Классификация текстов в Weka

Авторы: Лиана Бакрадзе, Анастасия Моисеева, Александра Михайлова

1. Классификация на данных своей команды

В прошлом задании нужно было заготовить по 1 статьи на человека из каждой из десяти категорий. Нас в команде трое, поэтому получаем 30 статей, по 3 статьи на класс.
При таком маленьком объёме обучающей выборки мы не ожидали выдающихся результатов от классификатора.

Подготовка данных

При выполнении предыдущего задания мы уже разбили тексты по категориям, поэтому оставалось только поместить каждый документ в отдельный текстовый файл и завернуть в папки с соответствующим названием.

На выходе получили arff-файл с такой структурой:

@attribute text string
@attribute @@class@@ {science,politics,sport,internet,hi-tech,culture,economics,incident,society,auto}

Далее данные были обработаны, как описано в презентации: преобразованы к виду вектора из слов. При этом были такие параметры: IDFTransorm=True, TFTransform=True, lowerCaseTokens=True, minTermFreq=3 и добавлен список русских stop-слов (везде в этом отчёте параметры указываются, если они отличны от дефолтных).

Был произведён выбор атрибутов с критерием хи-квадрат и Ranker.
Для оценки выборка была поделена на обучение и контроль в соотношении: 2:1

J48

При дереве с параметром minNumObj=1 (у нас всего по одному объекту на каждый класс) 4 из 10 текстов классифицируются правильно:

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0	1	1	1	science
1	0	1	1	1	politics
0	0	0	0	0	sport
0	0.1	0	0	0	internet
0	0	0	0	0	hi-tech
1	0.222	0.333	1	0.5	culture
0	0.2	0	0	0	economics
0	0	0	0	0	incident
0.5	0.125	0.5	0.5	0.5	society
0	0	0	0	0	auto
0.4	0.047	0.333	0.4	0.35	Weighted Avg

Построенное дерево выглядит вполне адекватно для данной выборки: например, если есть слово “полиция”, то классификатор отправляет статью в класс “происшествия”.

SMO

Использовалась SVM с полиномиальным ядром степени 2.
Правильно были классифицированы всего 2 из 10 объектов.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0	0	0	0	0	science
1	0	1	1	1	politics
0	0	0	0	0	sport
0	0	0	0	0	internet
0	0	0	0	0	hi-tech
1	0.778	0.125	1	0.222	culture
0	0.1	0	0	0	economics
0	0	0	0	0	incident
0	0	0	0	0	society
0	0	0	0	0	auto
0.2	0.078	0.113	0.2	0.122	Weighted Avg.

Logistic

Логистическая регрессия с дефолтными параметрами правильно классифицирует 4 из 10 объектов.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0	1	1	1	science
1	0	1	1	1	politics
0	0	0	0	0	sport
0	0	0	0	0	internet
0	0	0	0	0	hi-tech
1	0.444	0.2	1	0.333	culture
0	0.2	0	0	0	economics
1	0	1	1	1	incident
0	0	0	0	0	society
0	0	0	0	0	auto
0.4	0.044	0.32	0.4	0.333	Weighted Avg.

2. Классификация на данных всех команд.

На момент выполнения этого задания ни у одной из команд не было целой подборки размеченных по разделам статей. Два комплекта (всего 20 статей) в репозитории всё-таки нашлось. Наша команда дополнила получившуюся подборку так, чтобы получалось по 10 статей в каждом классе.
Следует заметить, что выборка всё равно слишком маленькая для того, чтобы нормально обучить классификатор, но результаты получились уже более радостные.

Данные были подготовлены аналогичным образом. Только частота minFreq была увеличена до 4 ввиду бо́льшего размера выборки.

Logistic
Логистическая регрессия с дефолтными параметрами правильно классифицирует 22 из 35 объектов.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.667	0	1	0.667	0.8	science
0.5	0	1	0.5	0.667	politics
1	0	1	1	1	sport
0.667	0.031	0.667	0.667	0.667	internet
0	0.061	0	0	0	hi-tech
1	0.129	0.5	1	0.667	culture
0.25	0.032	0.5	0.25	0.333	economics
0.75	0.097	0.5	0.75	0.6	incident
1	0.059	0.333	1	0.5	society
0.25	0	1	0.25	0.4	auto
0.629	0.037	0.724	0.629	0.616	Weighted Avg.

J48

При дереве с параметром minNumObj = 1 14 из 35 текстов классифицируются правильно:

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0	0	0	0	0	science
0.25	0.065	0.333	0.25	0.286	politics
1	0	1	1	1	sport
0.333	0	1	0.333	0.5	internet
0	0.121	0	0	0	hi-tech
0.75	0.226	0.3	0.75	0.429	culture
0	0.129	0	0	0	economics
0.25	0.032	0.5	0.25	0.333	incident
1	0.088	0.25	1	0.4	society
0.25	0	1	0.25	0.4	auto
0.4	0.061	0.508	0.4	0.391	Weighted Avg.

Для некоторых классов условия в дереве выглядят очень правдопободно: например, встречая слова google и яндекс, классификатор отправляет текст в класс “internet”

SMO

Использовалась SVM с полиномиальным ядром степени 1.
Правильно были классифицированы всего 15 из 35 объектов.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.333	0.063	0.333	0.333	0.333	science
0.5	0	1	0.5	0.667	politics
0.5	0.034	0.75	0.5	0.6	sport
0.333	0	1	0.333	0.5	internet
0	0.061	0	0	0	hi-tech
0.75	0.387	0.2	0.75	0.316	culture
0	0	0	0	0	economics
0.25	0	1	0.25	0.4	incident
1	0.088	0.25	1	0.4	society
0.75	0	1	0.75	0.857	auto
0.429	0.061	0.616	0.429	0.442	Weighted Avg.

Анализ

Ошибки классификатора вызваны, прежде всего, неточностью предметной области. Некоторые классы разделяются лучше, другие хуже. Например, сложно различить hi-tech и интернет.

Маленький объём выборки тоже негативно повлиял на качество классификации. Его можно было существенно улучшить, если бы были хорошие методы удаления “шума”: союзов, слов типа “сообщает” и тд.

Качество классификации можно было бы существенно улучшить, подобрав более оптимальные параметры, но тут, опять же, выборка для этого слишком маленькая.