Notes of Pattern Discovery in Data Mining

Jiawei Han

May 12, 2015

1 Week 1 on May 11 2015

1.1 Introduction

1. Data mining can be viewed from multiple angles, i..e, multi-dimensional view
   - Data to be mined
   - Knowledge to be mined
   - Methodologies or techniques utilized
   - Applications adapted
2. From data view
   - Structured and semi-structured data, e.g., relational data/object-relational data, data warehouse data, and transactional data
   - Unstructured data, e.g., text and web data, spatial and spatial-temporal data, multi-media data, data streams and sensor data, time-series data, graphs, social networks and information networks
3. From knowledge view
   - Data summary in multidimensional space, e.g., data cube and OLAP (online analytical processing)
   - Pattern discovery, e.g., frequent patterns, association and correlations
   - Classification and predictive modeling
   - Cluster analysis
   - Outlier analysis
   - Trend and evolution analysis
4. From methodology or technique view: DM is a multi-disciplinary subject, including statistics, machine learning, pattern recognition, visualization, algorithms, database technology, and distributed/cloud computing.
5. From application view
   - Mining text data and web pages
   - Mining business data

- Mining biological and medical data
- Mining social and information networks
- Invisible data mining

1.2 Basic concepts of pattern discovery

1. Patterns are a set of items, subsequences, and substructures that frequently occur together in a data set.

2. Pattern discovery aims to find the inherent regularities in a data set and it is the foundation of many essential data mining tasks, e.g., association, correlation, causality analysis, sequential patterns, classification (discriminative pattern-based analysis).

3. Conventional association rules has the format of $X \to Y$, where support probability of this rule is $X \cup Y$ and the confidence is $\frac{P(X \cup Y)}{P(X)}$. Note that the support of the rule is denoted as $P(X \cup Y)$, meaning the count of both X and Y occurs, but not $X \cap Y$, as it will be $\emptyset$.

4. Compression forms

   - A pattern (itemset) X is a closed-pattern if X is frequent, and there is no super-pattern $Y \supset X$, with the same support of X. A closed pattern is a lossless compression of frequent patterns. The sub-pattern of a closed pattern has the same support.
   - A pattern is a max-pattern if X is frequent and there is no frequent super-pattern $Y \supset X$. It is a lossy compression of frequent patterns and the sub-pattern of a max-pattern is unknown.
   - Mining closed-patterns is more desired than mining max-patterns.

1.2.1 Efficient pattern mining methods

1. The downward closure property (a.k.a. Apriori) of frequent patterns: Any subset of a frequent itemset is frequent. If any subset of an itemset S is infrequent, then S is infrequent.

   - Level-wise, join-based approach (Apriori, 1994)
   - Vertical data format approach (Eclat, 1997)
   - Frequent pattern projection and growth (FP-growth, 2000)

2. The Apriori algorithm

   - Procedures
     - Scan database once to get frequent 1-itemset.
     - Repeat: generate candidate itemsets of length $k+1$ from the frequent itemsets of length k. Then test the candidate to generate frequent itemsets of length $k+1$.
     - Until no frequent or candidate set can be generated.

    - Return all the frequent itemsets of different lengths.

- How to generate candidates, i.e., self-joining and pruning. Self-joining generate candidates of length $k+1$ by combining two frequent itemsets of length k. Pruning deletes candidates of length $k+1$ if infrequent subset exists.
- Improvements of Apriori
    - Reduce the passes of transaction database scans
        - ∗ Partitioning: Any itemset that is potentially frequent must be frequent in at least one of the partitions of transactional database (TDB). It scans the database only twice.
            - · Scan 1: Partition the database and find local frequent patterns
            - · Scan 2: Consolidate global frequent patterns
        - ∗ Dynamic itemset counting
    - Shrink the number of candidates, e.g., direct hashing and pruning.
    - Exploring special data structures

3. Exploring the vertical data format ECLAT (Equivalence Class Transformation); It is a depth-first search algorithm using set intersection

4. FP-growth algorithm

    - Procedures
        - Find frequent single items and partition the DB based on each of them
        - Recursively grow frequent patterns by doing the above for each partitioned databse (a.k.a conditional database)
        - An efficient data structure FP-tree is constructed for efficient processing.
    - Recursively construct and mine conditional FP-trees until the resulting FP-tree is empty, or until it contains only one path, which will generate all the combinations of its sub-paths, each of which is a frequent pattern.

2 Week 2 on May 11 2015

1. Interestingness measures

    - Lift is more telling than support and confidence
        - $Lift(B,C) = \frac{confidence(B \to C)}{support(C)} = \frac{support(B \cup C)}{support(B) \times support(C)}$
        - Lift=1 indicates independence
        - Lift > 1 indicates positive correlation
        - Lift < 1 indicates negative correlation
    - χ^2 is used to test correlated events
        - $\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$
        - $\chi^2 = 0$ means independence
        - $\chi^2 > 0$ means correlated, either positively or negatively.

