электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.9

Обзор аналитической платформы KNIME

***Сардарова М. Д.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Системы обработки информации и управления»

Научный руководитель: Гапанюк Ю.Е., к.т.н, доцент,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
gapyu@bmstu.ru

Современные компаниям необходимо хранить и обрабатывать огромные объемы разнородной информации, например публикации в социальных сетях, интернет-трафик, записи звонков в службу поддержки, электронные письма, показания датчиков. Для использования этих данных необходимо уметь их анализировать, часто в режиме реального времени.

Концепция Big Data – это серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов, которыми сложно управлять с помощью традиционных средств, в виду их структурных различий и большой скорости прироста.

Благодаря технологии Big Data у бизнеса появилось множество возможностей, например теперь можно узнавать почти все о поведении клиента. И на основании этих данных предлагать потребителю максимально подходящий продукт, разрабатывать новые продукты, проводить маркетинговые кампании, отслеживать действия конкурентов.

На сегодняшний момент существует немало систем анализа данных или аналитических платформ - KNIME, RapidMiner, Deductor, STATISTICA и др. Все они позволяют задействовать при анализе данных довольно большой набор методов как статистического, так и нестатистического характера.

В данной статье рассмотрим одну из платформ - KNIME .

KNIME представляет собой модульную платформу анализа данных.

KNIME обеспечивает более 1000 операции по анализу данных встроенными средствами или с помощью языка R и сторонних средств Weka для таких тем как

статистика, интеллектуальный анализ данных, веб-аналитика, интеллектуальный анализ текста, анализ социальных и медиа-данных.

Среда позволяет пользователю визуально создавать поток данных, выборочно выполнять анализ шагов, а затем исследовать результаты посредством интерактивного просмотра данных и моделей [1]. Общая схема работы с потоками данных представлена на рис. 1.

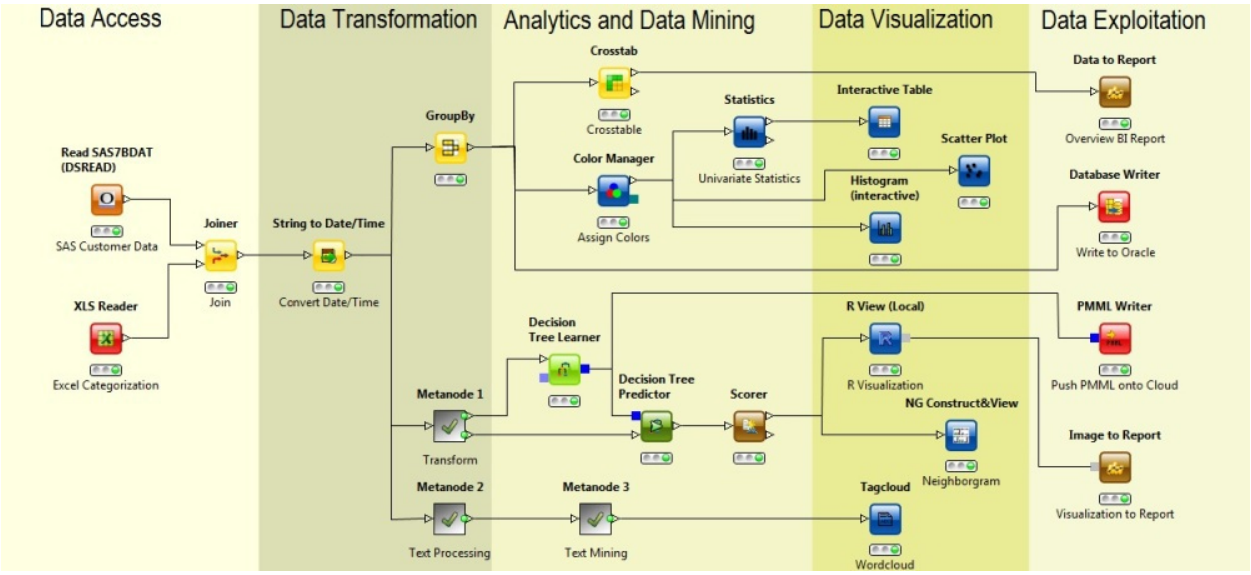


Рис. 1. Общая схема работы с потоками данных в среде KNIME

Потоки данных могут быть запущенны через интерактивный интерфейс пользователя и обрабатываться в пакетном режиме, упрощая интеграцию процесса анализа данных для менеджмента и выполнения на периодической основе.

