ASSOCIATION RULE MINING

We have studied the classification problem and have developed an ensemble-based classifier for disease diagnosis. This week we will focus on a complementary topic: *discovering relationships between features*. This kind of analysis can lead to significant insights in the application domain. For example, this is important in cancer research, because groups of genetic mutations may work collectively to cause the disease. Thus, the discovery of correlated genetic mutations provides knowledge that can be the starting point for research studies that investigate the mechanisms of the disease processes involved in cancer. While development of a diagnostic technique is important for correctly identifying a patient’s illness, the discovery of relationships between genetic mutations can lead to important insights into disease pathology, with potential implications for treatment and cure.

**Materials:**

Sections 12.1-12.3 of *Discovering Knowledge in Data*. D.R. Larose and C.D. Larose. Wiley. 2014, which is available at <https://alice.library.ohio.edu/record=b5187242?>.

**Concepts to learn from the materials:**

* Affinity analysis
* Association rules
* Support and confidence measures for association rules
* The *A Priori* algorithm for mining association rules

**Quiz**:

After learning the concepts listed above, complete the BlackBoard quiz no later than Monday November 8, 9:39 am. The quiz will cover your understanding of the *concepts to learn from the materials* (see above). The quiz may include multiple choice, true-false, fill-in-the-blank, and/or matching questions.

Data Mining Activity: (*to be started after you complete the quiz*)

Due date: no later than Friday November 12, 11:59 pm

1. Generate F1: *the set of frequent features*

Generate the set of features (genetic mutations) that occur in at least 4 samples

1. Generate F2: *the set of frequent feature pairs*
   1. Generate F2, the set of pairs of features that co-occur in at least 4 samples (consider ONLY the features in F1).
   2. Using F2, generate ALL possible association rules that contain a *single-item consequent*. Thus, for each pair (*A,B*) that is contained in F2, generate an association rule where *A* is the consequent and another rule where *B* is the consequent.
   3. Rank the association rules by support.
   4. Rank the association rules by confidence.
   5. Determine a minimum threshold value for selecting the most interesting association rules
      1. based on their support values
      2. based on their confidence values
   6. Select the association rules that meet your minimum support and confidence thresholds.
   7. Generate a table that shows
      1. the selected association rules, ranked by *support* \* *confidence*
      2. for each rule, report the values for support; confidence; and (support \* confidence)
2. Generate F3: *the set of frequent feature triplets*
   1. Use the *a priori* algorithm
      1. Construct C3, the set of candidate triplets:

Generate the set of candidate triplets by joining F2 with itself. That is, join feature pairs in F2 that have 1 item in common, to form the set of candidate feature triplets.

* + 1. Construct F3 by removing triplets from C3 using the *a priori* principle:

For each triplet in the candidate set, C3:

Generate all pairs of features that are subsets of the triplet

If one of the pairs generated in the previous step occurs in less than 4 samples, then remove the triplet from the candidate set

* 1. Using F3, generate ALL possible association rules that contain a single-item consequent. Thus, for each triplet (*A,B,C*) that is contained in F3, *A*, *B* and *C* would each be a consequent in an association rule.
  2. Rank the association rules by support.
  3. Rank the association rules by confidence.
  4. Determine a minimum threshold value for selecting the most interesting association rules
     1. based on their support values
     2. based on their confidence values
  5. Select the association rules that meet your minimum support and confidence thresholds.
  6. Generate a table that shows
     1. the selected association rules, ranked by *support* \* *confidence*
     2. for each rule, report the values for support; confidence; and (support \* confidence)

Submit an email to [welch@ohio.edu](mailto:welch@ohio.edu) that contains a brief report, including the following:

* The results from steps 1, 2, and 3.
* Your report should include the following sections:
  1. Research objectives (2-3 bullet points) – what are we trying to discover about cancer?
  2. Key results (2-3 bullet points) – what have we discovered about cancer?
  3. Summary of methods (2-4 bullet points) – describe your software design
  4. Key results (tables, figures, stats, lists, etc.)
  5. Discussion (2-4 bullet points) – how do your findings apply to cancer research?

Additionally, attach the computer program that you developed for this activity and the output of your program (either a screenshot(s) or a file).

**NOTE***: you must develop your own computer program to accomplish this assignment. You ARE NOT permitted to use pre-existing programs for performing association rule mining or any other component of this project.*

**NOTE**: I may respond to your email submissions with questions about your methods, results, and/or interpretation. Please respond promptly to my questions.