	$milk$	$\neg milk$	$\sum_{row}$
$coffee$	mc	$\neg mc$	c
$\neg coffee$	$m\neg c$	$\neg m\neg c$	$\neg c$
$\sum_{col}$	m	$\neg m$	$\sum$

- Lift and χ^2 may not be good for transactions with large null-transactions. In such a case, null-invariant measures are more proper for use, e.g., Jaccard, Cosine, Kulczynski, AllConf and MaxConf.
  - Null-invariant measures is crucial for massive data containing many null transactions
  - Imbalance-ratio and Kulczynski measures are promising to bring useful insights.

2. Other types of association rule mining
   - Multi-level association rules
   - Multi-dimensional association rules
   - Quantitative association rules
   - Mining negative associations
     - Rare patterns are those with low support but interesting
     - Negative patterns indicate negative correlation.
     - How to distinguish these two patterns
       * Support-based: If itemsets A and B are both frequent but rarely occur together, i.e., $sup(A \cup B) << sup(A) \times sup(B)$, then A and B are negatively correlated.Note that such measure's performance is poor for large null-transaction cases. Whether two itemsets A and B are negatively correlated should be be influenced by the number of null-transactions.
       * Kulczynski measure-based: If A and B are frequent but $(P(A|B) + P(B|A))/2 < \epsilon$, where ϵ is a negative pattern threshold, then A and B are negatively correlated. Such measure is a null-invariant measure.

3. Constraint-based pattern mining
   - Two types of constraints, i.e., pattern space pruning constraints and data space pruning constraints.
   - Pattern space pruning strategies
     - Pattern anti-monotonic, i.e., if an itemset S violates constraint C, so will its superset and hence, the mining on itemset S can be terminated. Note that Apriori pruning is anti-monotonic.
     - Pattern-monotone constraint: if an itemset S satisfies the constraint C, so does any of its superset. Hence, no need to check C in subsequent mining.
     - Data anti-monotone constraint: If a data entry t cannot satisfy a pattern p under C, then t cannot satisfy $p's$ superset either.
     - Succinct constraints are those can be enforced by directly manipulating the data.

    - Convertible constraints are those can be transformed into (anti-) monotone constraints.

3 Week 3 on May 11 2015

1. Given a set of sequences, find the complete set of frequent subsequences satisfying the minimum support threshold. A sequence contains several element, each may contain a set of items.
2. GSP-Apriori-based sequential pattern mining
    - Get all singleton sequences
    - Scan DB to find frequent sequence of length k
    - Generate candidate of length k+1
    - Repeat until no frequent sequence candidate
3. SPADE for vertical data format, by Zaki, similar to Eclat in association rule mining
4. PrefixSpan
    - No need to generate candidate subsequences
    - Projected DB keeps shrinking
5. Constraint-based sequential pattern mining
    - Anti-monotonic constraints
    - Monotonic constraints
    - Data anti-monotonic constraints
    - Succinct constraints: Enforce constraint by explicitly manipulating the data
    - Convertible constraints
    - Timing-based constraints
        - Order constraints
        - Min-gap/max-gap constraint
        - Max-span constraint
        - Window size constraint, i.e., how to merge items into element
6. Graph pattern mining
    - Apriori-based approach
        - A size-k subgraph is frequent if and only if all of its subgraphs are frequent
        - Candidate generation is based on either vertex or edge growing, and the latter is shown to be more efficient
    - gSpan approach, a depth-first growth of subgraphs from k-edge to (k+1)-edge, and then (k+2)-edge.
    - CloseGraph for mining closed graph patterns. It aims to handle the pattern explosion problem.

- gIndex is a graph indexing method.
- SpiderMine is an algorithm to mine top-k largest structural patterns in a massive networks

7. Pattern-based classification
    - The pattern extracted are then used as inputs for classification.
    - Associative classification
        - CBA, i.e., classification based on association rules; It uses high confidecne and high support rules to build classifiers.
            - Mine high confidence and high support association rules, while the right hand side is that class label.
            - Rank the rules in descending order of confidence and support.
            - Apply the first rule which match the test case to make prediction.
            - Can be more accurate than traditional classification methods, e.g., C4.5. One possible explanation is that by exploring high confidence and support associations among multiple attributes, it overcomes some constraints introduced by some classifiers which only considers one attribute at a time.
        - CMAR, uses multiple rules to make prediction.
            - CMAR is classification based on multiple association rules
            - The rules are pruned in two ways, first based on support and confidence (only use the more general rules), second based on χ^2 which aims to eliminate rules have non-positive correlation.
            - CMAR can improve model construction efficiency and classification accuracy.
    - Discriminative pattern-based classification, i.e., PatClass
        - Feature construction by frequent itemset mining
        - Feature selection using maximal marginal relevance (MMR)
        - Integrate with a general classification methods
        - k-itemsets are often more informative than single feature (1-itemsets) in classification
    - Direct mining of discriminative patterns (DDPMine) (2008)
        - Unlike the discriminative pattern-based classification, which first perorm frequent mining and then use certain criteria to select high quality discriminative patterns for classification, DDPMine directly mines high quality discriminative patterns.
        - Can be integrated with FP-growth algorithm
        - It can substantially increase the computational efficiency and high accuracy