Senior Project, 2014, Spring

FIU School of Computing & Information Sciences

Venture Hive Regional Miner

Student: Gian Piero Spicci, Florida International University
Mentor: Luis Amat, CTO
Instructor: Masoud Sadjadi, Florida International University

Problem

Venture Hive is a South Florida company incubator –it provides other businesses, usually start-ups, with the guidance and tools they need to grow and be successful.

To incubate efficiently, Venture Hive collects large amounts of data about regional companies; however, it needs a system that mines and sifts the data to retrieve useful information.

Frequent mining leads to the discovery of associations among items in large relational data sets. Venture Hive could benefit from finding these association patterns in its databases, especially in order to assess the replicability of practices that are correlated with desirable results.

Current System

Currently, Venture Hive does not have a data mining system in place; the company's incorporation of the Regional Miner into its operational structure will mark its incursion into new territory. The idea is, on the one hand, for Venture Hive to identify salient correlations between specific business attributes in order to make informed recommendations to its clients; on the other hand, to sort businesses and entrepreneurs into groups, or "clusters", based on shared features, with the purpose of tailoring its incubation approach according to the cluster to which a client belongs.

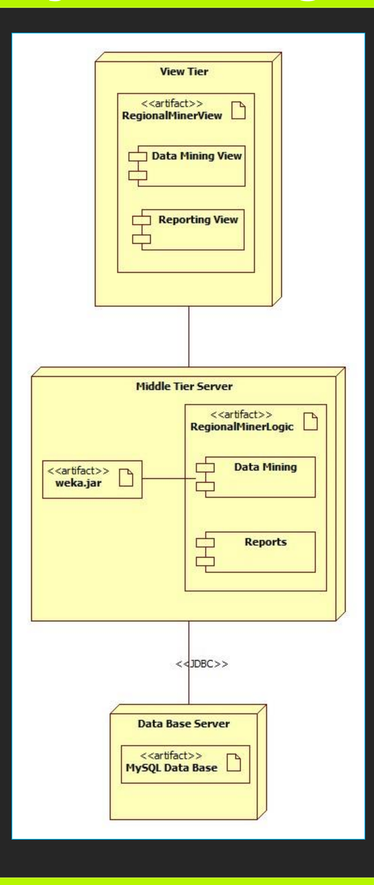
Having a human sort vast amounts of data in search of these associations and clusters can be very costly, even unfeasible; having a program perform this task is much more manageable and effective.

Requirements

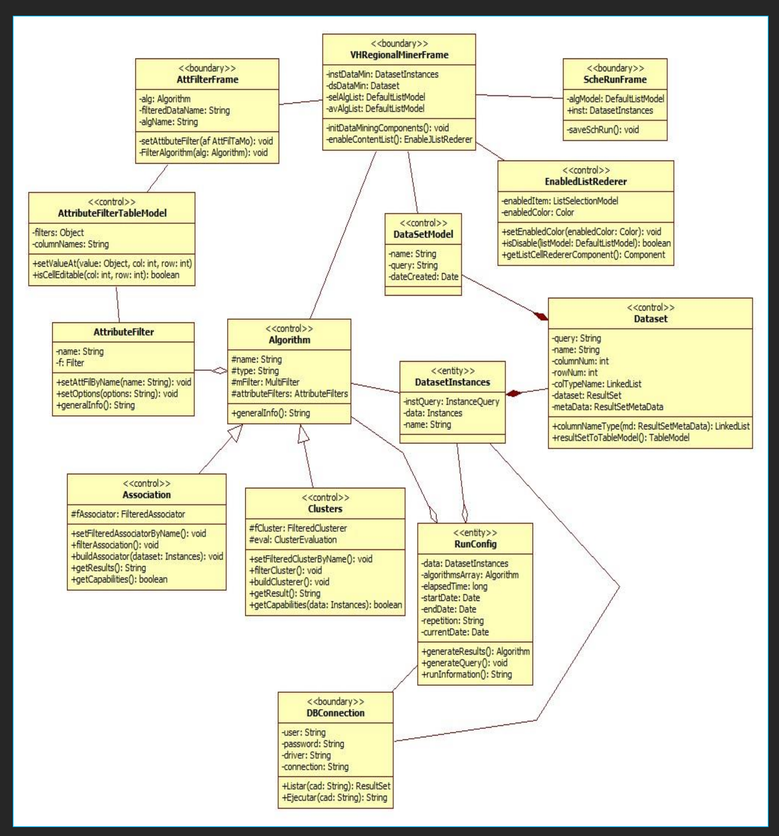
Venture Hive Regional Miner shall:

- Allow the user to select filtered data and view its details.
- Display which algorithms require data adjustments in order to be executed.
- Allow the user to perform data cleaning routines (e.g. replace missing values) to the selected data set for a specific algorithm.
- Allow the user to perform data transformation to provide better results (e.g. discretize and normalize) given a data set and specific algorithm.
- Allow the user to run multiple algorithms, regardless of type (i.e. association and clustering), in a single instruction.
- Allow the user to save a run configuration (i.e. data set, data adjustments and algorithms) for future execution.

