DOI 10.37882/2223–2966.2023.02–2.07

АНАЛИЗА БОЛЬШИХ ДАННЫХ ДЛЯ ЗАДАЧ УПРАВЛЕНИЯ В ДИСТАНЦИОННЫХ СИСТЕМАХ ВЫСШЕГО ОБРАЗОВАНИЯ

BIG DATA ANALYSIS METHODOLOGY FOR MANAGEMENT TASKS IN REMOTE HIGHER EDUCATION SYSTEMS

A. Belyakova
O. Romashkova

Summary. The article is devoted to the study of processing and analysis of big data for control tasks in remote sources of higher education. The methodology of data processing and analysis is described, which includes removing the curse of data dimensionality, setting an analysis goal, applying a machine learning algorithm and evaluating efficiency.

Keywords: educational organization, data processing, data analysis, big data.

Белякова Анна Вячеславовна
Аспирант, ГАОУ ВО «Московский городской педагогический университет (МГПУ)» г. Москва
itwhitelight@mail.ru
Ромашкова Оксана Николаевна
Доктор технических наук, профессор, ФГБОУ ВО «Российская академия народного хозяйства и государственной службы при Президенте РФ (РАНХиГС)»
г. Москва
ox-rom@yandex.ru

Аннотация. Статья посвящена исследованию процессов обработки и анализа больших данных для задач управления в дистанционных системах высшего образования. Описана методика обработки и анализа данных, включающая в себя снятие проклятия размерности данных, постановки цели анализа, применения алгоритма машинного обучения и оценку эффективности.

Ключевые слова: образовательная организация, обработка данных, анализ данных, большие данные.

В дистанционных системах высшего образования существует множество данных, и за счёт своих особенностей они также являются и большими данными, то есть обладают такими свойствами, как:

- многообразие (не все данные хранятся структурированно в базах данных, а какие-то данные разбиты по нескольким базам данных);
- рост объёма (количество данных увеличивается практически экспоненциально, чем более совершенна дистанционная система высшего образования, тем больше данных в неё попадает);
- количество данных (данные буквально большие: их объёмы переходят от измерений в терабайтах до измерений в петабайтах) .

Поэтому большие данные и сложно обработать. Кроме того, есть и проблема с разными типами данных: что-то хранится только в таблицах Excel, а что-то – в БД (базах данных), всё время появляются новые данные, а руководству для решения задач управления просмотреть за один раз такие объёмы информации не представляется возможным [1].

Описанная далее методика процесса обработки и анализа больших данных для задач управления позволяет справиться с большими данными и использовать результаты их обработки для принятия управленческих решений вследствие того, что лица, принимающие решения, будут обладать возможностью оценить результаты деятельности за счёт всей полноты информации, извлечённой из больших данных.

И подобные данные будут необходимы на всех уровнях принятия решений в организации высшего образования.

Для начального уровня (преподаватели) для проведения занятий и воспитательной деятельности (когда педагог выступает в качестве наставника) очевидно необходимо понимать, каков контингент обучающихся, в чём заключаются их интересы, это достигается за счёт рейтинговой оценки обучающихся и выявления закономерностей в больших данных — ведь на их основании можно сделать выводы об интересах обучающихся и приготовить гибкую программу занятий, которая, оставаясь в рамках учебного плана, поможет максимальному числу обучающихся успешно завершить курс обучения по программе, либо перевестись на обучение по другой, более подходящей именно для них специальности. За счёт обнаружения аномалий в данных преподаватели могут выявить обучающихся, показывающих невероятно высокие показатели и помочь им

