DiverseRank in Classification and Clustering

Mayank Gupta, Ankush Jain

Introduction - Problem

Exploring **Diverse-Frequent Patterns** in **Clustering**

- **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!!

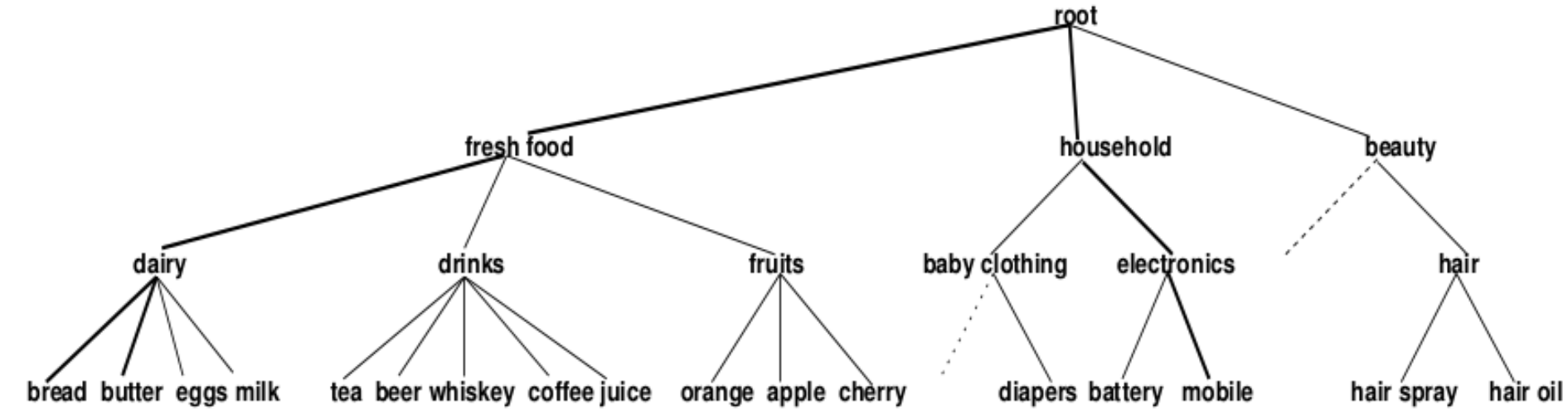
- Support (of a pattern)
  - (No. of transactions a rule appears in) / (Total number of transactions)
  - Patterns with support greater than minSupport are called frequent patterns

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

DiverseRank - New

- e’ - Number of edges in minimum subtree containing items
- e - Number of edges in tree of MINIMUM size that contains these items
- E - Number of edges in tree of MAXIMUM size that contains these items

$$dr(Y, C) = \frac{e' - e}{E - e}$$

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	?

