Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

Александров М.А., Стефановский Д.В.

Формирование представительной выборки при анализе больших данных

Москва 2017

Аннотация. Препринт содержит результаты исследований сотрудников Международной лаборатории математических методов исследования социальных сетей в 2016 г по одному из направлений работы лаборатории, связанному с обработкой больших данных

Александров М.А. доцент кафедры системного анализа и информатики экономического факультета Российской академии народного хозяйства и государственной службы при Президенте РФ

Стефановский Д.В. Старший научный сотрудник Международная лаборатория математических методов исследования социальных сетей ИПЭИ Российской академии народного хозяйства и государственной службы при Президенте РФ

Данная работа подготовлена на основе материалов научно-исследовательской работы, выполненной в соответствии с Государственным заданием РАНХиГС при Президенте Российской Федерации на 2016 год.

СОДЕРЖАНИЕ

Глава 1. Проблема больших данных

1.1 Новая парадигма

1. Особенности обработки больших данных

Тематика исследования предполагает: с одной стороны доступ к многочисленным сайтам, локализованным и в различных регионах и в различных организациях региона, а с другой использование полученной информации экспертами в реальном масштабе времени. В некоторых задачах вычислительная среда пользователя позволяет проводить такой сбор и обработку. Здесь мы можем говорить о коллекциях (наборах) данных, даже о больших коллекциях (больших наборах) данных. Но в других задачах это становится крайне трудным – данные либо одновременно не доступны, либо доступны, но время обработки не позволяет реализовать режим реального времени, либо и то и другое вместе. Технологии обработки данных, связанные именно с такими обстоятельствами, получили название технологий Больших Данных (Big Data).

Данная проблема стала предметом рассмотрения Национального Исследовательского Совета США (NRC) с конца 90-х годов прошлого столетия [1]. Тогда большие коллекции данных называли массивными (огромными) данными. Но широкое внедрение паутины WEB создало новую ситуацию, о которой сказано выше. Теперь понадобилось развитие специальных технологий, изначально ориентированных на сверхбольшие массивы данных. Эти технологии, чтобы отличить их от просто больших данных Large Data стали называть Big Data. В Национальном Совете был создан специальный Комитет по этой проблематике, который подготовил несколько лет назад обширный доклад [2]. Сравнивая эти две публикации, можно проследить эволюцию и постановок задач и предлагаемых решений. Отметим, что ключевая идея всех подходов к решению проблемы Big Data – это идея масштабируемости. Идеи масштабируемости и их реализации рассмотрены в разных статьях. Упомянем здесь работу [3].

Наблюдения, экспериментальные исследования и численное моделирование во многих областях науки и бизнеса в настоящее время требуют генерации терабайт данных. В некоторых случаях эти объемы затрагивают уже первые петабайты и идут

далее. Анализ информации, содержащейся в этих наборах данных, уже привели к крупным прорывам в областях, как геномика, астрономия дальних миров, физика высоких энергий и обработка материалов социальных сетей.

Традиционные методы анализа были основаны главным образом на предположении, что аналитики могут работать с данными в рамках своей собственной вычислительной среды, но Big Data меняет эту парадигму, особенно в тех случаях, в которых имеется огромное количество распределенных источников данных.

В то же время бизнес сообщество уже предприняло «меры защиты», чтобы противостоять вызовам Big Data. Лидером здесь стала электронная коммерция. Например, Google, Yahoo !, Microsoft и другие интернет-компании работают с базами, измеряемыми в эксабайтах (10^{18} байт). Социальные медиа, как, например, Facebook, YouTube, Twitter имеют сотни миллионов пользователей. Их информация это те самые эксабайты, а извлекаемые знания превосходят самые смелые ожидания недавнего прошлого. Интеллектуальный анализ данных из этих огромных хранилищ отражает то, что думают пользователи об управлении кризисами, маркетинге, развлечениях. Они помогают решать проблемы кибербезопасности и Национальной разведки. Коллекции документов, изображений, видео сейчас мыслится не просто как набор битовых строк, которые будут сохранены, проиндексированы, и когда нужно загружены в память компьютера. Это потенциальный источник знаний и открытий, требующих сложных методов анализа. Все это выходит далеко за рамки классической индексации через ключевые слова. Здесь уже нужны реляционные и семантические интерпретации явлений, лежащих в основе данных.

Ряд проблем, как в области управления данными, так и в области анализа данных требуют новых подходов для поддержки Big Data. Эти проблемы охватывают формирование данных, подготовку данных к анализу, разработку стратегии совместного использования.