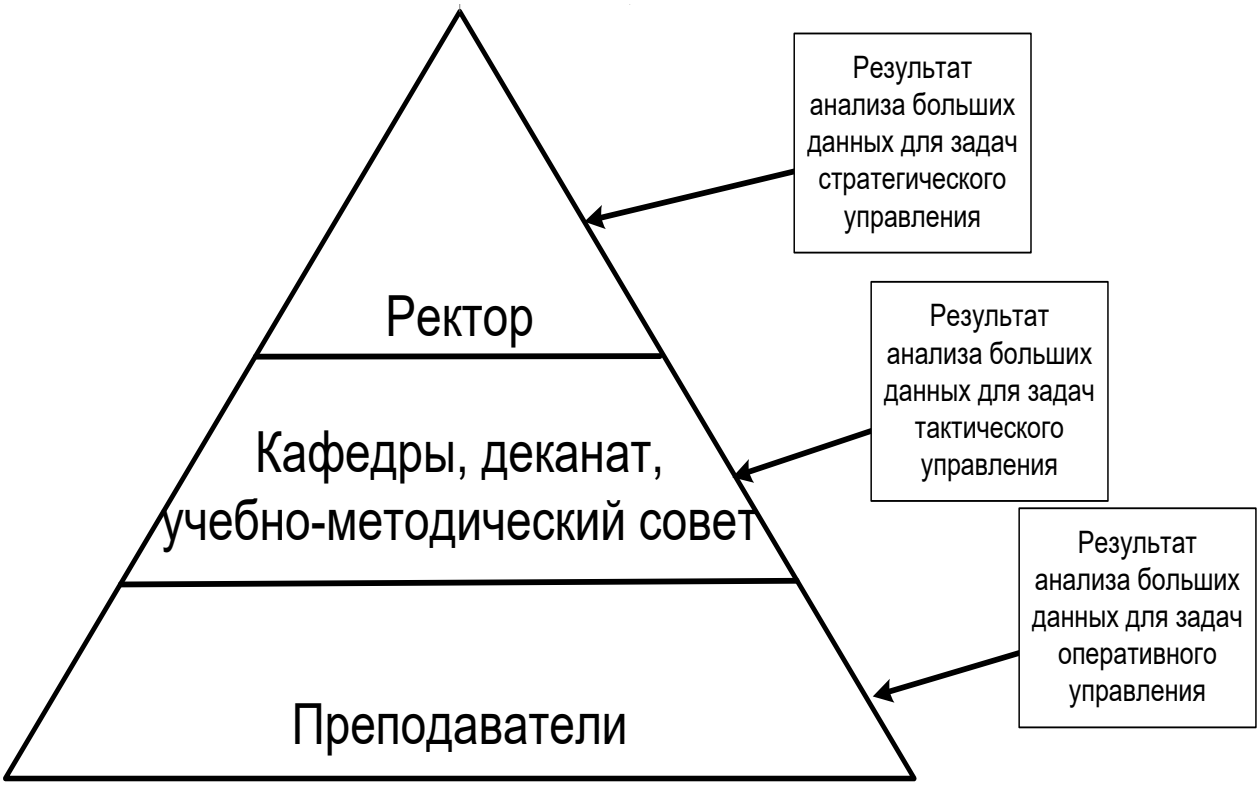

Рис. 1. Модель управления в дистанционной системе высшего образования

в самореализации, упрощая педагогам индивидуальный подход к каждому, когда созданы благоприятные условия как для быстро и очень успешно осваивающих программу обучающихся, так и для отстающих.

На среднем уровне (кафедра, деканат) управление учебным процессом опять же упрощается, за счёт возможности нанять востребованных преподавателей, именно по тем областям, которые интересны обучающимся и за счёт этого повысить востребованность программ дистанционного обучения. Для этого можно изымать из больших данных сведения о количестве просмотров онлайн-лекций, степень вовлеченности обучающихся, опросы об их интересах, учёт достижений обучающихся, сведения об успеваемости и дополнительно осваиваемых материалов при помощи дистанционной системы высшего образования [2].

