УДК 378.147

В.В.Братищенко
Байкальский государственный университет
г.Иркутск, Российская Федерация

Использование в учебном процессе визуальных методов конструирования информационных технологий анализа данных

Аннотация. В статье рассматривается программное обеспечение дисциплин, связанных с технологиями анализа данных (Online analytical processing, Data Mining). Предлагается использовать программное обеспечение (MS Power BI и Orange3), предлагающее визуальные методы определения процессов аналитической обработки данных. Применение таких программ позволяет снизить затраты на программирование компьютерных экспериментов и увеличить наглядность процесса обучения.

Ключевые слова. Информационные технологии анализа данных, Online analytical processing, Data Mining, MS Power BI, Orange4, визуальное проектирование.

Визуальные методы обучения [1,2] находят широкое применение в учебном процессе благодаря возможности демонстрации любых эффектов на экране компьютера. Кроме этого, визуальные методы применяются в конструировании самых разных объектов информационных систем: язык UML используется для конструирования различных компонент информационных технологий, модель «сущность-связь» применяется для проектирования структур базы данных, графические построители используются для формирования запросов к данным, построения интерфейса и определения форм документов. В данной работе демонстрируется применение этого принципа для конструирования процедур анализа данных и оценивается эффект его применения в учебном процессе.

Современный этап развития информационных технологий характеризуется накоплением больших массивов учетной информации по самым разным областям деятельности, начиная с управления экономическими объектами и заканчивая социальными процессами. Компьютерный учет, электронная коммерция, цифровая экономика, социальные сети – практически все действия пользователей учитываются и сохраняются в компьютерных системах. Все это приводит к необходимости и возможности анализа этих данных для решения различных задач.

Для выполнения процедур анализа данных разрабатывается все больше математических моделей, алгоритмов и компьютерных технологий. Инструменты Data Mining активно применяются для обработки информации в информационных системах. Умение применять

соответствующие компьютерные технологии становится необходимой частью компетенций современного ИТ-специалиста или сотрудника, применяющего цифровую обработку данных для решения профессиональных задач. Это определяет необходимость включения в образовательные технологии формирование соответствующих компетенций. Актуальным при этом является возможность проектирования аналитической обработки с минимальным погружением в технические детали программирования.

Современные учебники и руководства по аналитическим технологиям часто содержат подробное описание алгоритмов машинного обучения. Для практического применения готового программного обеспечения нет необходимости в детальном знании таких алгоритмов. Применение моделей Data Mining на практике требует точного понимания исходных данных для машинного обучения, видов выделяемых зависимостей, их характеристик и возможностей практического применения. Фундаментальную роль в таком понимании играет не знание алгоритмов, а смысл полученных результатов.

Целью анализа данных является определение влияния входных переменных на выходные или исследование зависимостей между переменными. Важно понимать для чего выполняется анализ и как будут использоваться найденные зависимости. Анализ данных включает следующие этапы:

1) Определение статистических характеристик и многомерный анализ данных применяют для оценки качества наблюдений. На этом этапе, кроме основных характеристик, таких как среднее, среднеквадратичное отклонение, мода, медиана, исследуют законы распределения вероятностей переменных. Все это позволяет оценить разброс каждой переменной, выделить пропуски, выбросы и ошибки. Для исследования взаимного влияния переменных используют различные методы: корреляционный анализ, регрессионный анализ, а также многомерный анализ (OLAP технологию) – исследование изменения переменных по нескольким направлениям.
2) Преобразование данных может включать замену пропусков, преобразование шкал, замены категориальных шкал числовыми, масштабирование переменных (приведение значений к сопоставимым интервалам), исключение ошибок и выбросов. Снижение размерности возможно за счет количества переменных: преобразование координат (метод главных компонент), исключение входных переменных, не оказывающих влияния на выходные. Другой вариант уменьшения размерности – применение сэмплинга – уменьшения количества наблюдений за счет случайного выбора части данных.
3) Выделение зависимостей и оценка точности моделей. Достаточно часто можно применять разные модели для решения одной задачи. В этом

случае появляется возможность выбора наиболее подходящей модели. Оцениваются перспективы практического применения найденных зависимостей – может оказаться так, что сильные зависимости характерны для крайне малой части наблюдений.

