Смирнов Е.В. Описание проблемы обработки и использования результатов анализа больших данных (Big Data) // Информационно-экономические аспекты стандартизации и технического регулирования, 2018. № 6(46).

УДК 004.85

ОПИСАНИЕ ПРОБЛЕМЫ ОБРАБОТКИ И ИСПОЛЬЗОВАНИЯ РЕЗУЛЬТАТОВ АНАЛИЗА БОЛЬШИХ ДАННЫХ (BIG DATA)

Смирнов Е.В., соискатель, ФГБУН ВИНИТИ РАН

Аннотация: Широкий спектр задач, решаемый нейронными сетями сегодня, первостепенно базируется на исходных данных для этих сетей. Качество, актуальность, данных и скорость их обработки определяют конечный результат. Увеличение объемов информации и возрастающая тенденция к анализу больших данных требует снижения временных и трудозатрат на формирование первичной информации для анализа и заставляет искать новые технологии обработки данных.

Ключевые слова: искусственные нейронные сети, большие данные, построение модели.

УДК 004.85

DESCRIPTION OF THE PROBLEM OF PROCESSING AND USING THE RESULTS OF THE ANALYSIS OF BIG DATA (BIG DATA)

Smirnov E.V., applicant, FGBUN VINITI RAS

Abstract: A wide range of tasks solved by neural networks today, is primarily based on the source data for these networks. The quality, relevance, data and speed of their processing determine the final result. The increase in the volume of information and the increasing trend towards the analysis of big data require the reduction of time and labor costs for the formation of primary information for analysis and makes it necessary to look for new data processing technologies..

Keywords: artificial neural networks, big data, model building.

Говоря об интеллектуальных информационных технологиях невозможно обойти такое понятие как искусственные нейронные сети (ИНС), сформировавшееся в 40-е годы. Моделирование системы как модели мозга с множеством нейронов, принимающих на вход ряд входных параметров с определенными коэффициентами перед переменными или «весами», позволило решать задачи математически не формализованные, имеющие

большие объемы входной информации, характеризующиеся избыточностью или перенасыщением данных, а также обладающие высоким показателем шума в них.

Эффективность ИНС бесспорна и не маловероятно, что экстенсивный рост их применения в различных сферах можно будет наблюдать уже в ближайшие годы, а их влияние затронет все возможные стороны жизни как общества в целом, так и частных его элементов. В тоже время, на сегодняшний день ИНС являются лишь подобием искусственного интеллекта, а на данном этапе являются частным решением прикладных задач [1]. Следствием этого является ряд проблем, возникающих при работе с ИНС, которые могут привести не только к искажению модели, но и к ее полной несостоятельности. Основными проблемами можно выделить:

– проблему подготовки данных достаточного объема и качества для обучения ИНС [2];

– поиск «золотой середины» формируемой модели, во избежание таких проблем как «overfitting» или «переобучение модели», а также «underfitting» или проблема «недостатка данных». Эти и ряд других проблем компенсируются большими временными и трудозатратами для формирования модели, обеспечивающей оптимальное качество прогнозирования;

– интерпретация выходных результатов специалистом может быть затрудненно, т.к. процедура решения поставленной задачи не является очевидным;

– «сигмоидальный характер передаточной функции нейрона может явиться причиной того, что если в процессе обучения несколько весовых коэффициентов станут слишком большими, то нейрон попадёт на горизонтальный участок функции в область насыщения; при этом изменения других весов, даже достаточно больших, практически не сказываются на величине выходного сигнала такого нейрона, а значит, и на величине целевой функции» [3];

– свойства моделируемых ИНС обусловленные частными особенностями.

Если все указанные проблемы за исключением первой сводятся к формированию ИНС и ее «обучению», а также интерпретации получаемых результатов с внесением изменений в модель различными методами, то качество исходных данных, является фундаментом модели. Ежегодный рост объемов информации вплотную подводит к понятию Big Data.

