Exploring diverse-frequent patterns in classification

Ankush Jain (201101010)
Mayank Gupta (201101004)

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

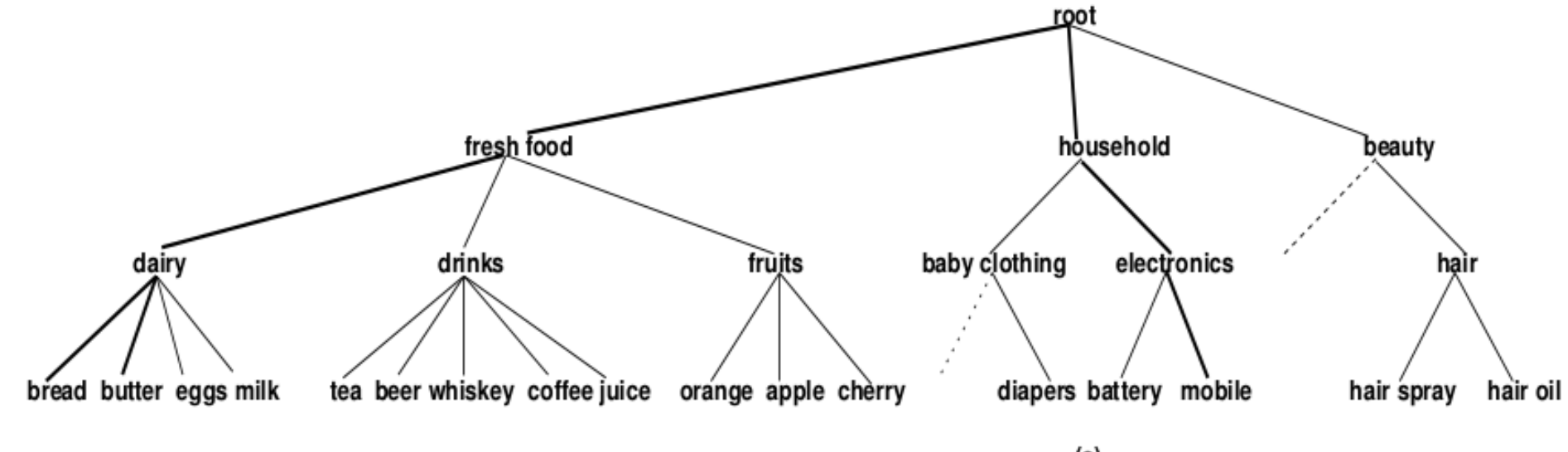
- Every pattern can be converted to a rule
  - {bread, butter, eggs}
    - {bread} => {butter, eggs}
    - {butter, eggs} => {bread}
    - ...
- Lots of frequent patterns are generated, how to find more important ones?
  - Confidence
    - Confidence(A => B)
      - Support(A union B) / Support(A)
 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



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 - 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 - 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

Project Plan

- 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 with Unbalanced Concept Hierarchy (the above paper)
- Extend the DiverseRank implementation and explore its usage in classification

Deliverables

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

Thank You