На высшем уровне (ректорат) для стратегического управления высшей образовательной организацией необходимы оценка эффективности деятельности на определённый момент для планирования стратегии развития; результаты анализа данных в виде рейтинговой оценки; получение достоверной картины о возможных затратах и прибыли — а это, в свою очередь, помогает регулировать политику ценообразования; о предполагаемом количестве обучающихся, которые не смогут получить высшее образование; предполагаемом количестве абитуриентов, поступающих в бакалавриат, магистратуру, аспирантуру и специалитет; необходим анализ смещения акцентов рынка образовательных услуг на основе уже имеющихся данных, вследствие чего ожидается снижение рисков за счёт более востребованных, возможно, междисциплинарных направлений, основанных на выявленных интересах обучающихся, как результат — повышение эффективности воспитательной работы и заинтересованности обучающихся как в образовательном процессе, так и в большей обратной связи, которая даст ещё больше данных, а они, в свою очередь, позволят принимать более эффективные решения, отвечая задачам современности [3, 4].

Востребованность результата анализа больших данных на разных уровнях управления дистанционной системой высшего образования для решения задач управления (рисунок 1).

Сбор и обработка больших данных для задач управления

Для сбора и последующей обработки больших данных обычно применяют либо озера данных, где данные обычно не структурированы, либо хранилище данных,

где данные структурированы определённым образом и вследствие этого их легче обработать. В случае с озером данных нет необходимости поддерживать определенную структуру данных, что является вроде бы преимуществом, над хранилищем данных, однако это усложняет последующую обработку данных.

Кроме того, при хранении и обработке больших данных всегда стоит проблема актуальности и достоверности этих данных, например: как долго есть необходимость хранить посещение лекцией обучающимися, отражают ли они картину действительности — достоверные ли они. А данные о том, кто и когда был онлайн: это актуальные ли данные, нужны ли они и насколько они достоверны, может быть, обучающийся просто зашел в дистанционную систему высшего образования, а затем отошел от устройства, через которое осуществлял соединение.

Какие-то данные могут являться так называемыми «скрытыми» или «темными» данными — то есть, когда мы для анализа больших данных некоторые данные не используем, не можем обработать или найти эти данные — то они становятся скрытыми. Если данные о посещении дополнительных, факультативных занятий по определенным дисциплинам не обрабатываются, не учитываться при анализе, то вследствие этого администрация, делая дополнительные занятия платными сталкивается с отсутствием спроса, так как оказывается, что даже бесплатные занятия по этим дисциплинам были не актуальны, однако если бы эти данные перестали быть скрытыми, темными, то администрация не приняла бы подобного решения. Хранение же скрытых данных без их обработки приводит лишь к издержкам.

Процесс сбора и обработки больших данных для анализа представлен в модели процессов (рисунок 2).

Таким образом процесс включает этапы сбора и извлечения данных с серверов дистанционной системы высшего образования для мгновенной обработки (обработка больших данных может производиться при помощи нейросети), при поступлении новых данных и загружаются на долгосрочное хранение (при этом есть возможность выполнить очистку или дополнение данных), затем данные загружаются в OLAP-куб для удобства аналитики, если не происходит какой-либо ошибки во время обработки данных, а в случае успешной обработки происходит получение результатов анализа данных.

Также ненужные на данный момент данные, которые не сохраняются в хранилище данных и на данный момент не будут обработаны, можно опционально за-