Понятия Big Data достаточно полно описывается следующим определением: «Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.» [4] дающим нам понимание, что big data – это термин, описывающий большие объемы высокоскоростных, сложных и переменных данных, требующих передовых методов и технологий для сбора, хранения, распространения, управления и анализа информации. Объемы информации, формирующие big data могут составлять гига-, тера-, пета- и даже зета-байты информации, и проблема сбора данных трансформировалась в проблему их обработки. Это вынуждает исследователей и специалистов в сфере машинного обучения искать способы и инструменты для анализа больших массивов данных, позволяющих исключить случайные зависимости, возникающие при применении традиционных подходов к анализу массивов данных, выявлять новые подходы к технологиям подготовки данных к анализу, получения доступа к всё большим вычислительным мощностям [5].

Процедура сбора и формирования объемов данных в big data за редким исключением не находятся под управлением специалиста, осуществляющего их анализ. Отсутствие этой возможности порождает ряд вопросов требующих решения перед дальнейшей работой: соблюдается ли целостность собранной информации и не исключается ли часть данных в процессе записи, какое влияние на данные вносит инструмент их собирающий, сохраняется ли

«однородность» данных и уровень вносимых «шумов», одинаков ли процесс сбора данных в каждый момент времени.

Существующие технологии анализа данных традиционно используют ряд следующих инструментов: Online analytical processing (OLAP) технология комплексного многомерного анализа данных, описанная Эдгаром Коддом, регрессионный анализ, кластерный анализ.

OLAP используется для анализа данных представленных в виде многомерного куба. Данная структура рабочих данных называется OLAP-кубом. Количество «мер» или измерений куба соответствует количеству атрибутов, описывающих данные, в то время как каждое измерение характеризуется соответствующими параметрами атрибутов. Удобство OLAP-кубов можно продемонстрировать, определив понятия среза, когда необходимо выполнить фильтрацию по определенным осям и понятие проекции, агрегируя данные в кубе на определенную ось (проецируя ее). Сложность применения OLAP состоит в обращениях к базам данных и требованиях к полноте и корректности данных. Так, при работе с большими данными может быть неопределимым оптимальное распределение базы данных по серверам.

Помимо базовой технологии OLAP существуют еще концепции: Multidimensional OLAP (MOLAP), Relation OLAP (ROLAP), Hybrid OLAP (HOLAP).

Регрессионный анализ служит для определения влияния независимых переменных на зависимую. Это достигается построением параметрических функций изменения числовых величин от времени. Корректировка функции осуществляется подстройкой ее для уменьшения стандартной ошибки. При работе с большими данными использование регрессионного анализа может быть распараллелено между несколькими вычислительными объектами. Операция вычисления стандартной ошибки разделяется прямым способом, корректировка параметров функции, основанная на градиентном спуске,

может быть распараллелена по причине вычислений частных производных каждого отдельного параметра. Эти вычисления базируются на дискретном дифференцировании, которые могут быть распараллелены, т.к. основаны на вычислении взвешенных сумм.

Кластерный анализ лучше всего описать как задачу разбиения «множества информационных сущностей на группы, при этом члены одной группы более похожи друг на друга, чем члены из разных (классификация относит каждый объект к одной из заранее определенных групп). Проблема кластеризации Big Data состоит в том, что имеющиеся алгоритмы предполагают возможность непосредственного обращения к любой информационной сущности в исходных данных (заранее невозможно предугадать, какие именно сущности понадобятся алгоритму). Решение проблемы может быть следующим. На каждом сервере запускается свой алгоритм, который оперирует только данными этого сервера, а на выходе дает параметры найденных кластеров и их веса, оцениваемые исходя из количества элементов внутри кластера. Затем полученная информация собирается на центральном сервере и производится метакластеризация - выделение групп близко расположенных кластеров с учетом их весов». [6]. Это приводит к выводу неприменимости кластеризации в прямом виде для анализа big data.

