Алгоритм ленивой классификации

Жиенбай Олжас

М15НоД ИССА

1. Введение

В данной работе реализован алгоритм ленивой классификации объектов, представленные бинарными признаками, а также проверка метрик качеств с “Random Forest Classifier”

1. Исследование зависимости точности классификации от значения порога (превышения отношения количества примеров одного класса к другому) и нахождение оптимального порога для достижения лучшей точности классификации;

Исследование зависимости точности классификации от объема тренировочных и тестовых данных (количества файлов).

1. Тренировочные данные разделяются по классам, первый класс с положительным результатом и отрицательным результатом(т.е определяем положительный и отрицательный контекст)  
   Для классификации объектов использовались формулы и . Значение порога определяем как , . Если оно выполнено то значение классифицируется положительно, если , то значение классифицируется отрицательно, иначе неопределенно.   
   Сравнивая каждое из значений в тестовой выборке со значениями из обучающей, подсчитывается количество TP(или true positive rate), FP(или false positive rate), TN(или true negative rate), FN(или false negative rate), и Contradictory (порог не пройден).
2. TP – правильные положительные, если мы предсказали положительный класс и реальное значение класса положительное (причем отношение положительных к отрицательным превышает порог)

FP – неправильные положительные, если мы предсказали положительный класс, а реальное значение класса отрицательное (причем отношение положительных к отрицательным превышает порог)

TN – правильные отрицательные, если мы предсказали отрицательный класс и реальное значение класса отрицательное (причем отношение отрицательных к положительным превышает порог)

FN – неправильные отрицательные, если мы предсказали отрицательный класс, а реальное значение класса положительное (причем отношение отрицательных к положительным превышает порог)

1. Алгоритм находит лучшее значение порога и соответствующую ему точность.  
   Экспериментальный путем рассматривается отрезок [0.6, 2] лучший порог оказался 1.0 с точностью accuracy = 0.921, однако random forest classifier показывает точность accuracy = 0.991