По итогам проделанного исследования принимается решение о встраивании аналитической технологии в систему обработки данных предприятия.

Исследование зависимостей может не выявить каких-либо закономерностей просто потому, что выходные переменные не зависят от входных, либо зависимость сильно зашумлена случайными воздействиями. Высокие риски предложенного процесса проектирования аналитических технологий желательно компенсировать снижением затрат на программирование. Применение готового программного обеспечения существенно снижает затраты на компьютерные эксперименты, но не исключает их совсем. Для организации пробных расчетов нужно ознакомиться с языком программирования, изучить интерфейсы и порядок применения нескольких десятков объектов и процедур. Для учебного процесса в условиях ограничений на объемы учебной работы такие затраты уменьшают количество рассмотренных моделей и алгоритмов.

Для снижения затрат на программирование следует применять методы визуального проектирования процесса обработки. Удачным инструментом такого рода являются Power BI [4] и Orange3 [5]. Power BI Descktop является приложением, свободно распространяемым Microsoft. Применяя визуальный построитель, можно определять запросы табличных данных из разных источников (рис. 1), соединять их вместе для аналитической обработки, результаты которой можно визуализировать с помощью разных графических инструментов в стиле Dashboard, когда изменение параметра или выбор значения приводят к цепочке визуальных эффектов, иллюстрирующих выполненное действие. Каждый визуальный элемент при указании сопровождается демонстрацией соответствующих пояснений.

Power BI реализует, прежде всего, многомерный анализ, иллюстрируя его адекватной деловой графикой. Конечно, построение запросов, связывание таблиц и визуализация требуют знания соответствующих принципов и способов определения всех этих действий, а построенные процедуры реализуются специальными языками Power Query и DAX, в детальном изучении которых чаще всего нет необходимости.

