity
  - Proposes a measure called DiverseRank
- *Extracting Diverse-Frequent Patterns with Unbalanced Concept Hierarchy, Kswamy et al*
  - Improves upon the previous paper, proposes algorithms to counter the limitation of balanced trees required by the previous paper

Work Done - Sem 1

- Read the relevant research papers
- Understood the related concepts - frequent patterns, common algorithms (*apriori* and *FP-growth*), DiverseRank (with both *balanced* and *unbalanced* concepts)
- Implemented the Apriori algorithm
  - Tested it against standard implementations
- Implemented DiverseRank
  - Tested it using CHT given in research papers

Work Done - Sem 1

- Read the relevant research papers
- Understood the related concepts - frequent patterns, common algorithms (*apriori* and *FP-growth*), DiverseRank (with both *balanced* and *unbalanced* concepts)
- Implemented the Apriori algorithm
  - Tested it against standard implementations
- Implemented DiverseRank
  - Tested it using CHT given in research papers

Classification

- Frequent pattern based classification
  - We try to mine rules to predict the class of unseen data
- Rules are of the form

$$(AT_{i1} = x_{i1}) \wedge (AT_{i2} = x_{i2}) \wedge \ldots \wedge (AT_{in} = x_{in}) \rightarrow p_{i1}$$

where AT_i represents an attribute and p_i represents a class label

Classification - Example

Table 2. Classification Data Set

Client #	Name	Current Job/year	Income	Criminal History	Loan
1	Sam	5	35K	No	Yes
2	Sara	1	40K	No	No
3	Smith	10	55K	Yes	No
4	Raj	5	40K	No	Yes
5	Omar	1	35K	No	No
6	Sandy	2	25K	No	No
7	Kamal	6	40K	No	Yes
8	Rony	5	34K	No	Yes

Table. 3. Sample of unclassified data set

Client #	Name	Current Job/year	Income	Criminal History	Loan
24	Raba	3	50K	No	?
25	Samy	3	14K	No	?
26	Steve	25	10K	Yes	?
27	Rob	0	45K	No	?

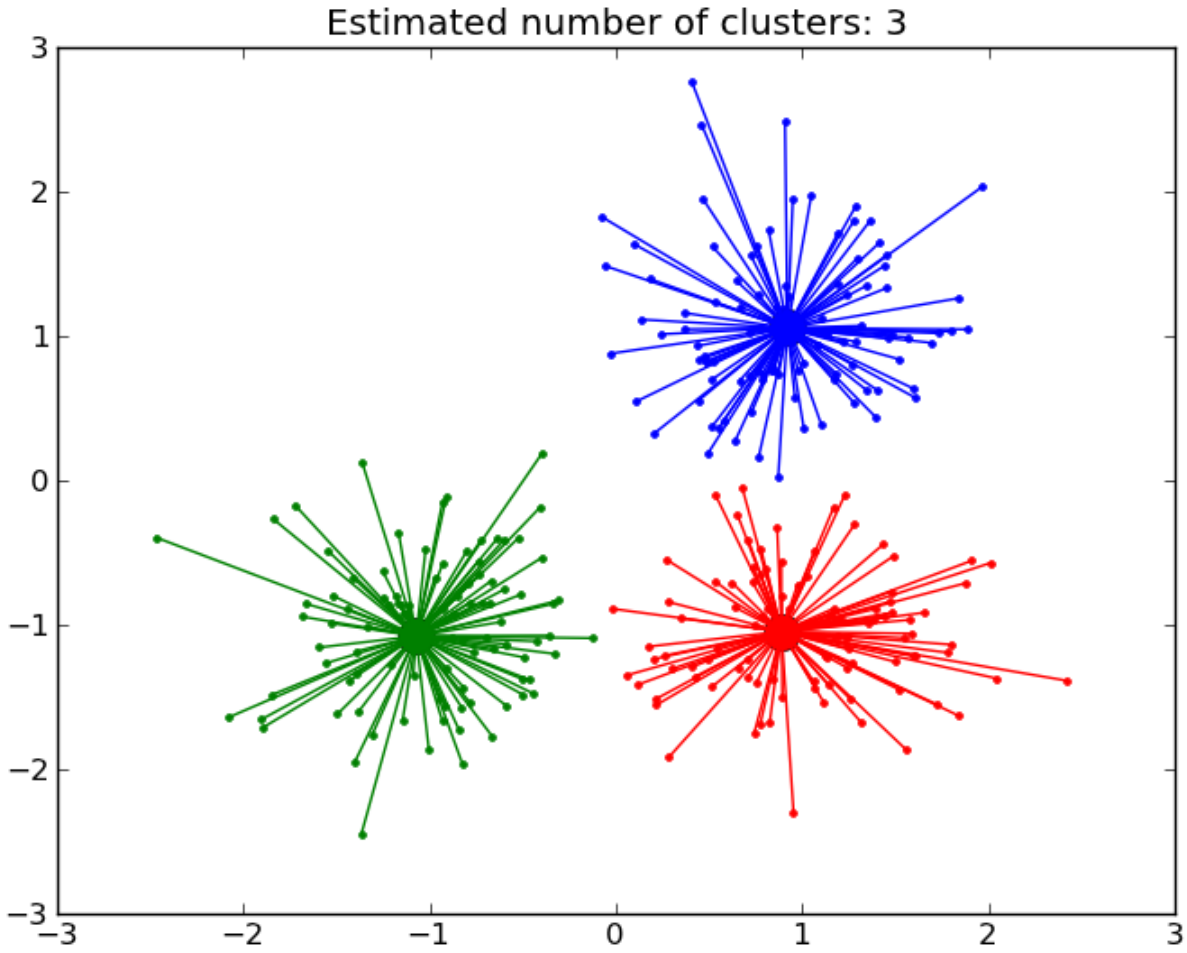
Classification - Diverse Frequent Patterns