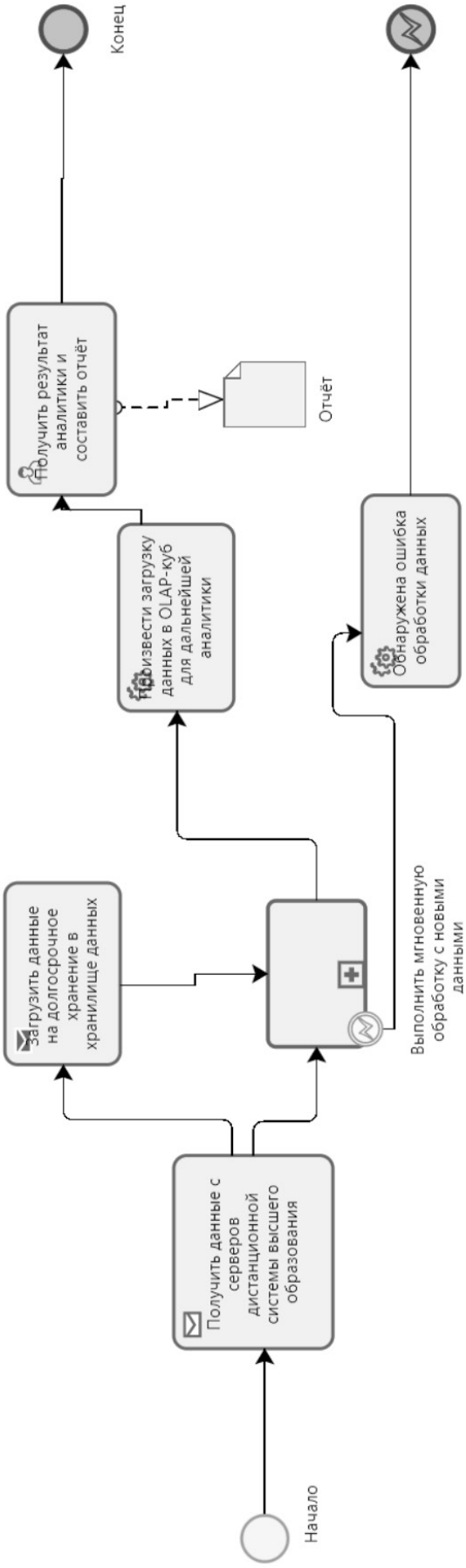

Рис. 2. Достижения в базе данных

Таблица 1. Сравнительный анализ сервисов

Название/Параметр	Snowflake	Azure	AWS
Конфиденциальность данных	Шифрует данные	Шифрование данных, цифровая подпись	Шифрование данных с собственным ключом
Ценовая политика	Тарифные планы, предзаказ, гибкая ценовая политика	Отдельно плата за пользование разными сервисами	Цена по запросу
Архитектура	Базы данных + виртуальная машина обработки запросов + набор сервисов в облаке	Масштабируемая архитектура SynapseSQL	Без общего доступа — независимые узлы в кластере
Обращение в службу поддержки	На сайте	В зависимости от оплаченного плана поддержки, в Twitter	В зависимости от оплаченного плана поддержки

гружать в озеро данных, из которого впоследствии, при извлечении этих данных, их придется структурировать, тогда как в хранилище данных хранятся уже структурированные данные, кроме того, необходимо будет рано или поздно вынимать оттуда данные для использования, и чем скорее, тем лучше, так как со временем некоторые данные теряют свою актуальность, без дополнительных усилий по обслуживанию озера данных теряют смысл, так что данную опцию можно реализовывать только при наличии возможностей по развертыванию, например, на «Snowflake» или «AWS». Данные сервисы могут предоставить также и хранилище данных (таблица 1) .

На последнем этапе жизненного цикла данных их необходимо удалять из архива, если есть уверенность в том, что они никогда не пригодятся, при этом надо убедиться, что это именно невостребованные данные, а не скрытые данные, которые можно было бы обработать, но за счёт их слабой структурированности их приняли за мусор. Обычно на данном этапе удаляют ненужные копии, поврежденные или очень старые данные.

Дополнительную сложность создаёт обеспечение безопасности больших данных, которое требует в первую очередь шифровать данные для сохранения конфиденциальности.

Анализ данных при помощи нейросети

Анализ данных может производиться различными методами, такими как дерево классификаций или регрессионный анализ. Суть методики процесса обработки больших данных для задач управления в дистанционных системах высшего образования как раз и состоит в том, чтобы применить каждый из этих методов на своём этапе обработки больших данных, так как на одном этапе определенный метод может быть не лучшим решением, но на другом этапе только он может помочь качественно обработать и проанализировать данные. Далее описана методика проведения анализа и обработки данных.