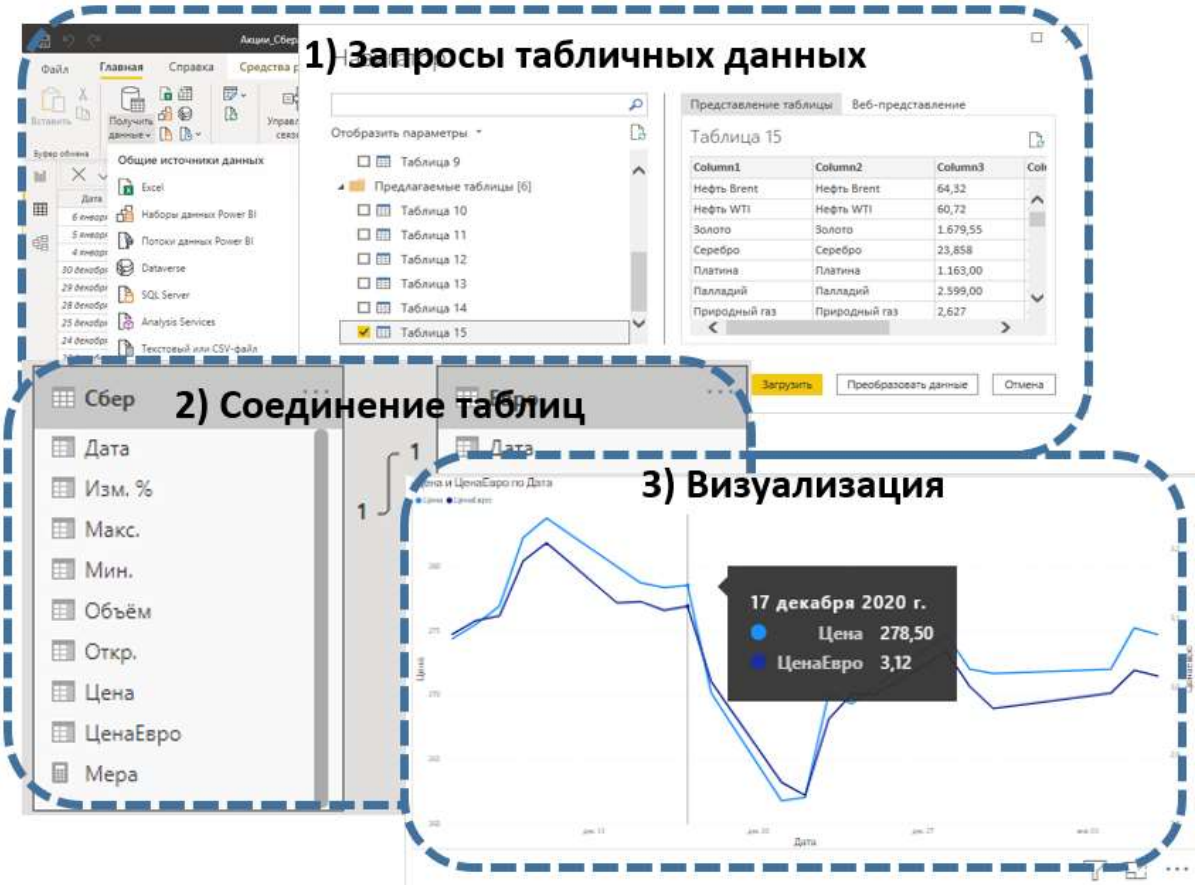

Рис. 1. Основные этапы анализа данных в PowerBI Descktop

В Orange обработка определяется блоками – виджетами (рис.2), соединенными линиями передачи данных, причем к данным могут относиться и применяемые модели. Связав виджеты в цепочку, разработчик задает процесс обработки данных, который может включать все рассмотренные этапы компьютерного эксперимента. На рис. 2 виджеты выполняют загрузку данных (File), замену пропусков (Impute), замену шкал (Continuize), настройку моделей классификации (Logistic Regression и Random Forest), измерение качества моделей (Test and Score) с визуализацией матрицы ошибок (Confusion Matrix) и ROC-кривой (ROC Analysis), выбор данных для прогнозирования (Select Columns) прогнозирование классов (Predictions) и сохранение результатов (Save Data). Обученные и настроенные модели могут быть сразу применены для обработки новых данных.

Применение этого подхода выявило недостаток свободно доступных образцов данных, которые можно применять для решения задач в разных областях. Известные репозитории (https://www.kaggle.com/datasets) таких данных, как правило, являются англоязычными.

