Data Mining and Exploration - Interim Report

Thorvaldur Helgason (s1237131)
Daniel Stanoescu (s0838600)
Maria Alexandra Alecu (s1255964)

February 28, 2013

1 What we have done so far:

1.1 Data Preprocessing

The data we have chosen for this project is the Orange telecom customer marketing database. We have been provided with a labelled training set and unlabeled test set for the data focusing on predicting the following three target values: customer churn (switch providers), appetancy (tendancy to buy new products) or purchase upgrades. Given the proposed dataset, we are aiming at solving a classification problem. To respect customer privacy, both the order of the customers as well as the variables in the dataset have been shuffled.

There are two datasets for this task, one small with 230 features and another big with 15000 features. We have decided to work solely on the small dataset to begin with because of the lack of proper computational resources, but we might try working with the big dataset if there will be time. Also, we are only going to be using the training data and labels because we do not have labels for the test data and can therefore not use it for evaluation purposes. Instead we are splitting the labeled training set up so 80% of it will be used for training our classifier and 20% of the date to evaluate it. We also make sure that the split is balanced in terms of classes.

After inspecting the data, we have identified that it contains both numerical and categorical values. We used R to explore as well as preprocess the data. We began with preprocessing the numerical data. We replaced all the entries that were empty with zero. Secondly we replaced the zero entries with the mean of the values of each variable in the dataset. In the case of the categorical data, we changed the empty strings to Not Available. We then converted all the strings to numeric values so we could handle data operations easier. Additionally, we created a binary bag-of-features dataset to see if it would work better with the **Naive Bayes** classifier.

1.2 Classifiers

We began experimenting with a series of algorithms in order to figure out which ones would perform better. We looked at employing **SVMs** but we quickly abandoned the idea as they would require a higher computational requirement than other methods. We have also familiarized ourselves with **Naive Bayes** and **Decision Trees**.

We are looking at **Naive Bayes** because that was used as a basis in the competition and we want to create our own baseline for our modified dataset. So far, it looks like **Naive Bayes** is performing better on the raw and binary datasets than the ones were missing values are replaced, although the overall performance is pretty poor (AUC ~ 0.5).

We have implemented **Decision Trees** using the package "rpart" from R and tested it on the preprocessed data, using the first 50 features of the data set.

In order to evaluate the performance of the algorithm, we have performed a 2-cross validation on the set. We have obtained a very good accuracy, namely of 91%. However, the function "printcp" function available in the "rpart" package seems to contradict this result. The function gives the error from 10-fold cross validation performed on the data set, using different values for the number of nodes in the decision tree. With this classifier, we obtain an error greater than 100%, meaning that no matter how many nodes we include in our decision tree, there is no decision tree that fits the data (our current interpretation).

2 What we plan on doing:

We plan to experiment with a linear classification method, more exactly, employing logistic regression. If there is time, a look into a mixture of gaussians to model the date will be attempted.

With respect to **decision trees**, we firstly want to see the performance of the classifier on the unprocessed data, given that decision trees can handle missing data and categorical data. Our next goals are to use all the features of the data set and perform cross-validation using more folds.

In what concerns the methods we want to use, our main goal is to use the classifiers we obtained at the previous step in **ensemble** methods, such as **random forests**, **bagging** and **boosting**. We plan to use the built-in function in R, such as "cforest" and also experiment with the same functions in Weka.

3 What comparisons we want to run:

To evaluate the performance of our algorithms on the data set we are aiming at running a 5-fold cross-validation on the training set and then run a final evaluation on the test set. For each classifier we are going to store their accuracy, confusion matrix, ROC curve and AUC.