System Design



Object Design



Implementation

Weka 3.7 developer version, a suite of machine learning software written in Java, is employed by our system –also written in Java for compatibility– to offer the user a range of algorithm choices from among the following two categories: association and clustering.

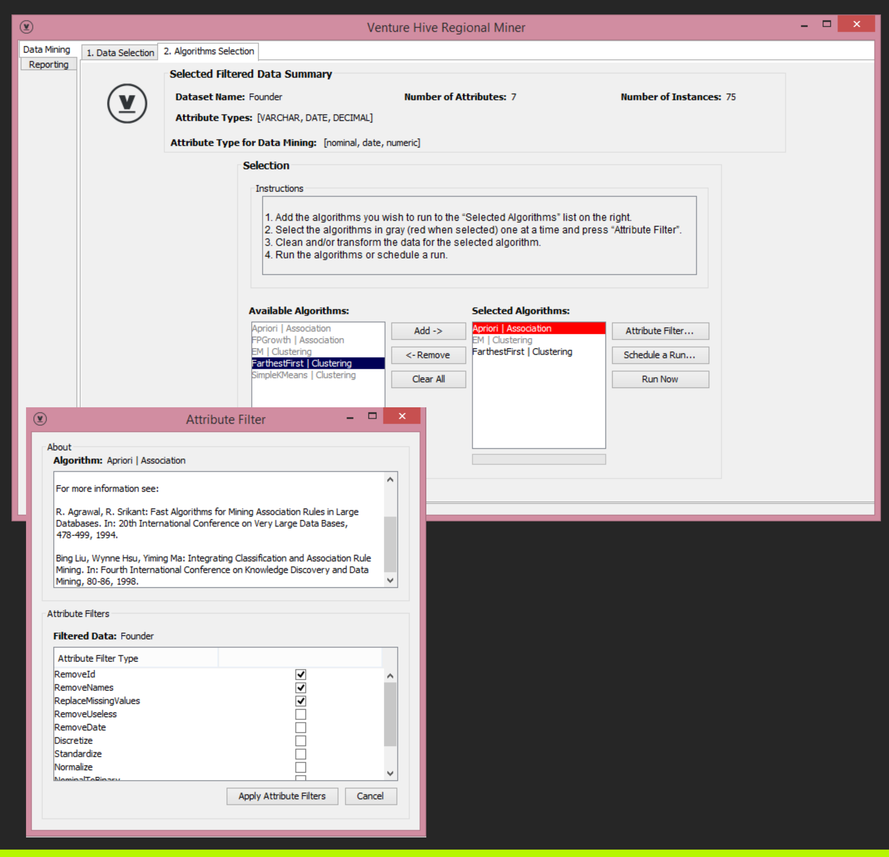
The algorithms are run on data sets to generate results that are intelligible to the user; multiple algorithms can be run at once, with each algorithm producing an individual result.

Now, because some data sets need to be refined before they are ready to be subjected to algorithm runs, the system uses tools provided by Weka for data pre-processing. Whenever pre-processing is necessary, the user will be guided by system recommendations. All the user-initiated tasks mentioned above can be scheduled by the user to for future execution.

Verification

Test ID	VHRM_system_runAlgorithm_002
Test Purpose	This test verifies that the system can correctly handle "Out Of Memory" errors.
Test Setup	• The user selected a data set containing over 1,000,000 instances. • All the algorithms are selected with the assigned attribute filters.
Inputs	• User clicks "Run Now". • Apriori started to run with the "Nominal to binary" and "Numeric to binary" attributes filters and was fed the "Municipality" data set…
Expected Output	"You run out of memory! An algorithm might be taking too much java heap space". No results are stored in the database.

Screenshots



Summary

The Regional Miner can help Venture Hive incubate its clients in two main ways.

First, it picks out correlations or associations between specific attributes; this allows Venture Hive to make recommendations to its clients based on what sorts of business practices are correlated with good results (e.g., number of founders being positively correlated with annual revenue).

Second, it assembles groups of businesses or entrepreneurs based on key features that they have in common, no matter how apparent the features may be to a human observer. Different incubation methods will suit different companies, yet companies in the same cluster will likely benefit from the same method. Venture Hive can "profile" a client according to the cluster to which it belongs in order to reduce the range of appropriate incubation methods.

Acknowledgement

The material presented in this poster is based upon the work supported by Luis Amat. I am thankful for the help and the support that I received from my family, Adriana Mujica, Ricardo Abend and Elias Eskenazi.