Итак, на первом этапе, который можно назвать «снятием проклятия размерности» [5], нам необходимо выделить те признаки, которые нам необходимы для анализа. Для этого нам необходимо провести понижение размерности, ведь из больших данных мы извлекаем огромное количество параметров об обучающихся, а избыточность бесполезных данных ничем нам не поможет при анализе, без них результат анализа будет получен быстрее и сам будет более точным. Таким образом, мы уже на первом этапе боремся с такой проблемой, как переобучение, когда результат излишне завязан на обучающий набор данных. Конечно, при последующем решении, к примеру, конкретно задачи кластеризации, могло бы, казалось, сработать обратное — повышение размерности выборки, однако в случае больших данных это не лучшая тактика, кроме того, тут существует опасность превышения размеров обучающей выборки, потому именно понижение размерности будет самым эффективным решением.

Понижение размерности может происходить либо за счёт объединения признаков, внутри пространства признаков O, либо за счёт сокращения количества признаков в пространстве O. Методика предполагает использование второго варианта, так как при использовании первого мы уже не будем понимать, какие конкретно признаки обучающихся мы анализируем, и при загрузке данных для дальнейшей аналитики в OLAP-куб могут возникнуть проблемы. В результате прохождения первого этапа методики мы должны получить из исходного пространства $O_1 - O_2$, где исходное пространство — O_1, с вектором

$$o = (f_1(o), \cdots, f_n(o),$$

а полученное пространство — O_2, с вектором [6]

$o = (f_1(o), \cdots, f_k(o)$, где $n > k$.

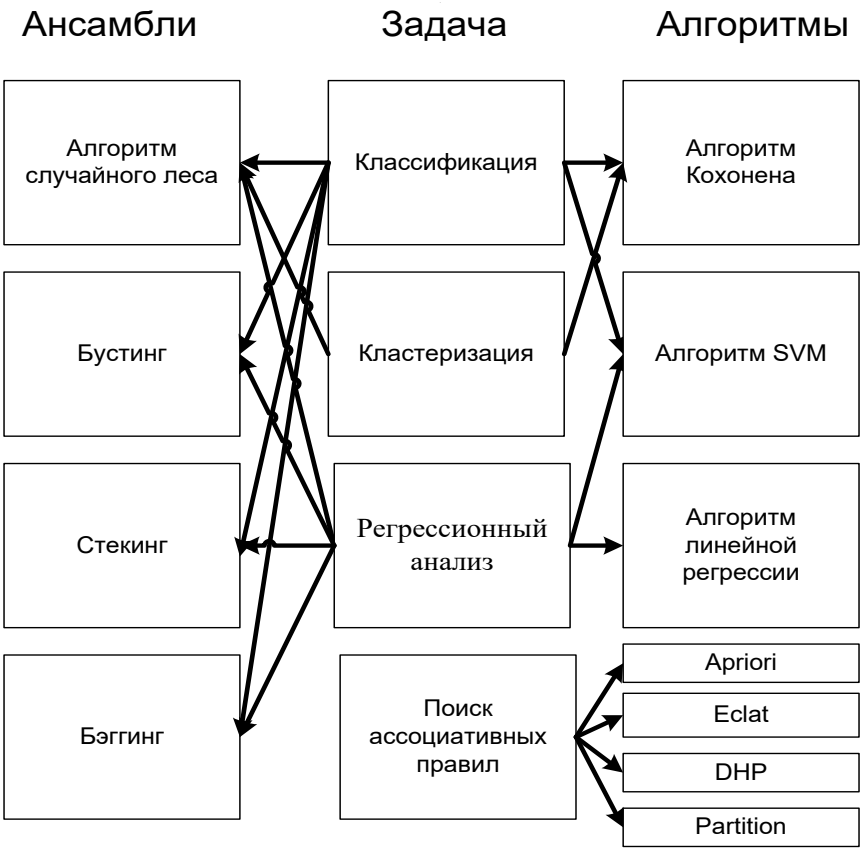

Рис. 3. Возможные алгоритмы для решения задач

Необходимо упомянуть о том, что первый этап в методике значительно упрощает оркестрацию [7] данных за счёт того, что образовательной организации не придётся хранить излишние аналитические данные и тратить вычислительные мощности на их обработку, ведь сами по себе большие данные в дистанционных системах высшего образования требуют значительных ресурсов по их хранению, приёму и передаче.