Основная часть системы включает в себя сотни модулей для обработки данных:

- узлы доступа к данным – для чтения и записи данных из БД и файлов;
- узлы управления данными – для управления таблицами данных, которые используются между узлами;
- инструменты для построения графиков и графических изображения;
- поддержка циклов, временные ряды, дистанционные матрицы;
- узлы для статистики и интеллектуального анализа данных, такие как кластеризация, нейронные сети, деревья принятия решений.

Разработаны так же дополнительные плагины, которые реализуют методы для анализа теста и изображений.

Платформа KNIME написана на языке Java и базируется на IDE Eclipse. Система является open-source проектом и интегрируется с другими проектами, такими как набор средств и алгоритмов для анализа данных и решения задач прогнозировании Weka, пакеты

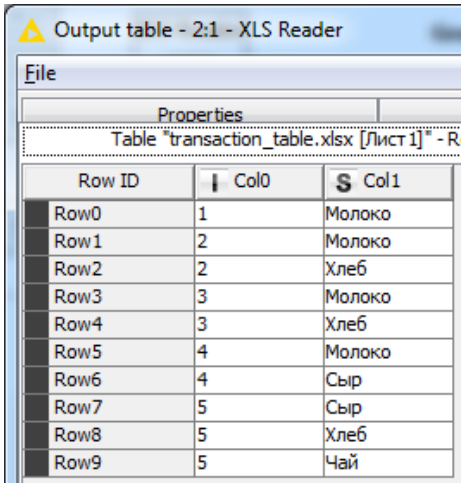
для проведения статистических расчетов на языке R, а так же LIBSVM, JFreeChart, ImageJ, Chemistry Development Kit [2]. А так же есть возможность расширять имеющийся функционал и добавлять необходимые узлы в качестве плагинов.

Рассмотрим несколько примеров моделирования и анализа в среде KNIME .

Алгоритм «Априори»

Алгоритм «Априори» применяется в интеллектуальном анализе для изучения последовательности событий. Для примера возьмем пары (или набор) товаров. Если человек приобрел один товар, то существует некоторая вероятность, что он купит и другой. Данный алгоритм применяется в продажах с целью увеличения сопутствующих покупок.

Входные данные – транзакционная БД (в Excel файле), где хранится id – номер транзакции и product – название товара (см. рис. 2)



Row ID	Col0	Col1
Row0	1	Молоко
Row1	2	Молоко
Row2	2	Хлеб
Row3	3	Молоко
Row4	3	Хлеб
Row5	4	Молоко
Row6	4	Сыр
Row7	5	Сыр
Row8	5	Хлеб
Row9	5	Чай

Рис. 2. Входные данные – транзакционная БД

Поток работ, реализующий предварительные шаги по чтению и группировке входных данных, а так же работу алгоритма, приведен на рисунке 3.

