УДК 004.89

***В.С. Лебедев*[1]**

[1] *Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация*

ИНДУКТИВНЫЙ ВЫВОД В АНАЛИЗЕ БОЛЬШИХ ДАННЫХ

Аннотация. Статья посвящена анализу больших данных и одному из его направлений интеллектуальному анализу данных. Проведен обзор некоторых методов интеллектуального анализа данных и акцент сделан на методе индуктивного вывода.

В статье приведено описание активно развивающейся технологии анализа больших данных, описаны характеристики, применяющиеся для описания больших данных. Представлен интеллектуальный анализ данных и несколько его методов, применяемых в технологиях BigData. Описан индуктивный вывод как метод получения новых знаний, его виды и их отличия и особенности.

Ключевые слова: большие данные, анализ больших данных, методы интеллектуального анализа данных, индуктивный вывод, индукция.

***V.S. Lebedev*[1]**

[1] *Irkutsk State Transport University, Irkutsk, Russian Federation*

INDUCTIVE INFERENCE IN BIG DATA ANALYSIS

Abstract. The article is devoted to big data analysis and one of its directions is data mining. Some methods of data mining are reviewed and emphasis is placed on the method of inductive inference.

The article describes the actively developing technology of big data analysis, describes the characteristics used to describe big data. The intellectual analysis of data and several of its methods used in BigData technologies are presented. Inductive inference is described as a method of obtaining new knowledge, its types and their differences and features.

Keywords: big data, big data analysis, data mining methods, inductive inference, induction.

Введение

За последнее десятилетие в отечественной и зарубежной научной литературе понятие «больших данных» широко употребляется в различных выражениях («анализ больших данных», «инженерные проблемы больших данных», «аналитика больших данных», «большие данные в логистике», «перспективы больших данных» и др.). Термин «большие данные» или «Big Data» обрел широкую известность совсем недавно – согласно «Google Trends», уровень его употребления резко возрос в 2011 году.

Большие данные – это колоссального объема массивы информации, которые обладают высокой скоростью накопления и могут быть представлены как в виде структурированной, так и неструктурированной информации. Структурированная информация представлена в универсально понятной форме, удобной для обработки и анализа. Неструктурированная информация не обладает таким свойством. Вместе с этим большие данные также включают в себя совокупность инновационных и передовых методов хранения и обработки информации для автоматизации и оптимизации бизнес-процессов.

Таким образом, большие данные можно характеризовать тремя основными особенностями:

1. Большой объем информации;
2. Высокая скорость изменения информации;
3. Разнообразие и разнородность данных.

Технологии BigData позволяют обработать большой объем неструктурированных данных, систематизировать их, проанализировать и выявить закономерности там, где человеческий мозг никогда бы их не заметил. Это открывает совершенно новые возможности по использованию данных [1].

Как правило, для описания больших данных используются характеристики, представленные ниже.

1. Объем (Volume) – количество собранных и хранящихся данных. Объем данных определяет могут ли данные рассматриваться как большие, а также возможную ценность и потенциал данных.

2. Разнообразие (Variety) – тип данных. Вид представления у больших данных может иметь разные формы, такие как текст, изображение, аудио, видео. Данные могут быть структурированными, неструктурированными или структурированными частично. Сложность структуры определяет возможность анализа.

3. Скорость (Velocity) – скорость, с которой происходит накопление и обработка данных. Обычно работа с большими данными осуществляется в режиме реального времени.

4. Изменчивость (Variability) - потоки данных могут иметь пики и спады, сезонности, периодичность. В ходе управления неструктурированной информацией могут возникать трудности при возникновении всплесков данных.

5. Достоверность (Veracity) – данные могут поступать из большого числа источников из чего следует, что достоверность и качество данных напрямую оказывают влияние на результат анализа данных и его дальнейшее применение. В случае использования недостоверных данных для анализа результаты этого анализа не будет представлять какой-либо ценности для практического применения.

Помимо вышеперечисленных сейчас в рассмотрение добавляют еще пару характеристик.