Отметим особенности:

- Работа с существенно распределенными источниками данных,
- Работа с данными, прошедшими несколько стадий предобработки, так что их источник неизвестен,
- Контроль полноты данных,

- Учет смещения выборки относительно всей совокупности и учет неоднородности,
- Работа с различными форматами данных и структур,
- Разработка алгоритмов, использующих параллельные и распределенные архитектуры,
- Обеспечение целостности данных,
- Обеспечение безопасности данных,
- Обнаружение данных и обеспечение их интеграции с ранее накопленными,
- Обеспечение совместного использования данных,
- Разработка методов визуализации больших массивов данных,
- Разработка масштабируемых алгоритмов
- Борьба с временными ограничениями при анализе данных и принятии решений в реальном времени

Чем лучше удастся использовать Big Data, тем больше будет возможностей науке расширить свое влияние, и технологиям стать более адаптивными, персонализированными и надежными. Если обратиться к здравоохранению, то здесь можно представить себе, что нам даст хранение обширной информации о каждом человеке. Имеется в виду – его геном, его место в социальной среде, данные об окружающей среде. Эти данные могут быть объединены с данными других людей, и с результатами фундаментальных биологических и медицинских исследований. Таким образом, могут быть оптимизированы методы лечения для каждого человека. Можно также представить себе многочисленные возможности для бизнеса, которые появляются при наличии данных о запросах людей на товары и услуги и детальными описаниями товаров и услуг. Все это позволит создавать новые рынки.

Оптимизм по поводу перспектив применения Big Data имеет законные основания. Несколько десятилетий исследований и разработок в области баз данных и поисковых систем дали богатый опыт в области создания масштабируемых систем передачи и обработки данных. В частности, эти направления вызвали появление облачных вычислений и других параллельных и распределенных платформ, которые, как считают специалисты, хорошо подходят для анализа Big Data. Кроме того, инновации в области машинного обучения, интеллектуального анализа данных, статистики и теории алгоритмов, привели к созданию методов, которые могут быть применены ко все более крупным наборам данных.

Тем не менее, такой оптимизм должен быть сдержанным и основываться на понимании основных трудностей, возникающих при попытке достижения поставленных задач. Отчасти, эти трудности знакомы тем, кто имел дело с реализацией крупномасштабных баз данных, создавал инструменты для смягчения узких мест в обработке данных, пытался достигнуть простоты и универсальности интерфейса программирования, проектировал системы, устойчивые к аппаратном сбоям, а также эксплуатировал параллельные и распределенные системы на аппаратном уровне. Но проблемы Big Data, как мы указывали выше, выходят за пределы хранения, индексации и выполнения запросов, что составляло классический перечень задач в системе управления базами данных. Здесь логическая цепочка от данных к знаниям, как правило, прослеживалась весьма слабо.

2. Трудности, связанные с логическим выводом

Умозаключение (логический вывод) это процесс превращения данных в знания, где знание часто выражается в терминах сущностей, которые не присутствуют в данных сами по себе, но которые присутствуют в моделях, которые можно использовать для интерпретации данных. Статистическая строгость необходима, чтобы оправдать выявленный переход от данных к знаниям. И здесь возникают многие трудности при попытке привнести статистические принципы нести на большие массивы данных. Здесь мы можем получить результаты, которые в лучшем случае не будут полезны, а в худшем случае и вовсе вредны. При всяком рассмотрении логического вывода на Big Data необходимо иметь в виду, здесь можно получить что-то, напоминающее знания, но что знанием не является. Проблема в том, что часто довольно трудно понять, что это произошло.

В самом деле, многие вопросы сводятся к оценке качества логического вывода. Главным здесь является феномен ‘смещения выборки’. А именно, данные могут быть собраны в соответствии с определенным критерием отбора, но выводы и решения на этих данных могут относиться к другому критерию формирования выборки. Эта проблема, кажется, будет особенно острой в технологиях Big Data, которые часто состоят из множества поднаборов данных, каждый из которых был сформирован в соответствии со своим конкретным критерием отбора при слабом контроле над общей композицией.

Другой важной проблемой является, т.н. ‘провенанс’ (происхождение). Многие системы включают слои логического вывода, где данные не являются оригинальными, а являются продуктами с процедур логического вывода каких-то других слоев. В большой системе, содержащей соединенные между собой умозаключения, это может приводить цикличности. Это может приводить к отклонениям от цепочки логического вывода и может усиливать шум.

