Exploring diverse-frequent patterns in classification

Mayank Gupta - 201101004
Ankush Jain - 201101010

Introduction

- Exploring **Diverse Patterns** in **Classification**
  - **Diversity**
    - Measure of variety
    - {Bread, Butter, Milk} vs {Soap, Chocolate, Rice}

  - **Classification**
    - Learn from some training data, and create a model that predicts the category of an entity

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

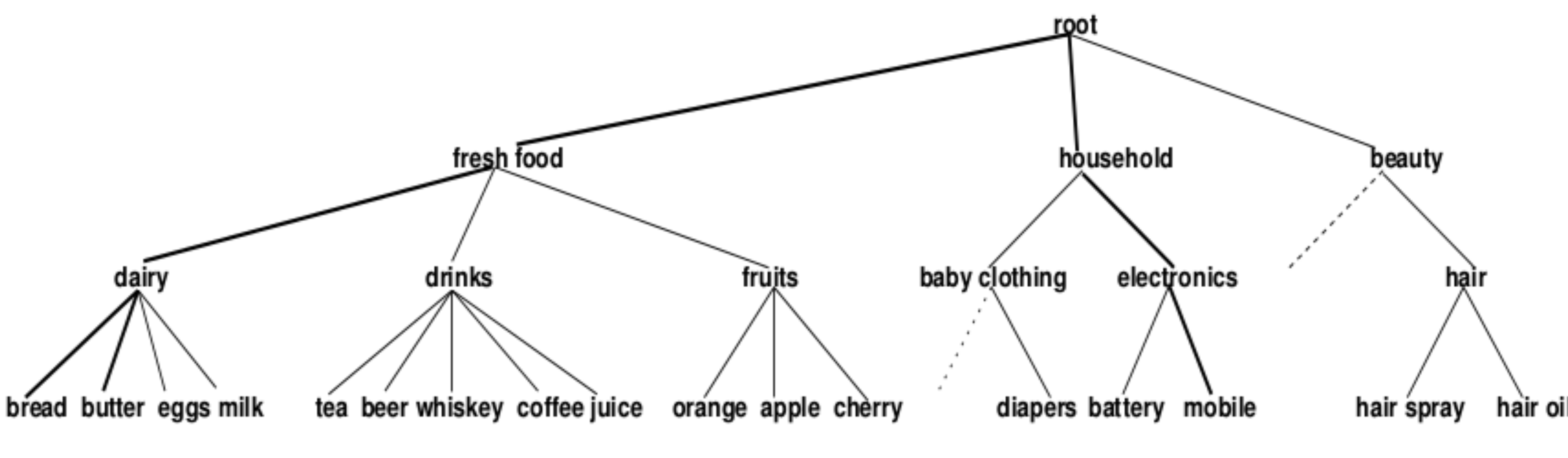
- Every pattern can be converted to a rule
  - {bread, butter, eggs}
    - {bread} => {butter, eggs}
    - {butter, eggs} => {bread}
    - ...
- Lots of frequent patterns are generated, how to find more important ones?
  - Confidence
    - Confidence(A => B)
      - Support(A union B) / Support(A)
 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

Issues

- Even filtering by minSupport and minConfidence produces a lot of patterns
- A better method of telling “which pattern is more interesting” is needed
  - {Bread, Butter, Eggs} is not very interesting
  - Hence, notion of diversity

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

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	?

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

Problem Definition

Exploring Diverse-Frequent patterns in classification

DiverseRank - Measure of diversity

Classification - Categorisation of data

Our task is to explore how the concept of DiverseRank can be extended to classification

Work Done Before Viva 1 - I

- Read the various research papers in the field, and understood the basic concepts about
  - Frequent Patterns
  - Association Rules
  - Apriori Algorithm
  - FP-Growth Algorithm
  - Diversity/DiverseRank

Work Done Before Viva 1 - II

- Implemented the Apriori Algorithm in Python
  - Takes CSV transaction files as input
  - Finds frequent patterns of all lengths with support greater than minSupport
  - For each frequent pattern, finds rules with confidence greater than minConfidence
- Tested the Apriori implementation on large datasets (100,000 transactions)
  - Dataset Source - FIMI Repository
- Verified results against standard implementation

Work Done After Viva 1

- Research papers
  - *Extracting Diverse-Frequent Patterns with Unbalanced Concept Hierarchy, M. Kumaraswamy et al*
    - Proposes algorithms to counter the limitation of balanced trees required by the previous paper
- Implementation of DiverseRank
- Extending the above implementation to support unbalanced concept hierarchy

Work Ongoing

- Test our implementation of DiverseRank against standard datasets and verify results
- Read up on classification

Project Plan

- Read up on classification from the book "Demand Driven Associative Classification"
- Understand how the concept of Frequent-Pattern based Association Rules can be extended to classification
- Extend the DiverseRank implementation and explore its usage in classification

Deliverables

- The extended DiverseRank implementation, with classification features
- Results and analysis

References

- Demand-Driven Associative Classification - *Adriano Veloso, Wagner Meira Jr.*
- Classification based on Associative Rule Mining Techniques - *“Alaa Al Deen” Mustafa Nofal and Sulieman Bani-Ahmad*

Thank You