1. Ценность (Value) – потенциальная ценность больших данных очень высока. На ценность оказывают влияние вышеперечисленные признаки. Наибольшую ценность имеют данные, которые могут быть использованы для решения конкретных задач либо которые способствуют получению новых сведений или идей.

2. Визуализация (Visualization) – визуальное описание и представление информация получаемой в результате анализа данных.

Все вышеперечисленные характеристики именуются как семь «V» («7-V») и при полном описании больших данных должны применяться все в совокупности для более точного определения рассматриваемого набора больших данных.

В настоящее время сверхсвязанный мир генерирует огромные объемы данных, хранящихся в компьютерной базе данных и облачной среде. Эти большие данные необходимо анализировать, чтобы извлечь полезные знания и представить их лицам, принимающим решения, для дальнейшего использования [2].

Интеллектуальный анализ данных

Под интеллектуальным анализом данных (ИАД) понимают обработку информации и выявление в ней тенденции, которая помогает принимать решения. Существует множество различных методов интеллектуального анализа данных, моделирования запросов обработки и сбора информации [3].

ИАД как направление обработки данных сформировался в результате объединения технологий анализа данных и искусственного интеллекта.

Основными методами ИАД в первую очередь являются методы, основанные на переборе. Обычный перебор всех вариантов занимает $O(2^N)$ операций (где N – общее количество объектов), из этого следует, что с ростом числа объектов перебора вычислительная сложность работы алгоритма растёт экспоненциально.

С целью уменьшения количества переборов и снижения вычислительной сложности таких алгоритмов, в методах интеллектуального анализа данных применяют различные эвристические подходы.

В результате применения методов интеллектуального анализа данных можно получить такие типы знаний как:

- математические функции;
- кластеры;

- деревья решений;
- ассоциативные правила.

Основными задачами, решаемыми с помощью методов ИАД, являются регрессионный анализ, задача классификации, задача кластеризации и поиск ассоциативных правил в базах данных.

К методам Data Mining относятся также методы статистики – корреляционный и регрессионный анализ, дисперсионный анализ, факторный анализ, анализ временных рядов, индуктивный вывод и др.

Корреляционный анализ – это количественный метод для определения тесноты и определения взаимосвязи между определенными переменными. Корреляция определяет отклонение для значений переменной. Она может быть парной или множественной, прямой или обратной.

Регрессионный анализ – это количественный метод для установления вида математической функции в причинно-следственной зависимости между переменными величинами. Основной целью корреляционного и регрессионного анализа является определение характера и степени между случайно или неслучайно изменяемыми величинами.

Главная цель дисперсионного анализа заключается в проверке статистической значимости различий между средними значениями переменных или групп переменных. Данная проверка осуществляется с помощью разбиения общей дисперсии на части, одна из которых обусловлена случайной ошибкой, а вторая связана с различием средних значений. Дисперсионный анализ наиболее эффективен для малых выборок.

Факторный анализ – это статистический метод, применяющийся для описания непостоянства значений выбранных коррелированных переменных в определениях потенциально меньшего числа ненаблюдаемых переменных, называемых факторами. Данный метод позволяет изучать взаимосвязи между значениями переменных и сокращать число переменных требуемых для описания данных.

Анализ временных рядов включает в себя методы анализа данных временных рядов с целью извлечения из них значимых статистических данных и других характеристик данных. Прогнозирование временных рядов – это использование модели для прогнозирования будущих значений на основе ранее наблюдаемых значений.

Индуктивное рассуждение — это метод рассуждения, при котором совокупность наблюдений рассматривается как вывод общего принципа. Оно состоит в том, чтобы делать широкие обобщения, основанные на конкретных наблюдениях.

Индуктивный вывод

Индукция – процесс логического вывода на основе перехода от частного положения к общему.

Различают полную индукцию, когда обобщение относится к конечно-обозримой области фактов, и неполную индукцию, когда оно относится к бесконечно или конечно-необозримой области фактов [4].