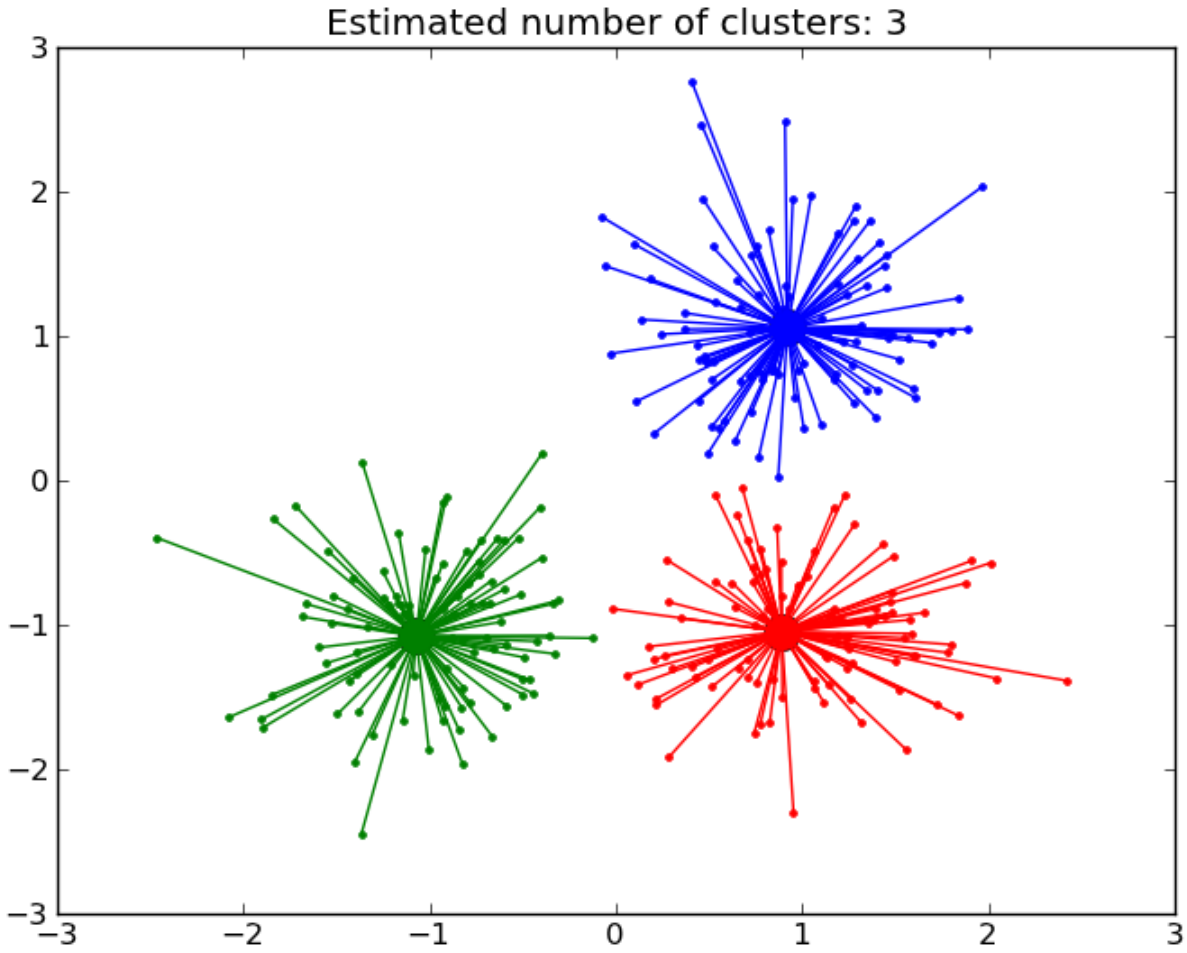
Classification - Diverse Frequent Patterns

- 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

Outlier Detection - DSRD

Given a cluster *{o_1, o_2, …, o_n}* with diversity values *{d_1, d_2, …, d_n}*, we define a measure *dsrd.*

