Exploring diverse-frequent patterns in classification

Mayank Gupta (201101004)
Ankush Jain (201101010)

1. Introduction

Frequent patterns are those that occur commonly in some data. They are commonly used in data mining and can isolate interesting patterns out of large datasets.

Traditionally, the importance of frequent patterns is gauged by their frequency of occurrence in the data. But this may lead to a large number of patterns, not all of which may be interesting. Researchers have proposed other metrics to measure the *interestingness* of a pattern, one of which is DiverseRank. DiverseRank tries to differentiate between patterns based on the amount of diversity in each of them.

Frequent patterns are also used in classification - patterns are found and converted into association rules, from which classes of unseen data can be predicted. The order of importance of these patterns is normally decided by some measures that reflect their frequency.

In our project, we try to explore if the notion of diversity can be somehow extended to the problem of classification.

2. Problem

a. Background - Frequent Patterns

Given a set of transactions, such as:

$Butter, Eggs, Milk, Orange$
$Bread, Butter, Eggs, Apple$
$Bread, Eggs, Battery, Milk, Tea$
$Bread, Eggs, Battery, Cherry$
$Butter, Diapers, Hair\ Spray, Whiskey$

The problem of mining frequent patterns is to find itemsets that exist in many transactions. For example, $Bread, Eggs$ is a pattern that occurs in many transactions.

For each pattern, we define a measure called *support* -

$$Support = \frac{Number\ of\ transactions\ a\ pattern\ appears\ in}{Total\ number\ of\ transactions}$$

Patterns whose *support* exceeds a user-defined threshold called *minSupport* are called frequent patterns.

These patterns are used to generate association rules. An association rule is defined as an implication of the form

$$A \Rightarrow B$$

where $A \cup B$ is a frequent pattern

We define another measure called *confidence* for each rule defined as -

$$Confidence = \frac{Support(A \cup B)}{Support(A)}$$

b. Background - Classification

Classification means to classify unseen data into some classes based on rules/observations that are automatically generated using some seen data.

Formally said, given records of the form:

$a_1 = v_1 \wedge a_2 = v_2 \wedge \ldots \wedge a_n = v_n$

and class labels $\{p_1, p_2, \ldots, p_m\}$,

the task is to assign one of the class labels to the record using some pre-classified data as a learning model.

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	?

An example classification data set

To create a learning model using association rules, we mine rules of the form $A \Rightarrow B$ where B is one of $\{p_1, p_2, \ldots, p_m\}$ (class labels). We then prioritize between the rules using a combination of *support* and *confidence*.

We then use these rules to predict the classes of unseen data, applying the rules in the order of priority until one fits.

In our project, we try to understand if classification is improved if, instead of prioritizing rules by *support* and *confidence* only, we also try to incorporate some percentage of diversity in our prioritizing function.

3. Progress till Viva 1

We got acquainted with the basics of frequent patterns and association rules and studied the following papers.

1. *Fast Algorithms For Mining Association Rules - Agarwal et al*
2. *Mining Frequent Patterns without Candidate Generation - Jiawei Han et al*
3. *Discovering diverse-frequent patterns in Transactional Databases - Kumaraswamy et al*

and studied the following algorithms.

1. Apriori algorithm
2. FP-Growth algorithm
3. Diversity/DiverseRank

We also implemented the Apriori algorithm in Python, and tested it against standard implementations. For this purpose, we used a FIMI repository dataset consisting of about 100,000 transactions.

4. Progress till Viva 2

We explored diversity further and understood DiverseRank in more detail. We studied the following paper:

1. *Extracting Diverse-Frequent Patterns with Unbalanced Concept Hierarchy - Kumaraswamy et al*

This paper proposed methods to remove the limitation imposed by the previous paper on the concept hierarchy (it had to be balanced).

We also implemented a version of DiverseRank in Python. We later extended that implementation to support unbalanced concept hierarchies as well.

5. Progress after Viva 2

We tested our implementation of DiverseRank against standard implementation.

We read up on classification from various sources including “Demand-driven Associative Classification” and “Data Mining: Concepts and Techniques”.

We encountered some issues while creating concept hierarchies for the classification data set, but eventually solved the issue after some discussions with Kumaraswamy sir.

We have progressed significantly with the implementation of an association rule-based classifier.

6. Future work

1. Complete the implementation of the classifier.
2. Take some standard datasets and analyse the performance of the classifier with various combinations of diversity + confidence.
3. Infer the reasons behind the improvement/regression
4. Tweak and improve the concept hierarchy for the test datasets

7. Conclusions

While it would be premature to offer concrete conclusions without thorough testing and analysis, preliminary observations suggest that introducing *diversity* could lead to slight gain in classification performance, although it would somewhat depend on the suitability of the concept hierarchy and the nature of the test data. The exact weightage that would be assigned to diversity scores and support/confidence scores would be decided once we begin with the analysis stage.

8. References

a. Fast Algorithms For Mining Association Rules, Agarwal et al
b. Mining Frequent Patterns without Candidate Generation, Jiawei Han et al
c. Discovering Diverse-Frequent Patterns In Transactional Databases et al
d. Extracting Diverse-Frequent Patterns with Unbalanced Concept Hierarchy, Kumaraswamy et al
e. Demand-driven Associative Classification
f. Classification based on Associative Rule Mining Techniques - “Alaa Al Deen” Mustafa Nofal and Sulieman Bani-Ahmad
g. Classification based on Association Rule Mining Techniques: A General Survey And Empirical Comparative Evaluation - “Alaa Al Deen” Mustafa Nofal