FEATURE SELECTION

We have developed a prototype ensemble-based classifier for disease diagnosis and we have performed association rule mining to discover relationships between feature pairs and triplets. In this final phase of the project we will focus on *feature selection*, the identification of the mutations that provide the most significant insights for cancer research. In part 1 of this activity, you will exploit the power of a fully developed random forest to identify a collection of relevant mutations (this is called feature selection) in the data set that you’ve been analyzing this semester. In part 2 of this activity, you will fully explore feature associations to discover groups of genetic mutations which may work collectively to cause cancer (this is called feature association selection); feature association analysis will likely contribute additional features to the feature selection process. These discoveries will provide a starting point for scientific research studies to investigate the mechanisms of the disease processes involved in cancer.

DATA MINING ACTIVITY: *Project Summary*

Part 1: *Deep classification analysis & Feature selection*

* completed

Part 2: *Deep association rule mining & Feature association selection*

* Due date: no later than Friday November 19, 11:59 pm

**PART 1**: *Deep classification analysis & Feature selection*

*Part 1 has been completed*

**PART 2**: *Deep association rule mining, Feature association selection, Feature selection update*

You have learned how to perform association rule mining in order to identify pairs and triples of genetic mutations that are strongly related. In this activity you will exploit the full power of association rule mining to help medical scientists to identify the full extent of relationships among sets of mutations that appear to work together in the cancer phenotype. This will involve the following major components:

* *Deep association rule mining*
* *Feature association selection*
* *Feature selection update*

*Deep association rule mining* will involve performing additional iterations of the *a priori* algorithm. Previously, you used the *a priori* algorithm to find the sets of frequent features (F1), feature pairs (F2), and feature triplets (F3), as well as important association rules. In this activity, you should perform additional iterations of the *a priori* algorithm to find F4, F5, … FN, as well as important association rules involving more than three features[[1]](#footnote-1). That is, find the set of frequent feature quadruplets (F4), the set of frequent feature quintuplets (F5), …, and the set of frequent feature N-tuplets (FN). Eventually, the *a priori* algorithm will produce an empty set of tuples, which will be the terminating point for your algorithm.

Due the combinatorial explosion of possible feature combinations, consider ways to reduce the search space, such as the following:

* be more restrictive in your frequent itemset threshold (e.g., only considering mutations that occur in at least 5 samples, or 6 samples, or 7 samples, etc.)
* increase the thresholds for identifying interesting association rules
* consider only cancer (**C**) samples, in order to identify feature groups that are positively correlated with the disease phenotype

*Feature association selection* will involve the interpretation of the results of deep association rule mining. You should select sets of related genetic mutations (from F4, F5, …, FN) that appear to function cooperatively in individuals with cancer and could be the subject of further scientific study. Justify your selections.

*Feature selection update*:

You should relate the selected feature associations to the features selected in PART 1. Are there any overlaps between selected features and selected feature associations? Are there any additional features that should be selected, based on the feature associations that you discovered in this activity?

Once you have selected a set of mutations, determine if any of the newly selected mutations are known to be associated with cancer (perform a Google search on the names of the mutated genes that you selected). Did you find anything interesting?

**NOTE***: you must develop your own computer program to accomplish deep association rule mining and feature association selection. You ARE NOT permitted to use pre-existing programs for association rule mining.*

Due date: no later than Friday November 19, 11:59 pm

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

* The results from deep association rule mining, feature association selection, and feature selection update.
* 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).

1. Note that in order to find quadruplets, you may need to relax the minimum threshold for *itemset frequency* (**φ**). In the previous assignment you set **φ** = 4, but if that value of **φ** results in an empty itemset F4, then you should try setting **φ** = 3. [↑](#footnote-ref-1)