И, наконец, главный вопрос контроля частоты ошибок при множестве гипотез. Действительно, массовые наборы данных обычно включают в себя рост не только числа объектов (строки базы данных), но и рост числа дескрипторов этих данных (столбцы базы данных). При формировании прогностической модели на этих дескрипторах приходится рассматривать многочисленные комбинации дескрипторов, обладающие наилучшей предсказательной способностью. При росте числа дескрипторов экспоненциально растет число гипотез относительно свойств этих дескрипторов. И это приводит к серьезным последствиями относительно частоты появления ошибок. То есть, наивный призыв к ‘закону больших чисел’ для Big Data вряд ли может быть оправдано. Во всяком случае, опасности, связанные со статистическими колебаниями, на самом деле могут возрастать по мере того, как набор данных увеличивается в размерах.

Статистика имеет инструменты, которые в принципе могут решать такие вопросы, но здесь надо принять во внимание: (1) Все статистические инструменты основаны на предположениях о характеристиках данных, которые установлены и принимаются во внимание вместе со знаниям, как эти данные отбирали. Однако, эти предположения могут быть нарушены в процессе сборки Big Data; (2) Все инструменты для оценки ошибок вычислений, включая задачи диагностики, сами по себе являются вычислительными процедурами. И эти процедуры могут быть неосуществимыми из больших объемов данных.

Несмотря на все опасности, указанные выше, специалисты Комитета по анализу больших данных полагают, логические выводы могут быть сделаны корректно и, тем самым оказаться весьма полезными. Для этого должны быть предприняты усилия по разработке масштабируемых вычислительных инфраструктур. Исследования должны учитывать циклы принятия решений в режиме реального времени и обеспечивать компромисс между скоростью и точностью. Необходимы новые инструменты, которые позволили бы включить экспертов в анализ данных на всех этапах обработки. При таком включении следует

учесть, что знания часто субъективны и зависят от контекста, и что некоторые аспекты человеческого интеллекта в ближайшем будущем не будут заменены функциями машин.

3. Возможные направления исследования

Спектр исследований, связанный с технологиями Big Data, можно представить так:

- Представление данных, в том числе атрибутация данных, а также преобразования, которые пытаются уменьшить репрезентативную сложность данных. Было бы полезно выявить наиболее часто используемые преобразования;
- Вопросы вычислительной сложности алгоритмов и, как понимание таких вопросов позволяет оценить вычислительные ресурсов. Обеспечение компромиссов между ресурсами;
- Статистическая модель потенциальных возможностей в коррекции (чистке) данных и проверки их достоверности;
- Отбор проб как в качестве составной части процесса сбора данных, так и в качестве ключевой методологии для сокращения объема данных;
- Методы, использующие возможности людей, включенных в цикл обработки данных. В частности, это может быть краудсорсинг, где люди используются в качестве источника обучающих данных для алгоритмов обучения и визуализации. Такое участие помогает людям не только понять выводы из анализа данных, но и пересматривать принятые модели.

1.2 Подходы к решению проблем Big Data

1. Междисциплинарность

Научные исследования и разработки, необходимые для анализа Big Data, выходят далеко за пределы границ какой-то одной дисциплины, и все специалисты сходятся в необходимости радикальной междисциплинарности в подходах к решению проблем Big Data. Компьютерные специалисты, участвующие в создании систем обработки Big Data, должны развивать более глубокое понимание умозаключений (логических

выводов), в то время как специалисты по статистике должны заботиться о масштабируемости при решении алгоритмических задач, а также о режиме реального времени при принятии решений. Математики также должны играть важную роль, имея дело прикладной линейной алгеброй и теорией оптимизации. Последняя уже вносит свой вклад в анализ крупномасштабных данных. Эта роль математиков, скорее всего, продолжит расти. Как уже было сказано, участие человека в процессе обработки данных имеет важное значение, и здесь требуется участие социологов, психологов, а также специалистов в области визуализации. И, наконец, огромный рост проектных решений требует участия как ученых, так и пользователей из данной предметной области на этапе создания конкретных технологий Big Data.

2. Опыт последних лет

Этот опыт может быть представлен следующими положениями:

В последние годы наблюдается быстрый рост параллельных и распределенных вычислительных систем, разработанных в значительной степени, чтобы служить в качестве основы для современных интернет-информационных экосистем. Эти системы подпитывают информационный поиск, электронную коммерцию, социальные сети и онлайн-развлечения, и они предоставляют платформу, на которой вопросы анализа Big Data вышли на первый план. Частью этой проблемы является проблема масштабирования этих систем и алгоритмов для обработки все больших коллекций данных. Важно, однако, признать, что цели анализа Big Data выходят за рамки вычислений и представлений данных, которые были содержанием классических поисковых систем и систем обработки баз данных и использовались там для решения задач статистического вывода. Сейчас же цель состоит в том, чтобы превратить данные в знания и эффективно поддержать принятие решений.

