A model to discover rare patterns using ML techniques

MOTIVATION

- **Research Problem**:
  For this project we intend to extract rare patterns from a database, we believe that though more work has been done in frequent pattern mining, discovering rare patterns is also very important and often times more useful as can be seen from its applications in different domains.

- **Importance**:
  Mining rare items is very important in diverse domains, below are some importance:
  - In the supermarket transactions, these rare items mainly represent the most profitable items in stock compared to frequently bought items like bread, chips etc.
  - In the medical field, a significant amount of data is collected from different devices such as electrocardiogram, positron emission tomography, diagnoses, or magnetic resonance imaging. The recognition of unusual patterns in such data often reflects disease conditions and provides useful information for experts
  - In Anomaly detection, mining rare itemsets can be helpful in detecting credit card fraud detection, call intrusion etc.

- **Challenges**
  Mining rare patterns is greatly affected by the varying frequencies between items in a transaction database, mining frequent and rare items from the database can become very tricky and we are faced with the following problems:
  - Important rare itemsets can't be exploited if min_sup is set too high
  - The itemset space become exploded when min_sup is set too low, because though it finds all the frequent and rare itemsets, it also finds useless itemsets that are frequent

  Example: In a supermarket transaction data, in order to find rules involving those infrequently purchased items such as food processor and cooking pan (they generate more profits per item), we need to set the minsup to very low (say, 0.5%). We may find the following useful rule:
  foodProcessor + cookingPun [sup = OS%, conf = 60%]
  However, this low minsup may also cause the following meaningless rule to be found:
  bread, cheese, milk + beer [sup = OS%, conf = 60%]
  Knowing that 0.5% of the customers buy the 4 items together is useless because a11 these items are frequently purchased in a supermarket. For this rule to be useful, the support needs to be much higher.

- **Existing Work**
  The existing work in rare pattern mining can be basically divided into 2 foundational approaches with other different methods stemming from these two foundational approaches: these 2 foundational approaches are the Apriori based approach (Breadth first search methods) and FP-Tree based approach (Depth first search methods).

    - ***Apriori based approach***: This approach explores the itemsets search space in a level wise manner, i.e, it finds 1-itemsets, then 2-itemsets, followed by 3-itemsets, till there are no more itemsets to be found, then it ends. The drawbacks of methods based on this approaches is the fact that They perform multiple scans over the dataset to count the support of candidate itemsets and also they employ a test-generate approach to generate candidate itemsets which leads to a rapid explosion of the itemset space.
    - ***FP-Tree based approach***: These approaches uses a top-down transversal method and thus have the advantage of not scanning the database multiple times and generating an explosive itemset space but most methods stemming from thjis approach has the disadvantage of not finding all rare itemsets and also not performing very well on small dense datasets.
- Papers
    - Modern Applications and Challenges for Rare Itemset Mining
    - Mining Frequent Patterns with Multiple Item Support Thresholds in Tourism Information Databases
    - Mining Association Rules with Multiple Minimum Supports