$$dsrd(k) = \sqrt{\sum_{j=1}^{n} (d_k - d_j)^2}$$

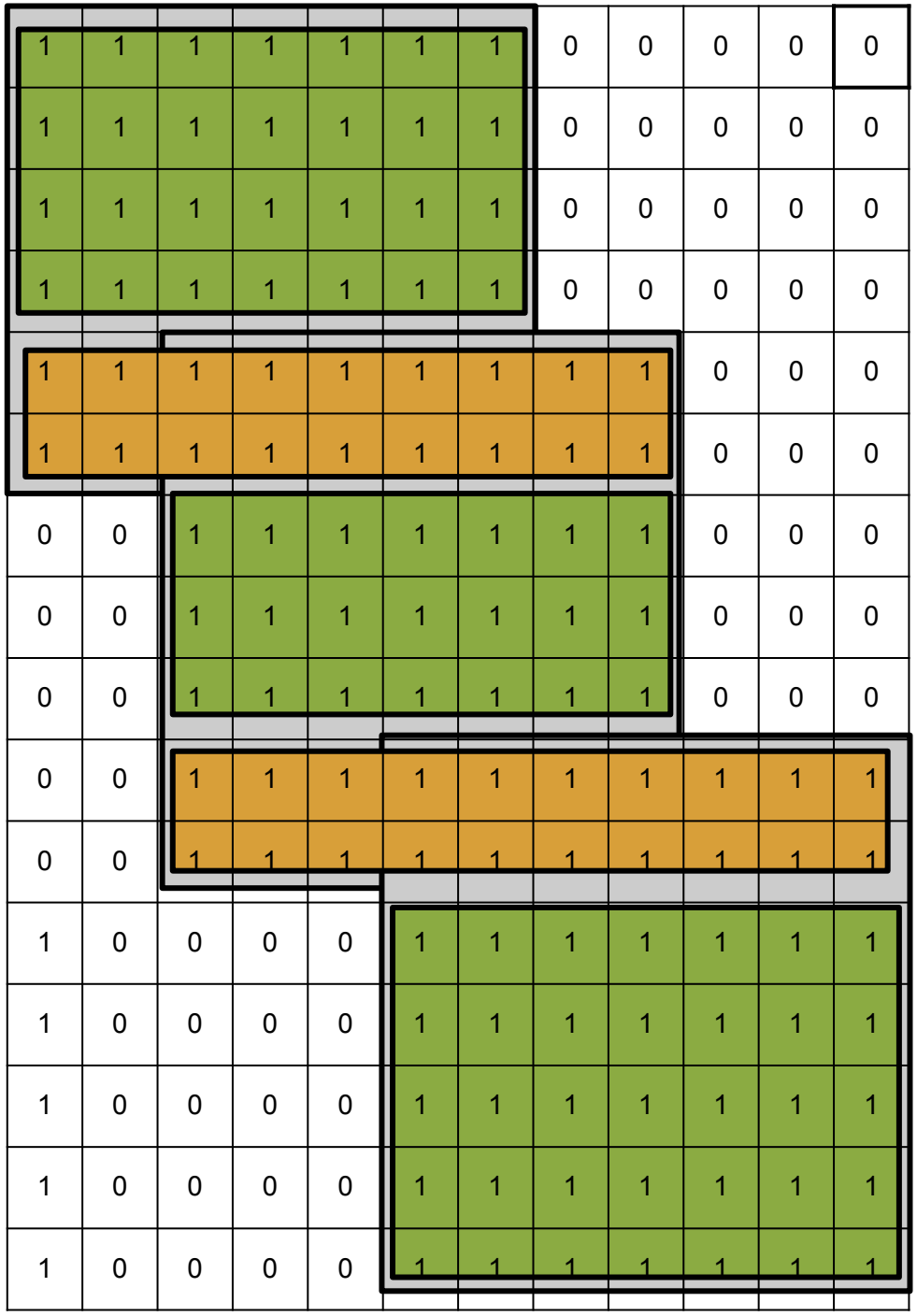
All elements with a *dsrd* value greater than *delta* are removed.

Outlier Detection - Binning

- We sort the cluster according to diversity values
- At each point where the difference between adjacent diversity values exceeds *delta*, a partition is marked
  - Partition divides current section into two
  - n partitions form n + 1 sections of the cluster
- The longest cluster so formed is selected and others are removed

Why Diversity works

Trans1	**Whole Milk**	**2% Milk**	**Pepsi**	Mango Milk	Badam Milk	Kesar Milk
Trans2	**Whole Milk**	**2% Milk**	**Pepsi**	Pepsi	Coke	Badam Milk
Trans3	**Whole Milk**	**2% Milk**	**Pepsi**	Lassi	Curd	Mango Milk
Trans4	**Whole Milk**	**2% Milk**	**Pepsi**	Orange Juice	Apple Juice	Mixed Fruit Juice
Trans5	**Whole Milk**	**2% Milk**	**Pepsi**	Shampoo	Soap	Hair Oil