Схема полной индукции: Множество **A** состоит из элементов: a1, a2, a3, …, an.

a_1 имеет признак **B**

a_2 имеет признак **B**

Все элементы от a_3 до a_n также имеют признак **B**

Следовательно Все элементы множества **A** имеют признак **B**.

Рис. 1. Схема полной индукции

В полной индукции умозаключение формируется на основании полного рассмотрения элементов исследуемого множества и распространяется абсолютно на все множество. Умо-

заключение, полученное данным видом индукции, можно считать достоверным на основании изучения всех элементов.

Схема неполной индукции: Множество **A** состоит из элементов: a1, a2, a3, … ak, … an.

a_1 имеет признак **B**

a_2 имеет признак **B**

Все элементы от a_3 до a_k также имеют признак **B**

Следовательно Вероятно, a_{k+1} и остальные элементы множества **A** имеют признак **B**.

Рис. 2. Схема неполной индукции

Умозаключение по схеме неполной индукции выводится на основе исследования части объектов рассматриваемого множества и распространяется на все множество. Неполная индукция может приводить к ошибочным заключениям так как с точки зрения формальной логики она не является доказательной. Однако, не смотря на подобный недостаток именно данный вид индукции является основным способом получения новых знаний в виду ограничений, возникающих при исследованиях в различных предметных областях. Заключение, полученное с применением неполной индукции, имеет вероятностный характер и может потребовать приведения дополнительных аргументов в свое доказательство, но при этом получить его гораздо проще чем при полной индукции.

Кроме этого, неполная индукция в свою очередь подразделяется на два вида: популярную и научную.

Логическое умозаключение о всем множестве элементов на основании информации о некоторых элементах подмножества исходного множества при условии, что отсутствуют случаи, противоречащие получаемому заключению, называется популярной индукцией. Данный вид индукции распространен в повседневной жизни общества поэтому и имеет такое наименование. Заключения, полученные данные видом неполной индукции, могут быть ошибочными из-за субъективных суждений конкретного человека или группы людей, которым они принадлежат.

Научная индукция – это вывод о всем множестве элементов на основании информации об элементах и причинных связях подмножества. Отличием научной индукции от популярной является рассмотрение выборки элементов с учетом возможных взаимосвязей элементов внутри множества или подмножества. Вероятность или правдоподобность научной индукции выше, чем популярной, так как в ней общий вывод о всем классе предметов или явлений осуществляется на основе имеющихся знаний о необходимых признаках и причинно-следственных связей лишь некоторых предметов или явлений данного класса, а не выбранных случайно или без определённой цели. Как понятно из названия данный вид индукции применяется в науке и может быть использован для получения достоверных умозаключений в анализе больших данных.

В конце XIX века Джон Стюарт Милль сформулировал основные принципы индуктивного рассуждения в процессе опытного исследования: методы сходства, различий, сопутствующих изменений и остатков. На основе этих принципов строятся все индуктивные умозаключения. Смысл принципов заключается в установлении причинно-следственных отношений, что основывается на идеях выделения сходства и различия в наблюдаемых ситуация окружающего мира. Указанные выше методы, как правило, применяются не изолировано, а в сочетании друг с другом, что в некоторой степени повышает правдоподобность получаемых логических умозаключений.

Важно заметить, что индуктивные выводы в научном творчестве выступают в единстве с дедуктивными выводами и умозаключениями по аналогии. Поэтому процесс обоснования истины нельзя строить только на одной форме вывода [5]. В некоторых случаях комбинации

методов способны давать результаты значительно превышающие возможности применяемых методов по одному.

Заключение

В завершении можно сделать следующие выводы:

1. Анализ больших данных новое и активно развивающиеся направление в области обработки больших массивов неструктурированной информации. На данный момент существует огромное множество методов применяемых в анализе больших данных. Число этих методов постоянно увеличивается.

2. Для решения поставленных задач в анализе больших данных применяются различные методы интеллектуального анализа данных, одним из таких методов является индуктивный вывод.