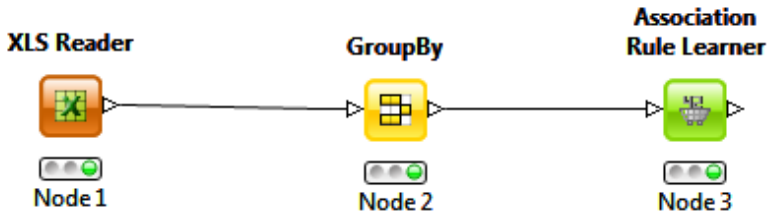
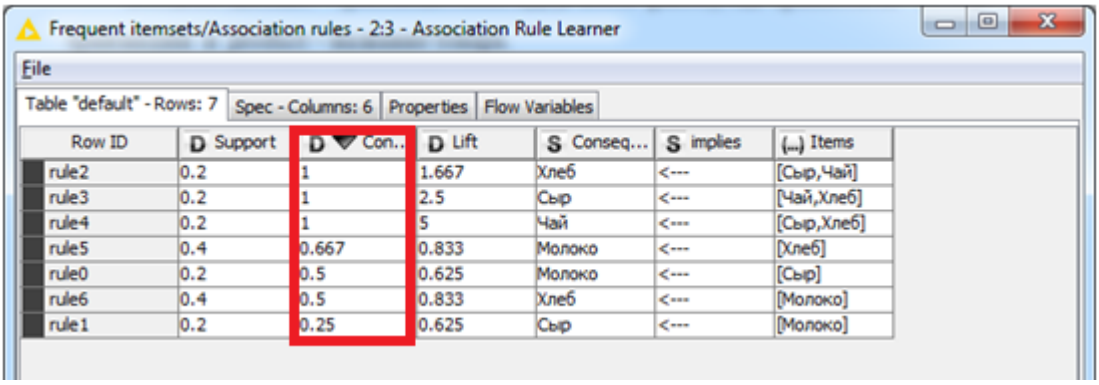


Рис. 3. Поток работ для реализации алгоритма «Априори»

Результаты представлены на рисунке 4. Здесь можно увидеть вероятность покупки определенных товаров при условии покупки других товаров.



Row ID	Support	Con..	Lift	Conseq...	implies	Items
rule2	0.2	1	1.667	Хлеб	<---	[Сыр,Чай]
rule3	0.2	1	2.5	Сыр	<---	[Чай,Хлеб]
rule4	0.2	1	5	Чай	<---	[Сыр,Хлеб]
rule5	0.4	0.667	0.833	Молоко	<---	[Хлеб]
rule0	0.2	0.5	0.625	Молоко	<---	[Сыр]
rule6	0.4	0.5	0.833	Хлеб	<---	[Молоко]
rule1	0.2	0.25	0.625	Сыр	<---	[Молоко]

Рис. 4. Результат работы алгоритма – вероятность приобретения товара

Кластеризация данных методом k-средних.

Кластеризация – это выделение групп объектов по сходным параметрам. Задача кластеризации является одной из фундаментальных в области интеллектуального анализа данных. Это задача находит применение во многих сферах – маркетинг, сегментация видео и изображений, анализ текстов, прогнозирование. Чаще всего, задача разбиения на кластеры является первым шагом при анализе данных. Затем для выделенных групп применяются другие методы.

Наиболее популярный метод – метод k-средних. Работа этого метода сводится к минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров [3].

На рисунке 5 изображен поток работ для кластеризации данных и ее визуального отображения.

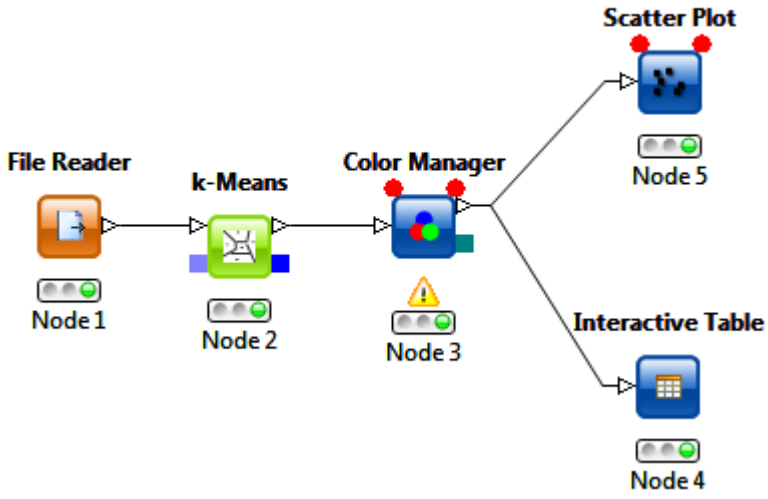


Рис. 5. Поток работ для кластеризации данных методом k-средних

На рисунке 6 изображен график разбиения данных на кластеры, каждый цвет – отдельный кластер.