Выявление знаний в процессе логического вывода требует контроля над ошибками. Поэтому в технологиях Big Data должны быть предусмотрены статистически обоснованные процедуры, которые обеспечивают контроль над ошибками при обработке больших массивов данных. При этом надо учитывать, что эти процедуры сами являются вычислительными процедурами, которые потребляют ресурсы.

Есть много источников потенциальной ошибки при анализе Big Data. Многие из них возникают из-за интереса к ‘длинным хвостам’, которые часто сопровождают сбор больших массивов данных. События в длинных хвостах могут возникать быть исчезающе редко, даже в объемах Big Data. Например, в применении информационных технологий, целью которых является предоставление детализированной персонализированной услуги, для многих пользователей может оказаться мало информации, даже в очень больших наборах данных. В науке цель состоит в том, чтобы найти необычные или редкие явления, хотя доказательства для таких явлений могут оказаться слабыми. Последнее вполне может случиться, если принять во внимание увеличение частоты появления ошибок, связанных с поиском при большом числе гипотез.

Другие источники ошибок, которые распространены в Big Data, включают в себя неоднородность в данных, и пристрастия, возникающих в результате неконтролируемых образцов выборки. Все это может быть следствием неизвестного происхождения товаров в базе данных.

Анализ данных при классической обработке всегда основан на установленных предположениях о характере данных. В Big Data предположения могут не совпадать, и часто не совпадают с предположениями. Поэтому, здесь часто приходится задаваться априорными предположениями о данных и проверять их одновременно с обработкой.

Анализ Big Data не является достоянием какой-либо одной предметной области, но есть достаточно междисциплинарная технология предприятие. Решения проблем Big Data требуют смещения идей из области информатики и прикладной статистики, в области прикладной и чистой математики, теории оптимизации, а также различных областей техники. Отметим здесь, в частности, теорию обработки сигналов и теорию информации. На протяжении всего процесса проектирования систем для анализа Big Data необходимо участие специалистов заданной предметной области, как ученых, так пользователей.

Есть также много вопросов, примыкающих к проблемам Big Data. В первую очередь, это вопросы неприкосновенности личных данных, решение которых требует участия ученых-юристов, экономистов и социологов.

В целом, с привлечением специалистов разных областей можно обсуждать компромиссные решения с учетом ограничений, связанных с вычислениями, статистикой, участием человека в процессе принятия решений.

Надо обратить внимание на подход, основанный на так называемых ‘наиболее вероятных решениях’. Речь идет о частных решениях, ориентированных на наиболее часто встречающиеся случаи, как контр подход к генерализации решения. Дело в том, что попытка решить проблему в более общей постановке, чем требуется, может дать решение, которое нереализуемо. И здесь междисциплинарность может направить усилия в нужном направлении. Например, наилучшее алгоритмическое решение, может дать с точки зрения статистики слабообоснованное решение. Поэтому, можно рассмотреть случаи часто встречающихся характеристик данных, и на них ориентировать разрабатываемый алгоритм. Аналогичным образом, знание типичных запросов может позволить ограничить анализ на именно этом множестве, не затрагивая все множество возможных запросов. Также можно подойти и к проблеме параллельного программирования, рассмотрев упрощенные алгоритмы, ориентированные именно на эти типичные случаи.

3. Формы данных

Хотя существует много источников данных, которые в настоящее время способствуют быстрому росту объема данных, несколько форм данных создают особенно интересные проблемы.

- Во-первых, много текущих данных включает в себя человеческую речь, и все чаще цель обработки таких данных состоит в извлечении смысловых значений, лежащих в основе этих данных. Примеры включают в себя как выявление тональности текстов, так полномасштабный семантический анализ, необходимый для вопрос-ответных систем.
- Во-вторых, изображения и видео становятся все более распространенными формами данных, что ставит целый ряд проблем крупномасштабного сжатия изображений, распознавания образов и семантического анализа.
- В-третьих, данные все чаще метятся гео-пространственными и временными тегами, создавая сложности согласования через пространственно-временные масштабы.
- В-четвертых, многие наборы данных включают сети и графики, где пользователя интересуют такие семантически богатые понятия, как ‘центральность’ и ‘влияние’.

Имеются более сложные формы данных, которые сводятся к решению проблем искусственного интеллекта, где требуются решения, выходящие за рамки краткосрочного масштабирования существующих алгоритмов. Технологии Big Data могут дать новые подходы к решению этих проблем. В качестве примера можно привести машинный перевод, где нужная форма перевода текста может быть получена на основе анализа больших коллекций данных по близкой тематике.