На втором этапе методики мы выбираем цель анализа данных: если необходимо выявить, есть ли тенденция увеличения с годами отчисления обучающихся платных отделений и почему — это будет одна задача анализа данных (дерево классификаций), если же нам необходимо предсказать затраты на покупку оборудования для реализации образовательного процесса в дистанционных системах высшего образования, то это будет другая задача анализа данных, а именно — регрессионный анализ, о котором упоминалось в начале данной главы: результат, выдаваемый нейросетью A при регрессионном анализе будет принадлежать множеству вещественных чисел $\mathbb{R}$.

Несмотря на то, что результатом регрессионного анализа будет конкретная цена, образовательной организации потребуется множество данных для обучающей выборки: данные о востребованности уже закупленного оборудования, данные о количестве и цене оборудования за прошлые годы, данные о цене покупки у поставщика на момент анализа и даже курс доллара к рублю.

На третьем этапе методики образовательная организация выбирает алгоритм или ансамбль алгоритмов машинного обучения для задачи в предыдущем этапе: к примеру, это может быть алгоритм случайного леса (как раз ансамбль, состоящий из деревьев решений), с которым на выходе получаем затраты на покупку оборудования образовательной организацией в будущем периоде для предотвращения издержек (рисунок 3).

В конце указанного этапа часть данных откладывается для тестирования качества модели, а остальные данные идут на обучение.

На четвёртом и последнем этапе реализации методики происходит получение результата и оценка качества полученного результата при помощи кросс-валидации, когда при помощи незадействованных при обучении данных происходит оценка эффективности.

Заключение

Таким образом, в рамках исследований проведен анализ процессов обработки и анализа больших данных для задач управления в дистанционных системах высшего образования. Описана методика обработки и анализа данных, включающая в себя снятие проклятия размерности данных, постановки цели анализа, применения алгоритма машинного обучения и оценку эффективности.

ЛИТЕРАТУРА

1. Белякова А.В., Пономарева Л.А., Чискидов С.В. Прототип информационной системы оценки качества учебного процесса в образовательной организации // В книге: Новые информационные технологии в научных исследованиях. Материалы XXIV Всероссийской научно-технической конференции студентов, молодых ученых и специалистов.— 2019.— С. 45–46.
2. Ромашкова О.Н., Пономарева Л.А., Василюк И.П. Применение инфокоммуникационных технологий для анализа показателей рейтинговой оценки вуза // В книге: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. 2018. С. 65–68.
3. Белякова А.В., Пономарева Л.А., Чискидов С.В., Василюк И.П. Программа для автоматизированного управления рейтинговыми показателями вузов // Свидетельство о регистрации программы для ЭВМ RU 2019611874, 05.02.2019. Заявка № 2018664941 от 24.12.2018.
4. Gaidamaka, Y.V., Romashkova, O. N., Ponomareva, L.A., Vasilyuk, I.P. Application of information technology for the analysis of the rating of university // В сборнике: CEUR Workshop Proceedings 8. Сер. "ITTMM 2018 — Proceedings of the Selected Papers of the 8th International Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems"" 2018. С. 46–53.
5. Белякова А.В. Новый подход к анализу достижений обучающихся образовательных организаций // В сборнике: Математика и информатика в образовании и бизнесе. Сборник материалов международной научно-практической конференции.— 2020.— С. 68–73.
6. Белякова А.В. Моделирование схем процессов обработки данных для задач управления в системах высшего образования // В сборнике: Открытая наука 2021. Сборник материалов научной конференции с международным участием.— Москва, 2021.— С. 89–91.
7. Белякова А.В., Пономарева Л.А., Ромашкова О.Н. Математическая модель оценки качества образовательного процесса // В книге: Новые информационные технологии в научных исследованиях. Материалы XXIV Всероссийской научно-технической конференции студентов, молодых ученых и специалистов.— 2019.— С. 43–44.

Журнал «Современная наука: актуальные проблемы теории и практики»

Российская Академия Народного Хозяйства и Государственной Службы при Президенте Российской Федерации