Интеллектуальный анализ больших данных

НГТУ НЭТИ

Новосибирский государственный технический университет НЭТИ

Кафедра Автоматизированных систем управления НГТУ

Доцент каф. АСУ

Муртазина М.Ш.

к.т.н.

Вторая лабораторная работа (методические указания)

Классификация – это метод, который заключается в построении моделей, выполняющих отнесение объекта к одному из нескольких на априорно заданных классов.

Классификации по признаковому описанию объекта является наиболее часто встречающейся задачей машинного обучения.

Классификация относится к стратегии обучения с учителем.

Вторая лабораторная работа (методические указания)

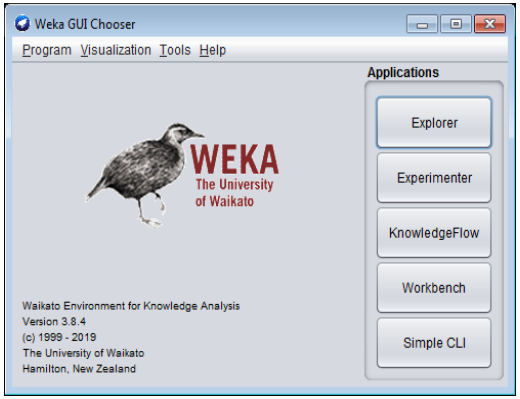
Существуют сотни методов классификации.

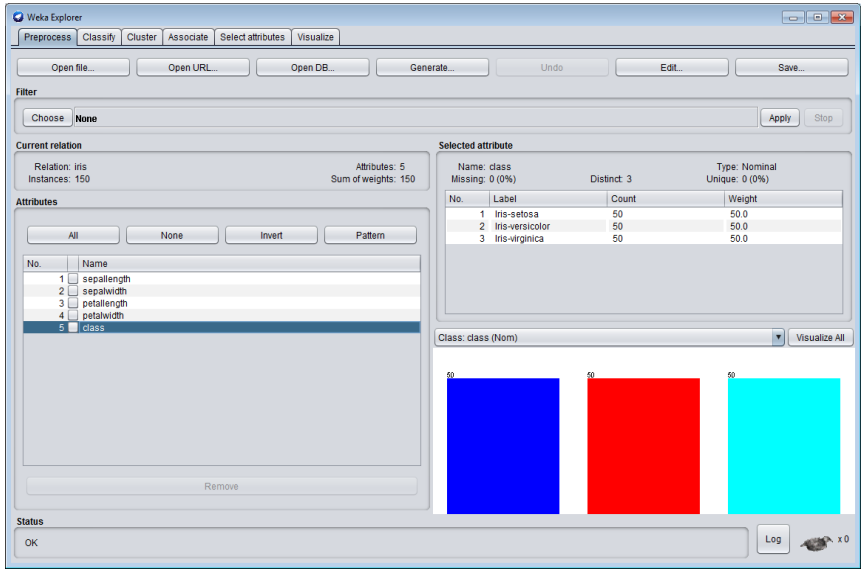
К наиболее часто используемым методам классификации относятся:

- Метод k-ближайших соседей (K-Nearest Neighbors),
- Классификатор дерева решений (Decision Tree Classifier),
- Наивный байесовский классификатор (Naive Bayes),
- Метод опорных векторов (Support Vector Machines),
- Логистическая регрессия (Logistic Regression),
- Многослойный перцептрон (Multilayer perceptron).

Вторая лабораторная работа (методические указания)

Целью выполнения лабораторной работы по дисциплине является закрепление у студентов теоретических знаний и выработка навыка применения этих знаний при решении практических задач классификации по признаковому описанию объекта в программной среде WEKA.





Вторая лабораторная работа (методические указания)

Задачи лабораторной работы:

- Описать набор данных.
- Описать назначение стандартных классификаторов пакета и средства отбора признаков WEKA.
- Исходя из набора данных, выбрать из стандартных классификаторов WEKA те, которые можно применить к вашему набору данных. Исследовать отбор признаков.
- На всех выбранных классификаторах провести обучение с настройками по умолчанию. Проанализировать результаты.
- Определить наиболее подходящие алгоритмы классификации для набора данных (3 шт.) и для них провести эксперимент с разными настройками.

Вторая лабораторная работа (методические указания)

Для решения задачи классификации применяется набор стандартных классификаторов следующих типов:

- bayes (классификатор, основанный на применении теоремы Байеса),
- functions (группа классификаторов, основанных на математических моделей),
- lazy (метод обучения на основе обобщения обучающих данных, которые задерживаются до тех пор, пока не будет сделан запрос к системе),
- meta (подход, который позволяет определить наиболее подходящий алгоритм и параметры к нем для конкретной задачи из портфолио алгоритмов),
- misc (разнообразные классификаторы),
- rules (классификационные правила),
- trees (обучение с помощью деревьев решений).

Вторая лабораторная работа (методические указания)

1. Составить таблицу вида для классификаторов.

	Классификатор	Описание	Опции
	bayes		
1	BayesNet	Обучение байесовской сети с использованием различных алгоритмов поиска и показателей качества. Базовый класс для классификатора байесовской сети, предоставляет структуры данных (структуру сети, условные распределения вероятностей и т.д.) и средства общие для алгоримов обучения сети Байеса, таких как K2 и B.	**numDecimalPlaces** - количество десятичных разрядов, которое будет использоваться для вывода чисел в модели. **batchSize** - предпочтительное количество экземпляров для обработки, если выполняется пакетное прогнозирование. **estimator** - выбор алгоритма оценки для поиска таблиц условной вероятности байесовской сети. **debug** - вывод дополнительной информации в консоль. **searchAlgorithm** - выбор метода, используемого для поиска сетевых структур. **useADTree** - когда используется ADTree (структура данных для увеличения скорости подсчета), время обучения обычно уменьшается.

2. Аналогично пункту 1 описать средства отбора признаков.
3. Описать набор данных

Вторая лабораторная работа (методические указания)

4. Результаты обучения представить в виде таблице вида

Название	Корректно классифицированные экземпляры	Некорректно классифицированные экземпляры	Коэффициент Каппа	Средняя абсолютная ошибка	Среднеквадратичная ошибка	Относительная абсолютная ошибка	Относительная среднеквадратичная ошибка
Bayes							
BayesNet	79,3	20,7	0,7577	0,0711	0,2139	29,08	61,16
NaiveBayse	67,41	32,59	0,6194	0,1088	0,2538	44,47	72,56
NaiveBayseMultinomialText	16,63	83,37	0	0,2446	0,3497	1000	100
NaiveBayseUpdateable	67,41	32,59	0,6194	0,1088	0,2538	44,47	72,56
Functions							
Logistic	95,2155	4,7845	0,9441	0,0149	0,1135	6,0876	32,4513
MultilayerPerceptron	94,3629	5,6371	0,9342	0,0207	0,118	8,4736	33,7464

Вторая лабораторная работа (методические указания)

5. Выбрать три алгоритма с наилучшими результатами и провести эксперимент с разными настройками.
6. Проверить можно ли сократить количество признаков для вашего набора данных. Если да, сравнить как изменился результат работы моделей.
7. Сделать выводы.