Homework 3 - Classification rules

UTFPR - CPGEI - Data Mining
Prof. Dr. Heitor Silvério Lopes

Vinícius Couto Tasso

October, 2019

Communities and Crime dataset

The Communities and Crime dataset[1] combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR of various American communities. The objective is to try and establish relations between these communities and the occurrence of violent crimes (murder, rape, robbery, and assault) within the United States.

The first five attributes were removed from the dataset because they are irrelevant for this task. The rest of the attributes, all of which were already normalized, were discretized. To lessen the disbalance of classes, SMOTE[2] was applied to generate data. Table 1 compares the original amount of instances per class on the original dataset versus the augmented dataset. In order to better represent the target attribute (*ViolentCrimesPerPop*), it was relabeled and classified into five classes, from left to right in order of interest: **Light, Low, Moderate, Intense and Extreme**. All numeric attributes were also discretized into 5 range of values as well.

	Light	Low	Moderate	Intense	Extreme	Total
Original	1188	432	192	95	87	1994
Augmented	1188	432	384	380	348	2732

Table 1: Comparison of the amounts of instances per class before and after data augmentation with SMOTE.

Although the dataset is still unbalanced, the classes of greater interest (**Moderate** and **Extreme**) are closer to being balanced in comparison with the remaining classes. In order to not have a dataset made mostly from synthetic data, SMOTE was not used any further.

Table 2 shows some of the experiments performed with the dataset. Although PART was a better classifier than the equivalent JRip classifier in both scenarios, it was responsible for generating a much more complex and less comprehensive set of rules. For this reason, JRip was the classifier chosen to draw conclusions from. Establishing a higher minimum of objects per rule allowed the algorithm to generate about 4 rules per class on average, without having much impact on the classification performance.

	Min. obj.*	# rules	Avg. TP Rate	Avg. FP Rate	F1 Score
JRip	2	50	0.705	0.144	0.681
JRip	15	21	0.673	0.187	0.634
PART	2	175	0.713	0.113	0.697
PART	15	43	0.690	0.114	0.670

Table 2: Results obtained with different experiments using JRip and PART as classifiers. Min. obj. was the only modified parameter and it is responsible for defining the least amount of objects each rule must include.

Another reason to use JRip for this problem is the very nature of the algorithm: it starts generating rules from the minority classes which are, in this particular case, the ones of higher interest.

Since the dataset has a pretty large number of attributes per instance, it is reasonable to assume that not every attribute is directly correlated with the crime rate of a community. From the classification rules obtained it is possible to make a few assumptions about possible reasons correlated to high crime rates:

[1] Available at http://archive.ics.uci.edu/ml/datasets/communities+and+crime
[2] Synthetic Minority Oversampling Technique

- Communities where the population density per square mile is either very low or very high;
- High percentage of kids born to never marry;
- Percentage of people under the poverty level above average;
- Poor drug control activities by the police (crimes could be often drug-related);
- Communities where the population is largely composed of younger people (12-21 years);
- Most of the people above 25 years old are not high school graduates;

Contraceptive Method Choice dataset

The Contraceptive Method Choice dataset is a subset of the 1987 Nation Indonesia Contraceptive Prevalence Survey. The data was gathered from women who were either not pregnant or didn't know they were at the time of the interview. The objective is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics.

OneR was used to establish the classification baseline. As Table 3 shows, not much improved was made by JRip's classification when compared to OneR. This may be an indicator that the attribute used by OneR to make predictions is much more important than any other attribute, that the rest of the attributes don't hold relevant information for the task or a consequence of the dataset imbalance.

	# rules	TP Rate	FP Rate	Accuracy
OneR	1	0.475	0.354	0.474
JRip	9	0.484	0.350	0.484

Table 3: Comparison between JRip and OneR(baseline) results for classification

Although there was only a slight improvement over the baseline, as any rule generator, there are rules with better coverage and accuracy. The OneR generates a rule with 100% coverage, but its accuracy was awfully low. JRip, on the other hand, presented 9 rules, some of which had great accuracy, but not very large coverage.

From the results, it is possible to draw conclusions about the usage of short and long-term contraceptive methods, although probably not very assertive given the classification results. Results show that women with higher education and mother of at least 3 kids are likely to use long-term contraceptive methods. Older women when the number of children born is above average and younger women seem to be more attracted by short-term methods. The standard of living appears to be another factor that influences the method of choice since women with higher standards usually choose short-term contraceptives.