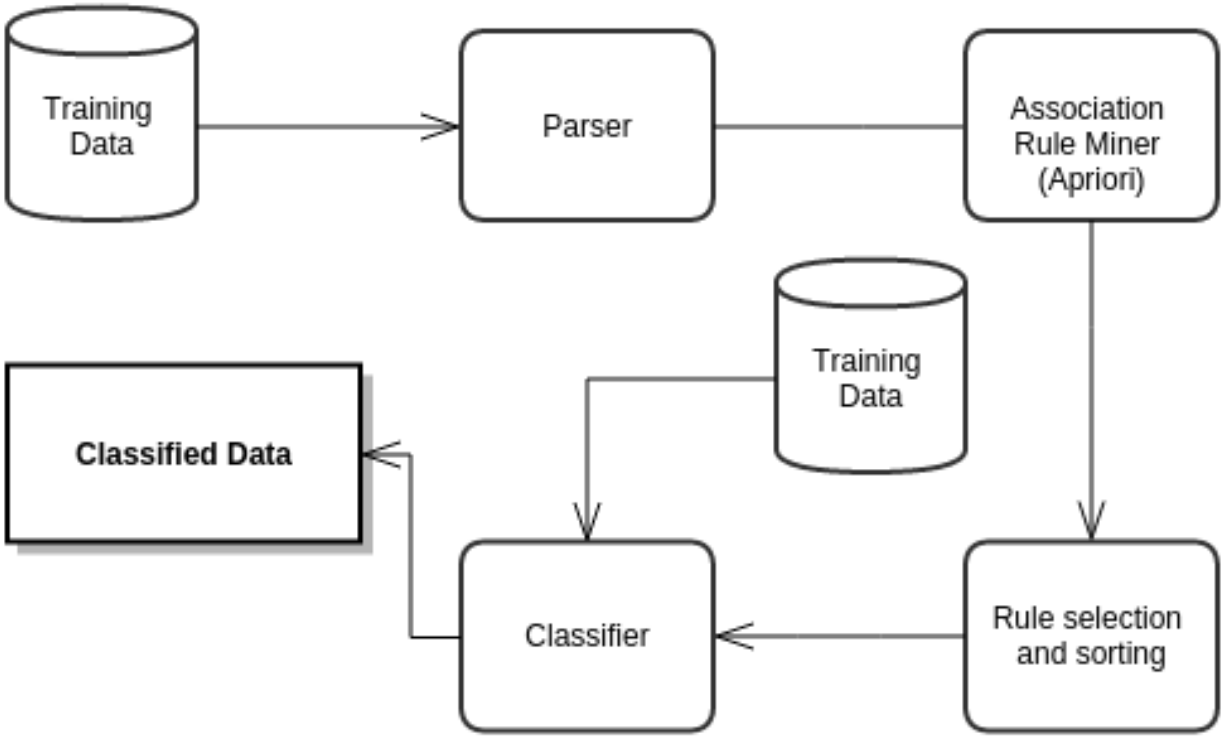

1 1 1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 1 1 0 0 0
1 1 1 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 1 1 1 1
0 0 1 1 1 1 1 1 1 1 1 1
1 0 0 0 0 1 1 1 1 1 1 1
1 0 0 0 0 1 1 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 - Sem 2

- 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 - Sem 2

- Wrote a classifier based on Association Rule mining



Work done - Sem 2

- 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 - Sem 2

- 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

Work done - Sem 2

- Implemented a merging stage for both original/diversity approaches
- Conducted extensive testing using standard UCI datasets (both with and w/o merging)
  - Iris, Seed, Zoo, Water
- Implemented a newer version of the Diversity approach and computed results

Clustering - evaluation

- Number of clusters is not pre-determined in *iterative frequent pattern based clustering*
- Calculate True Positives, True Negatives, False Positives, False Negatives
- Accuracy = TP + TN/(Total)
- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)

Clustering - evaluation

- Evaluation measure -

```
foreach pair(a, b) in dataset:
    if a and b lie in the same natural cluster:
        if they are predicted to be in the same cluster:
            truePositive++;
        else:
            falseNegative++;
    else:
        if they are predicted to be in the same cluster:
            falsePostive++;
        else:
            trueNegative++;
```

Clustering - results

Accuracy without merging (Diversity Old)

Dataset	MinClus	DSRD	Binning
Iris	82.2	85.58	85.58
Seed	79.11	78.96	78.96
Zoo	88.33	88.55	81.18
Water	60.23	62.28	60.24

Accuracy with merging (Diversity Old)

Dataset	MinClus	DSRD	Binning
Iris	82.78	91.80	91.80
Seed	82.51	83.21	83.21
Zoo	91.02	94.65	90.47
Water	60.29	59.80	60.50

Clustering - results

Accuracy without merging (Diversity New)

Dataset	MinClus	DSRD	Binning
Iris	82.2	82.43	77.54
Seed	79.11	79.96	75.06
Zoo	88.33	88.33	85.65
Water	60.23	63.35	60.27

Accuracy with merging (Diversity New)

Dataset	MinClus	DSRD	Binning
Iris	82.78	90.94	93.24
Seed	82.51	85.00	83.41
Zoo	91.02	91.96	95.78
Water	60.29	62.05	60.18

Conclusion

- Diversity can improve clustering performance, IF
  - A logical concept hierarchy is possible for the dataset (can not be done for all datasets)
- Newer diversity approach performs better
  - Binning + newer Diversity + Merging performs better
- Future possibilities
  - Combine DSRD and Binning to form a hybird approach