Лабораторная работа №2.

КЛАССИФИКАЦИЯ. ПОСТРОЕНИЕ МОДЕЛИ КЛАССИФИКАЦИИ. СРАВНЕНИЕ РАЗЛИЧНЫХ АЛГОРИТМОВ КЛАССИФИКАЦИИ И ВЫБОР ОПТИМАЛЬНОГО В ORANGE

Цель и задача работы Изучить основные методы классификации с использованием приложения «Orange Data Mining». Осуществить классификацию тестовых данных, используя разные алгоритмы. Научиться сравнивать результаты работы алгоритмов классификации и выбирать наиболее подходящий.

Пример построения модели data mining в Orange

Кластеризация с помощью дерева решений

Для загрузки файла используе м виджет **File**, который находится на вкладке **Data**. В качестве примера возьмём файл *zoo.tab* из набора тестовых данных, поставляемых вместе с Orange.

Далее выберем классификатор «**Classification Tree**» из вкладки **Classify**. Для анализа качества классификации используем виджеты **Test & Score** и **Confusion Matrix.**

Рассмотрим окно настроек виджета **Test & Score**. В нём можно указать режим работы виджета. Виджет поддерживает различные методы отбора проб (разбиения входных данных на обучающую и тестовую выборки).

1. **Cross validation** разбивает данные на заданное пользователем количество блоков (обычно 5 или 10). Алгоритм тестируется на примерах из каждого блока, при этом блоки, используемые для обучения и предсказания, постоянно меняются (сначала прогнозируется первый блок, потом второй и так далее, а остальные блоки используются для обучения).

2. **Leave-one-out** похож на **Cross Validation**, но он использует в качестве блока только один элемент (т.е. количество блоков будет равно размеру выборки). Этот метод, очевидно, очень стабильный, надежный и очень медленный.

3. **Random sampling** (случайная выборка) случайным образом разбивает данные на обучающую и тестируемую выборки в указанной пропорции (например, 70:30); вся процедура повторяется в течение определенного количества времени.

4. **Test on train data** (тест на тренировочных данных) использует весь набор данных для обучения, а затем для тестирования. Этот метод практически всегда дает неправильные результаты.

5. **Test on test data** (тест на тестовых данных): вышеуказанные методы используют данные только от одного источника данных. Чтобы ввести другой набор данных с примерами тестирования (например, из другого файла или некоторых данных, выбранных в другой виджет), мы выбираем отдельный сигнал проверки данных в канале связи и «Тестирование на тестовых данных».

Порядок выполнения работы

1) Загрузить данные с помощью виджета File.
2) Осуществить классификацию данных с помощью алгоритмов Classification Tree, Logistic Regression, Naive Bayes, SVM, CN2 Rule Induction, Nearest Neighbors, Random Forest Classification
3) Осуществить кросс-валидацию с помощью виджета Test & Score, используя различные виды разбиения входных данных на тестовые и проверочные (cross validation, random sampling, leave one out).
4) Проверить различные варианты выбора тестовой и обучающей выборки в виджете Test & Score.
5) Обосновать выбор наилучшего алгоритма для классификации исходных данных. Использовать виджеты Confusion Matrix и ROC Analysis.
6) Для алгоритма Classification Tree вывести дерево решений в графическом виде.
7) Вывести ошибки классификации для разных алгоритмов на точечную диаграмму вместе с результатами правильной классификации.
8) Вывести ошибки классификации для разных алгоритмов в виде таблицы.
9) Вывести и проанализировать ROC-кривые для разных алгоритмов.
10) Осуществить классификацию данных файла, используя виджет Predictions. Вывести полученные результаты в виде таблицы.
11) Обосновать выбор оптимального алгоритма классификации. Подготовить отчёт.