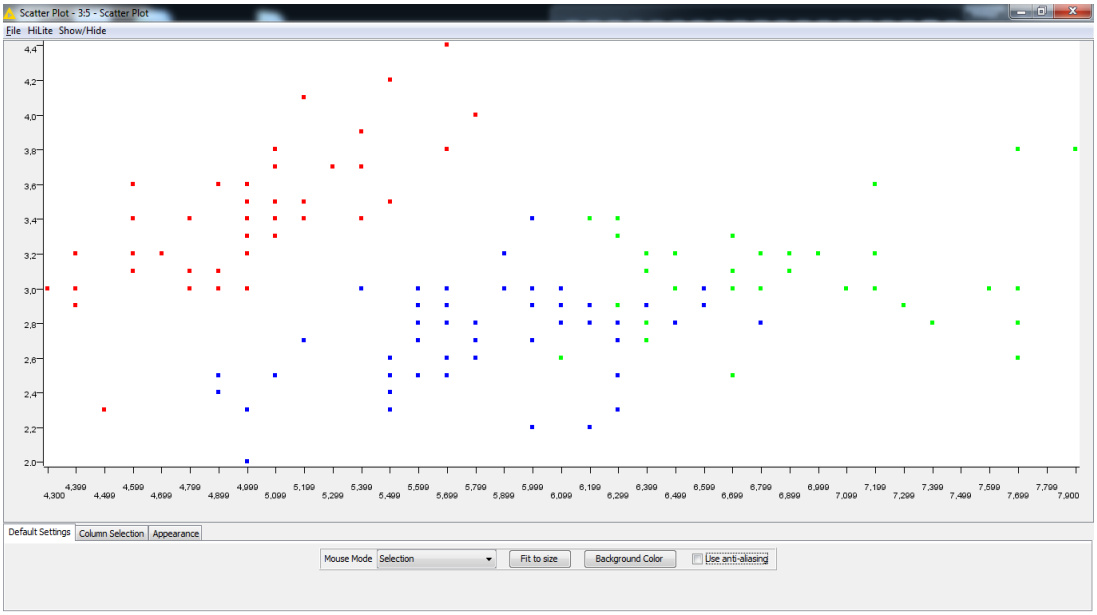


Рис. 6. График разбиения данных на кластеры

Заключение

Платформа KNIME предоставляет широкие возможности для анализа различного рода данных при помощи визуальных компонент, поэтому от пользователя требуются минимальные знания языков программирования или не требуются вовсе. Это делает работу по анализу данных более эффективной и ориентированной для широкого круга пользователей.

Список литературы

1. Официальный сайт платформы KNIME. Режим доступа: http://www.knime.org (дата обращения 26.03.2015).
2. Веб-издание Predictive Analytics Today. Режим доступа: http://www.predictiveanalyticstoday.com (дата обращения 25.03.2015)
3. Чубукова И. А. Data Mining. Интернет-Университет Информационных Технологий. 2-е изд., испр. М.; БИНОМ. Лаборатория знаний. 2008. 382с.
4. Терехов В.И. Применение когнитивной компьютерной графики для визуализации актуальной информации лицам, принимающим решение // Инженерный журнал: наука и инновации. Электронное научно-техническое издание. № 3(3). 2012. Режим доступа: http://engjournal.ru/articles/111/111.pdf (дата обращения 21.03.2015).
5. Виноградова М.В., Игушев Э.Г. Принципы организации структуры данных с произвольным доступом и быстрой операцией вставки или удаления // Инженерный

журнал: наука и инновации. Электронное научно-техническое издание, выпуск: № 3(3). 2012. Режим доступа: http://engjournal.ru/articles/101/101.pdf (дата обращения 22.03.2015).