**Reviewed proposal**

Due to time and resources constraint, the initial proposal has been changed. Instead of dealing with the large data set, we played with the small data set and used different methods rather than the ensemble selection. We applied three classifiers (Naïve Bayes, Lazy IBK, and Decision Table) on the small data set after preprocessing the data.

Preprocessing Small Data Set:

To have a quality decision, it is necessary to have a quality data and no one can deny that real world data is incomplete and noisy. Therefore, it is important to preprocess data before classify it to tackle the problem of incomplete data. Incomplete in missing attributes values, lacking attributes of interest and inconsistent containing different names or codes (categorical data).

Classification:

Naïve Bayes Classifier (NBC) is simple in structure and it is based on the assumption that predictors are conditionally independent given the target variable. Because of its simplicity, NBC was attractive choice since we have a large set of variables. The second classifier is Lazy IBK (Instance-Based K) which is one of the Nearest Neighbor algorithms. Thirdly, we used Decision Table Classifier which uses a simple decision table majority algorithm; it makes decisions on attributes for each instance.

**Framework implementation**

Preprocessing:

To make use of the data, we cleaned it up as the following:

1. Remove all attributes that has > 50% missing values
2. Use unsupervised filters to:
   1. Replace the missing values for nominal and numeric attributes with the modes and means
   2. Change numeric data type to nominal
   3. Apply discretization
3. Select features using InfoGainAttributeEval filter

Which evaluates the worth of an attribute by measuring the information gain with respect to the class, ranks attributes by their individual evaluations. Based on the ranking, we remove the least significant attributes that has < 0.001

Classification:

We applied the three mentioned classifiers for the training data set with the three labels (churn, appetency, and up selling). Table-1 and Table-2 show the results of each classifier twice, one by using the entire preprocessed data, and the other one using only the numeric data (categorical data was removed).

Whole data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Churn | | Appetency | | Up selling | |
| F-Measure | AUC | F-Measure | AUC | F-Measure | AUC |
| NBC | 0.857 | 0.568 | 0.963 | 0.523 | 0.834 | 0.682 |
| Lazy IBK | 0.878 | 0.533 | 0.97 | 0.5 | 0.882 | 0.564 |
| Decision Table | 0.893 | 0.53 | 0.982 | 0.5 | 0.94 | 0.756 |

No categorical

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Churn | | Appetency | | Up selling | |
| F-Measure | AUC | F-Measure | AUC | F-Measure | AUC |
| NBC | 0.82 | 0.529 | 0.951 | 0.497 | 0.794 | 0.658 |
| Lazy IBK | 0.88 | 0.532 | 0.97 | 0.525 | 0.892 | 0.619 |
| Decision Table | 0.893 | 0.53 | 0.973 | 0.5 | 0.94 | 0.757 |

Analyze the results …………. And compare them with the KDD results generally

~ would you mind completing this part please?

Let me know please , so if you mind I’ll analyze it

Thanks ☺