- We use frequent pattern mining to find out association rules
- These rules are then used to predict classes
- Different algorithms differ on how exactly these rules are selected and combined
  - C4.5
  - CBA
  - CPAR
  - CMAR

Clustering

- Divide data points into groups of clusters of similar elements
- Clustering is good if -
  - Data points within clusters are very similar
  - Data points across clusters are very different
- Clustering is form of unsupervised classification
  - Operates on unlabelled data
- Simplest algorithm: *k-means*

Clustering - Example



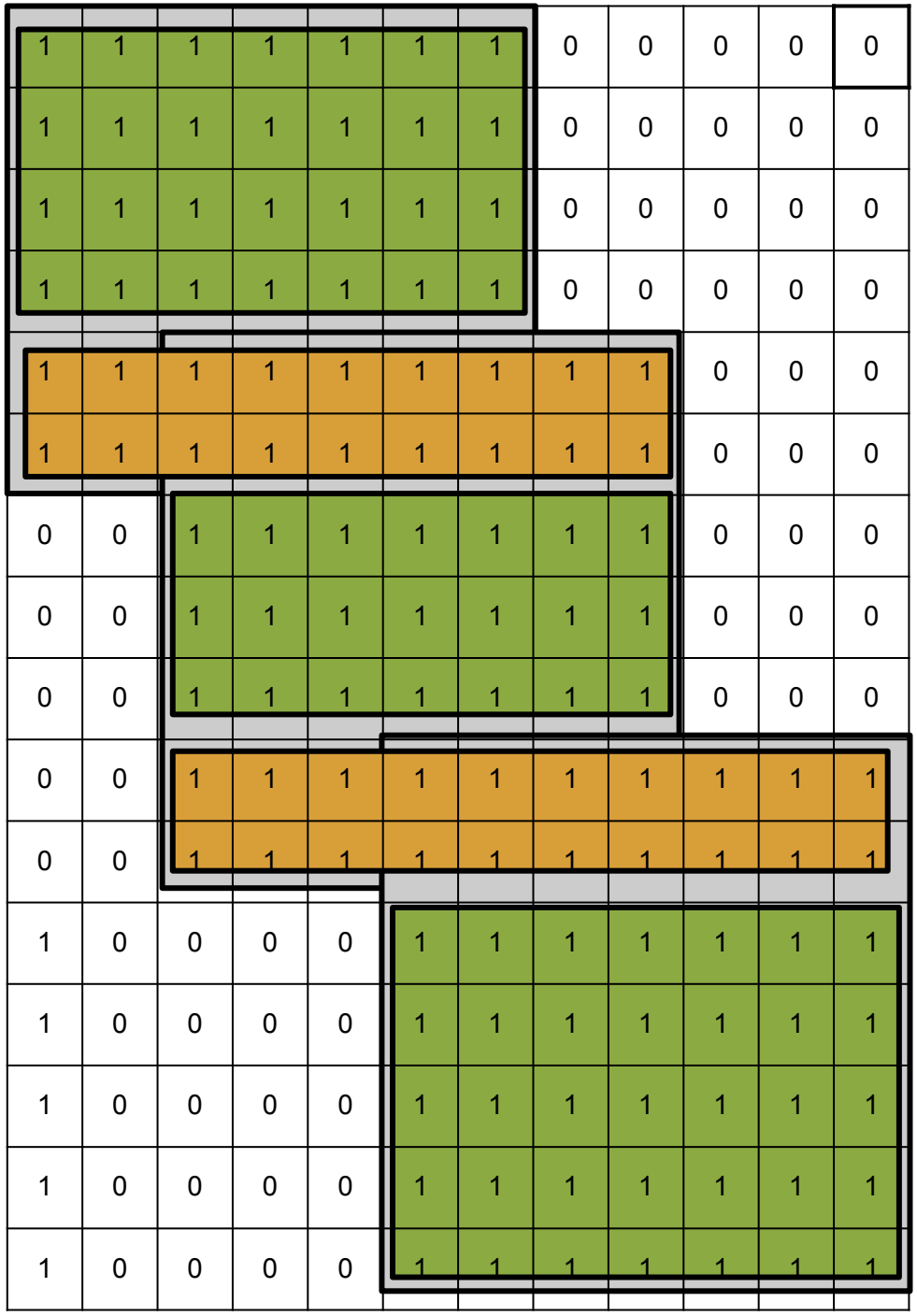
Clustering - Challenges

- Curse of dimensionality
  - Simple algorithms like *k-means* perform poorly with high-dimensional data
  - Distances between points not meaningful, poor results
- Solution
  - Find frequent patterns in transactions
  - Use the strongest frequent patterns to cluster data

Diversity and Clustering

How does diversity help clustering?

- Key idea - *find the odd one out*
  - A member of a cluster with very different characteristics from that cluster is likely not a part of the cluster
  - What characteristics?
    - Diversity!
- As clusters are iteratively formed, find if elements with different diversity exist
  - Remove them from cluster
  - Add to the pending items

