СОВРЕМЕННЫЕ МЕТОДЫ И ТЕХНИКИ АНАЛИЗА И ОБРАБОТКИ ДАННЫХ

Козинец А.Н., магистрант гр.976701

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Матвейчук Н.М. – канд. физ.-мат. наук, доцент

Аннотация. В работе представлен анализ основных методов и техник применяемых для анализа и обработки данных.

Ключевые слова. Аналитика, данные, статистика, визуализация, искусственный интеллект, машинное обучение, визуализации.

В наш век больших объемов данных понимание того, как анализировать и извлекать истинный смысл из цифровых данных окружающих нас, является одним из основных факторов успеха.

Несмотря на колоссальный объем данных, которые мы создаем каждый день, всего несколько процентов из них анализируется и используется для обнаружения, улучшения и аналитики. Хотя эти несколько процентов могут показаться незначительным, учитывая количество цифровой информации, которую мы имеем под рукой, эти данные по-прежнему представляют собой огромные объемы информации.

Существуют разные методы анализа данных в зависимости от поставленного вопроса, типа данных и количества собранных данных. Каждый из них фокусируется на стратегиях использования новых данных, анализа и детализации информации для преобразования фактов и цифр в параметры принятия решений. Эти процедуры позволяют нам сделать основной вывод на основе данных, устраняя ненужный хаос, создаваемый остальной частью данных.

Можно выделить два основных метода анализа данных таких как качественный и количественный. Качественный метод анализа в основном решается с помощью количественных методов, таких как анкеты, шкала отношения, опросы и многое другое. В количественном же анализе данные представлены в виде шкал измерений и расширяются для большего количества статистических манипуляций. К другим методам можно отнести такие как: анализ текста, статистический анализ, диагностический анализ, прогностический анализ и предписательный анализ.

Генерация данных — это не останавливающийся процесс, в котором сбор и анализ данных выполняются одновременно, часто при помощи комбинирования разных методов. Обеспечение целостности данных - один из важнейших компонентов анализа данных, но не менее важный как применение правильных техник для конкретно поставленных задач. Техники анализа данных можно разделить на следующие категории[1]:

- основанные на математике и статистике;
- основанные на искусственном интеллекте и машинном обучении;
- основанные на визуализации и графиках.

Разбирая каждую из этих категорий можно выделить основные методы применяемые в каждой из них. При использовании методов, основанных на математике и статистике как минимум 3 типа анализа выходят на первые позиции такие как дисперсионный, описательный и регрессивный. Описательный анализ принимает во внимание исторические данные, ключевые показатели эффективности и описывает производительность на основе выбранного эталона. Он учитывает прошлые тенденции и то, как они могут повлиять на то что будет в будущем. Дисперсионный же анализ представляет собой дисперсию в области, на которую распространяется набор данных. Этот метод позволяет аналитикам данных определять изменчивость изучаемых факторов. Метод регрессионного же анализа позволяет аналитикам данных определять изменчивость изучаемых факторов.

Искусственный интеллект — это довольно новое явление представляющее собой свойство интеллектуальных систем выполнять творческие функции. Искусственные нейронные сети, не смогли обойти в том числе и сферу аналитики данных. Современные системы искусственного интеллекта, могут хорошо обрабатывать данные имеющие слабую структуру или же вообще не имеющие ее. Такие методы очень надежны в приложениях для прогнозирования, а в качестве примера применения можно выделить целую сферу - маркетинга[2]. К методам основанным на искусственном интеллекте и машинном обучении так же можно отнести дерево решений. Как следует из названия, данная техника представляет такие модели как классификации и регрессии. Данные метод разделяет набор данных на более мелкие подмножества, одновременно развиваясь в связанное дерево решений. Не стоит также забывать про нечеткую логику. Это метод анализа данных, основанный на вероятности, который помогает справляться с неопределенностями в методах интеллектуального анализа данных.

Техники основанные на визуализации и графиках представляют собой наборы диаграмм. Столбчатая диаграмма и гистограмма используются для представления числовых различий между категориями. Линейная же диаграмма представления изменения данных за непрерывный интервал времени. Круговая диаграмма используется для представления соотношения различных классификаций. Она подходит в основном только одной серии данных, но ее можно сделать многоуровневой, чтобы отображать долю данных в разных категориях.

Анализ данных является ключом к любой сфере деятельности, будь то запуск нового предприятия, принятие маркетинговых решений, управления человеческими ресурсами или же продумывания дальнейшей стратегии развития бизнеса. Выводы и статистические вероятности, рассчитанные на основе анализа данных, помогают принимать наиболее важные решения, исключая всякую человеческую предвзятость. Различные аналитические методы имеют местами совпадающие функции и разные ограничения, но они также дополняют друг друга и использования их вместе позволяет значительно упростить процесс взаимодействия. Прежде чем выбрать методы анализа данных, важно принять во внимание такие факторы как объем работы который нужно сделать, целесообразность, цель и задачу. На основании проведенного анализа целей, можно выбрать оптимальные решения пригодные для конкретно поставленной задачи или же разработать сложную структуру в которой будут сообща использоваться разные методы обработки, анализа, поиска, визуализации данных для достижения общей цели.

Список использованных источников:

1. Big Data: A Revolution That Will Transform How We Live, Work, and Think / Kenneth Cukier, Viktor Mayer-Schönberger // An Hachette UK, 2013.

2. Artificial Intelligence for Big Data: Complete guide to automating Big Data solutions using Artificial Intelligence techniques / Anand Deshpande, Manish Kumar // Packt Publishing, 2018.