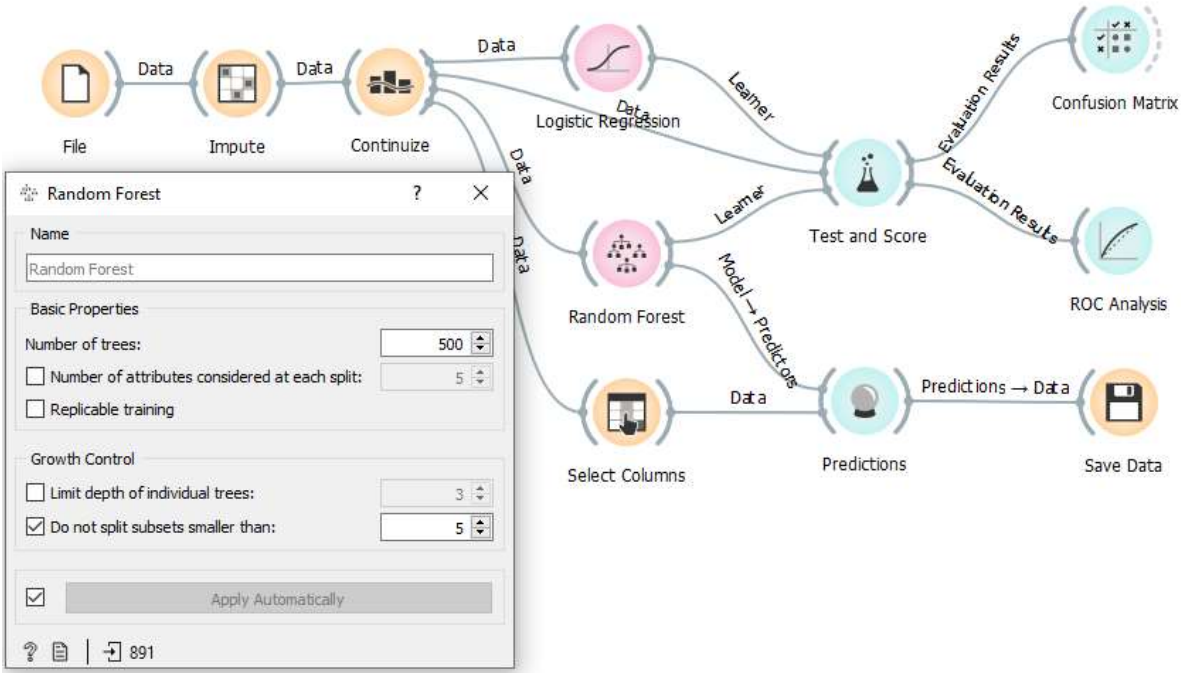

Рис. 2. Решение задачи классификации с помощью Orange

Применение методов и программ визуального конструирования аналитических технологий решает множество задач:

1) **В процессе обучения появляется возможность больше времени уделять моделям анализа и меньше техническим деталям применения соответствующего программного обеспечения.**
2) Теория иллюстрируется примерами успешного применения для решения практических задач.
3) Решение учебных задач существенно ускоряется.
4) Появляется возможность выполнения множества компьютерных экспериментов по конфигурированию и подбору параметров.
5) Визуализация результатов анализа позволяет точнее оценивать возможности и особенности применения настроенных моделей.

Список использованной литературы

1. Тихонова И.В., Иванов И.И., Омарова П.Г. Реализация принципа визуализации в процессе обучения // Проблемы современного педагогического образования. 2018. №60-1. URL: https://cyberleninka.ru/ article/n/realizatsiya-printsipa-vizualizatsii-v-protsesse-obucheniya.
2. Кротова И., Камоза Т., Донченко Н. Метод визуализации в системе инновационного обучения // Высшее образование в России. 2008. №4. URL: https://cyberleninka.ru/article/n/metod-vizualizatsii-v-sisteme-innovatsionnogo-obucheniya.

3. Буч Г. Язык UML. Руководство пользователя/ Г.Буч, А.Джекобсон, Д.Рамбо. М.: ДМК, 2000.
4. PowerBIBook.ru.– URL: https://powerbibook.ru/index.html
5. Orange documentation.– URL:https://orangedatamining.com/docs/

Информация об авторе

Братищенко Владимир Владимирович – кандидат физико-математических наук, доцент, ФГБОУ ВО «Байкальский государственный университет», 664003, г. Иркутск, ул. Ленина, 11, e-mail: vvb@bgu.ru

1. Тихонова И.В., Иванов И.И., Омарова П.Г. Реализация принципа визуализации в процессе обучения // Проблемы современного педагогического образования. 2018. №60-1. URL: https://cyberleninka.ru/article/n/realizatsiya-printsipa-vizualizatsii-v-protsesse-obucheniya.

2. Кротова И., Камоза Т., Донченко Н. Метод визуализации в системе инновационного обучения // Высшее образование в России. 2008. №4. URL: https://cyberleninka.ru/article/n/metod-vizualizatsii-v-sisteme-innovatsionnogo-obucheniya.

3. Буч Г. Язык UML. Руководство пользователя/ Г.Буч, А.Джекобсон, Д.Рамбо. М.: ДМК, 2000.

4. PowerBIBook.ru.– URL: https://powerbibook.ru/index.html

5. Orange documentation.– URL:https://orangedatamining.com/docs/