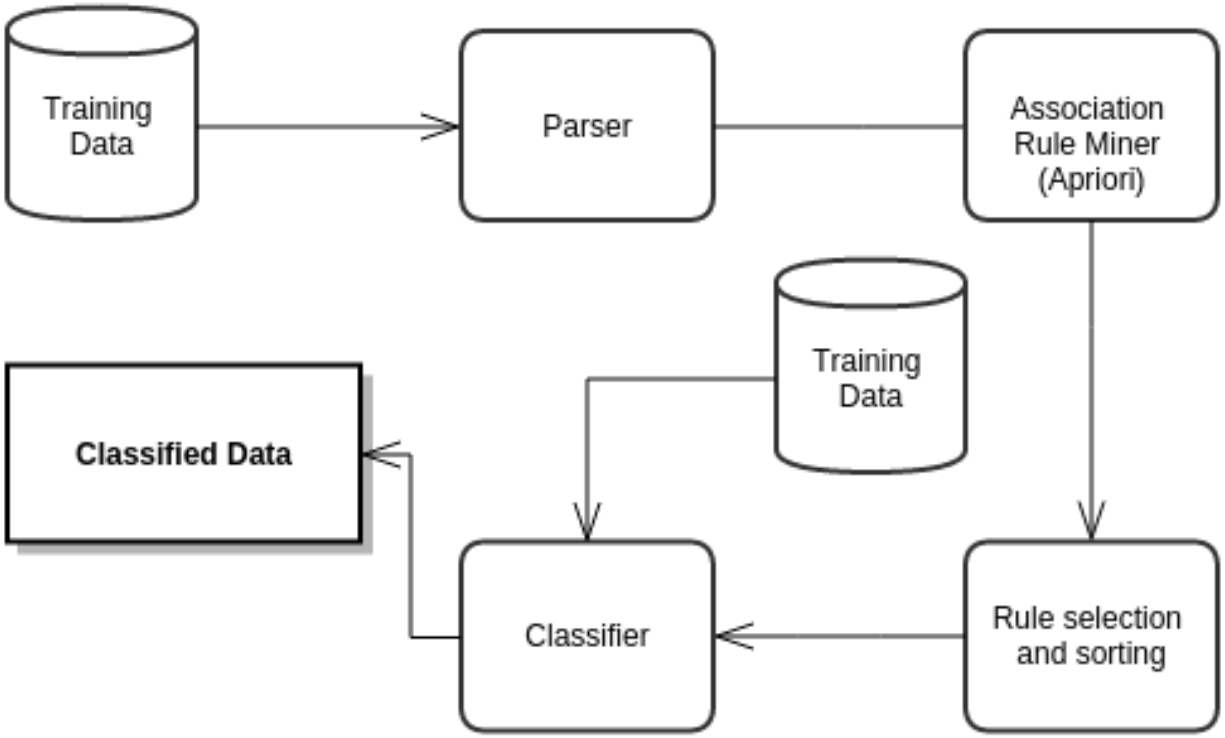

1 1 1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 1 1 0 0 0
1 1 1 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 1 1 1 1
0 0 1 1 1 1 1 1 1 1 1 1
1 0 0 0 0 1 1 1 1 1 1 1
1 0 0 0 0 1 1 1 1 1 1 1
1 0 0 0 0 1 1 1 1 1 1 1
1 0 0 0 0 1 1 1 1 1 1 1
1 0 0 0 0 1 1 1 1 1 1 1

Work done after Evaluation II

- Read relevant literature
  - Demand-driven Associative Classification - *Valoso, Meira Jr.*
  - Integrating Classification and Association Rule Mining - *Liu, Hsu et al*
    - Describes classification approaches using frequent pattern based mining
  - Iterative Projected Clustering by Subspace Mining *Yiu, Mamolis et al*
    - Describes clustering approaches using frequent pattern based mining

Work done after Evaluation II

- Wrote a classifier based on Association Rule mining



Work done after Evaluation II

- Tested the classifier against UCI datasets
  - Converted datasets to transaction dataset
  - Dataset used - *cars*
    - Tries to predict desirability of cars based on features
- Read up on various methods of Association Rule selection
  - C4.5
  - CBA
  - CPAR
  - CMAR

Work done after Evaluation II

- Implemented a clustering system
  - Takes as input dataset in transactional form
  - Used Apriori algorithm to find frequent patterns
  - Iteratively forms clusters using the strongest of the mined patterns
- Extended the clustering algorithm to consider diversity
  - After each iteration, data points having diversity different (*outliers)* from rest of the cluster are removed

Clustering - evaluation

- **N**umber of clusters is not pre-determined in *iterative frequent pattern based clustering*
- Evaluation measure -

```
foreach cluster in predicted_clusters:
    for all pairs of i, j in cluster:
        numPairs++
        if i and j share the same cluster originally:
            truePositive++
```

- Accuracy - truePositive/numPairs

Clustering - results

- Without Diversity
  - 1293 out of 1334 pairs predicted correctly
  - 96.9% accuracy
- With Diversity
  - Varies with the kind of filtering mechanism used
  - One approach
    - 1214 pairs out of 1276 pairs predicted correctly
    - 95.14% accuracy
  - Another approach
    - 1700/1727 pairs predicted correctly
    - 98.43% accuracy
- **Caveat** - evaluation metric is biased

Future Plan

- Investigate into high clustering
  - Try tuning various constants like beta, minimum cluster size
  - Try tweaking the diversity selection method
  - Concept hierarchy
- Use more evaluation measures
- Test using more datasets and draw conclusions