Достоинства и недостатки методов анализа приведены в таблице 1.

Таблица 1

Достоинства и недостатки методов анализа

Метод	**Достоинства**	**Недостатки**
MOLAP	– Высокая производительность относительно реляционных баз данных – Соответствие структуры и интерфейсов является наилучшим со структурой аналитических запросов – Легкая интеграция дополнительных функций	– Эффективность использования внешней памяти довольно низкая, относительно реляционных баз данных механизмы транзакции хуже – Отсутствие единых стандартов на языки описания и управления данными и интерфейсы – Не поддерживаемая репликация данных

Метод	Достоинства	Недостатки
ROLAP	– Возможность работы с очень большими БД – Возможность производить анализ непосредственно над хранилищем посредством инструментов – При условии изменяющейся размерности задач динамическое представление размерности является оптимальным решением (не требуется физическая реорганизация БД) – Меньшие требования к клиентским станциям – Высокий уровень защиты данных и лучшее разграничение прав доступа [7]	– Ограниченные возможности расчета значений функционального типа – Производительность по сравнению с MOLAP ниже
Регрессионный анализ	– Простота вычислительных алгоритмов – Наглядность результатов модели	– Низкая точность прогноза – Выбор вида зависимости является субъективным
Кластерный анализ	– Возможность производить разбиение по набору признаков – Отсутствие ограничений на виды изучаемых объектов	– Требуется возможность непосредственного обращения к любой информационный сущности в исходных данных [8] – Зависимость состава и количества кластеров от выбираемых критериев разбития

Таким образом, на сегодняшний день не существует единой универсальной методологии анализа данных в big data. Это требует от специалистов траты до 80% общего времени для получения пакета данных, пригодного для использования в ИНС. Отсутствие унифицированного метода анализа данных требует индивидуальный подход к каждому типу решаемых задач. Сейчас этот недостаток пытаются устранить увеличением скорости обмена данными между распределенными носителями данных, улучшением инструментов для фиксирования «равномерных» значений и снижения уровня шумов, но это не является решением долгосрочной проблемы подготовки данных.

Для решения этой задачи необходима разработка новой технологии, учитывающей преимущества и недостатки имеющихся методов, а также обладающей некой универсальностью по направлениям, обеспечивающей подготовку и анализ исходных данных.

Список использованных источников и литературы

1. Евтушенко Г.И., Отраднов К.К. К проблеме познания в нейронных сетях // Теоретические и прикладные аспекты современной науки. 2015. №7-3. С. 29
2. Бадамшин Р.А., Ильясов Б.Г., Черняховская Л.Р. Проблемы управления сложными динамическими объектами в критических ситуациях на основе знаний. – М.: Машиностроение, 2003. С. 240.
3. Охтилев М.Ю., Соколов Б.В., Юсупов Р.М. Интеллектуальные технологии мониторинга состояния и управления структурной динамикой сложных технических объектов. Москва. 2005. С. 90.
4. TechAmerica Foundation’s Federal Big Data Commission. (2012). Demystifying big data: A practical guide to transforming the business of Government. Retrieved from http://www.techamerica.org/Docs/fileManager.cfm ?f=techamerica-bigdatareport-final.pdf
5. Волков В.В., Скугаревский Д.А., Титаев К.Д. Проблемы и перспективы исследований на основе Big Data (на примере социологии права), Социологические исследования, 2016, № 1 (381). С. 48-58.
6. Магеррамов З.Т., Абдуллаев В.Г., Магеррамова А.З., Big Data: проблемы, методы анализа, алгоритмы // Радиоэлектроника и информатика. 2017.
7. Альперович М. Технологии хранения и обработки корпоративных данных (Data Warehousing, OLAP, Data Mining)
8. Казиев Г. З., Курдюков В.В. Модели и методы кластеризации big data для их анализа и обработки, Современный взгляд на будущее науки: приоритетные направления и инструменты развития, Санкт-Петербург, 2017, С. 155