3. Индуктивный вывод позволяет строить правдоподобные умозаключения о классе предметов или явлений на основе выборки объектов этого класса. Это значит, что данный метод может применяться в анализе больших данный с целью выявления зависимостей в огромных объемах данных. Выявленные зависимости могут содержать новые важные знаний об исследуемой предметной области, в том числе и ассоциативные зависимости, которые ранее не были определены при исследование конкретного объекта или явления.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Веретенников, А. В. BigData: анализ больших данных сегодня / А. В. Веретенников. — Текст: непосредственный // Молодой ученый. — 2017. — № 32 (166). — С. 9-12. — URL: https://moluch.ru/archive/166/45354/ (дата обращения: 01.05.2022).

2. Ait-Mlouk, A., Agouti, T. & Gharnati, F. Интеллектуальный анализ и приоритизация ассоциативных правил для больших данных: многокритериальный подход к анализу решений. J Big Data 4, 42 (2017). https://doi.org/10.1186/s40537-017-0105-4 (дата обращения: 01.05.2022).

3. Певченко, С. С. Методы интеллектуального анализа данных / С. С. Певченко. — Текст: непосредственный // Молодой ученый. — 2015. — № 13 (93). — С. 167-169. — URL: https://moluch.ru/archive/93/20875/ (дата обращения: 02.05.2022).

4. Советский энциклопедический словарь, Москва, издательство «Советская энциклопедия», 1981

5. Рахматуллин, Р. Ю. Истина как философская категория / Р. Ю. Рахматуллин. — Текст: непосредственный // Молодой ученый. — 2014. — № 13 (72). — С. 332-335. — URL: https://moluch.ru/archive/72/12329/ (дата обращения: 11.07.2022).

REFERENCES

1. Veretennikov, A. V. BigData: analiz bol'shikh dannykh segodnya [Big Data: Big Data analysis today]. Molodoy uchenyy [Young scientist], 2017, No 32 (166), pp 9-12.

2. Ait-Mlouk, A., Agouti, T. & Gharnati, F. Intellektual'nyy analiz i prioritizatsiya assotsiativnykh pravil dlya bol'shikh dannykh: mnogokriterial'nyy podkhod k analizu resheniy [Mining and prioritization of association rules for big data: multi-criteria decision analysis approach]. J Big Data 4, 42 (2017).

3. Pevchenko, S. S. Metody intellektual'nogo analiza dannykh [Data Mining Methods]. M Molodoy uchenyy [Young scientist], 2015, No 13 (93), pp 167-169.

4. Sovetskiy entsiklopedicheskiy slovar', Moskva, izdatel'stvo «Sovetskaya entsiklo-pediya» [Soviet Encyclopedic Dictionary, Moscow, publishing house "Soviet Encyclopedia"], 1981

5. Rahmatyllin, R. Yu. Istina kak filosofskaya kategoriya [Truth as a philosophical category]. Molodoy uchenyy [Young scientist], 2014, No 13 (72), pp 332-335.

Информация об авторах

Вадим Сергеевич Лебедев – аспирант, кафедра «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: lebedevvs97@yandex.ru

Information about the authors

Vadim Sergeevich Lebedev – post-graduate student, Department "Information systems and information security", Irkutsk State Transport University, Irkutsk, e-mail: lebedevvs97@yandex.ru

Для цитирования

Лебедев В.С. Индуктивный вывод в анализе больших данных // Информационные технологии и математическое моделирование в управлении сложными системами: электрон. науч. журн. – 2022. – №4(16). – С. 10-15 – DOI: 10.26731/2658-3704.2022.4(16).16-21 – Режим доступа: http://ismm-irgups.ru/toma/416-2022, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 17.12.2022).

For citations

Lebedev V.S Inductive inference in big data analysis // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2022. No. 4(16). P. 16-21. DOI: 10.26731/2658-3704.2022.4(16).16-21 [Accessed 17/12/22].