DiverseRank in classification

Ankush Jain (201101010)
Mayank Gupta (201101004)

Introduction: 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.
- Two main algorithms - Apriori and FP-Growth
- Support(A->B) - (# Transactions containing A & B)/(Total Transactions)
- Confidence(A->B) - Support(A U B) / Support(A)

Apriori Algorithm

- Simplest algorithm
- Initially, k = 0, and bigItemSet = itemSet
- While bigItemSet is not empty:
  - k = k + 1
  - bigItemSet = [Combinations of items of length k]
  - bigItemSet = FilterByMinSupport(bigItemSet)
  - frequentItems.append(bigItemSet)
- Inefficient, in both space and time

FP-Growth Algorithm

- Revolves around a data structure called FP-Tree
- FP-Tree: highly compact structure, possible to fit in memory
- Requires only two passes over data, instead of k passes required by Apriori

Diversity

- Both algorithms rank patterns by support
- Important to distinguish patterns by diversity, for some applications
- Diversity - measure of the number of categories, items of an item-set belong to
- DiverseRank - a parameter to measure diversity

DiverseRank

- Merging Factor and Level Factor
- Diverse Rank: f(MF, LF)

$$DR(Y) = \sum_{l=h-1}^{s+1} PLF(l) * MF(Y,l)$$

By replacing the corresponding equations, we get the formula for *DiverseRank*.

$$DR(Y) = \sum_{l=h-1}^{s+1} \left[\frac{2*(h-l)}{(h-1)*h} \right] * \left[\frac{|GFP(Y,l)|-1}{|GFP(Y,l+1)|-1} \right]$$

Work Done - I

- Read primary research papers in the field
  - *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)*

Work Done - II

- Research papers contd...
  - *Discovering Diverse-Frequent Patterns In Transactional Databases, Somya Srivastava et al*
    - Introduces the notion of diversity
    - Proposes a measure called DiverseRank
    - Describes an algorithm to mine patterns using DiverseRank, and proposes optimizations to improve its performance

Work Done - III

- 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
- Limitations
  - In-memory, cannot handle large files

Work Ongoing - I

- Test our Apriori implementation on large datasets
  - Current implementation does everything in-memory
  - Might have problems with large datasets
- Verify the results against standard tools
- Check performance

Work Ongoing - II

- Research papers
  - *Extracting Diverse-Frequent Patterns with Unbalanced Concept Hierarchy, M. Kumaraswamy et al*
    - Improves upon the previous paper, proposes algorithms to counter the limitation of balanced trees required by the previous paper

Problem Definition

Exploring DiverseRank 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

Project Plan

- Read and properly analyze the important research papers (3 done, 1 ongoing)
- Implement the Apriori algorithm and test against large datasets (Implementation done, testing ongoing)
- Attack the problem of using DiverseRank in classification, and compile results