4. Интерфейс с пользователем и потоковая обработка

Анализ Big Data создает новые проблемы в интерфейсе между людьми и компьютерами. Как только что упоминалось, многие наборы данных требуют такой обработки, которая находится за пределами досягаемости алгоритмических подходов, и для которой необходимо участие человека.

Это участие может быть определено аналитиком на основе имеющихся данных. Участие человека поможет сократить набор возможных гипотез и обеспечить компромисс при решении вопроса о продолжении или приостановке расчетов. Участие человека может быть реализовано в форме краудсорсинга. Это потенциально мощный источник ресурсов, которым надо пользоваться с осторожностью, принимая во внимание многие виды ошибок и предубеждений, которые могут возникнуть. В любом случае есть много проблем, с которыми придется столкнуться при разработке эффективных визуализаций и интерфейсов. В более общем плане, говоря об интерфейсах, подразумевается увязка человеческих суждений с алгоритмами анализа данных.

Многие пользователи проводят анализ данных в режиме реального времени, получая потоки данных, которые могут перегрузить каналы их передачи. Кроме того, часто возникает желание быстро принимать решения в режиме реального времени. Эти временнЫе проблемы дают особенно наглядный пример необходимости компромисса между точностью (качеством решения) и скоростью (временем получения решения). Для получения такого компромисса нужен диалог, соответственно, между специалистами в области прикладной статистики, и специалистами в области техники алгоритмирования и программирования. Между тем, при разработке алгоритмов анализа данных:

- исследования прикладных статистиков по оценке ошибок в решениях редко принимаются во внимание из-за ограничений времени их принятия в режиме реального времени;
- исследования специалистами-программистами вычислительной сложности алгоритмов редко принимаются во внимание из-за необходимости выполнять ограничения, связанные со статистическим риском принятых решений.

5. Программное обеспечение

Существует большая потребность в развитии т.н. ‘промежуточного программного обеспечения’ - программных компонентов, которые связывают спецификации анализа данных высокого уровня и низкого уровня в системах распределенной архитектуры. Большую часть компонентов можно позаимствовать из инструментов, уже разработанных в различных научных коллективах, но основное внимание нужно уделить алгоритмическим решениям в рамках ограничений, связанных со статистическими оценками.

Существует также острая потребность создания программного обеспечения, ориентированного на конечных пользователей. Здесь цель состоит в том, чтобы относительно наивные пользователи могли бы осуществлять анализ Big Data без детального знания алгоритмов вычислений и статистических оценок результатов. Тем не менее, это не говорит о том, что конечной целью масштабируемого программного обеспечения является получение решений под ключ.

Эффективное (продуктивное) участие человека при анализе Big Data всегда будет необходимо при анализе данных, и это участие должно быть основано на понимании сложности компьютерных вычислений и понимании статистических оценок результатов. Развитие систем анализа Big Data должно осуществляться параллельно с усилиями по обучению студентов и пользователей ‘вычислительному’ и ‘статистическому’ мышлению.

1.3 Алгоритмические проблемы

Специалистами Комитета по анализу больших данных подготовлен перечень некоторых основных алгоритмических проблем, возникающих при обработке Big Data. Предлагаемый перечень мог бы помочь организовать исследовательский ландшафт, а также обеспечить отправную точку для разработки промежуточного программного обеспечения, о котором говорилось выше. Этот перечень определяет основные задачи, решение которых оказалось полезным при анализе данных. Комитет определил следующие семь основных классов задач:

1. Основные статистические данные,
2. Обобщенная задача N тел,
3. Теоретико-графовые вычисления,
4. Линейные алгебраические вычисления,
5. Оптимизация,
6. Интеграция,
7. Проблемы выравнивания.

Для каждого из этих классов алгоритмов существуют вычислительные проблемы, которые возникают в рамках какой-либо конкретной предметной области. Задание предметной области формирует специализированную стратегию, ориентированную на эту предметную область. Большая часть алгоритмов в прошлом рассматривала вычислительную среду как один процессор, при этом предполагалось, что все данные располагаются в оперативной памяти (RAM). Сейчас вычислительная среда характеризуется следующими особенностями:

- Имеется установка для потоковой передачи, при которой данные прибывают в быстрой последовательности, и только подмножество этих данных может быть сохранено;
- Имеется установка на основе дисковой памяти, когда данные слишком велики, чтобы хранить их при расчетах в оперативной памяти, но эти данные помещается на диск одной машины;
- Имеется распределенная установка, в которой данные распределены между несколькими машинами либо, используя их RAM, либо используя их диски;

– Многопоточная установка, в которой данные находятся на одной машине, имеющей несколько процессоров, которые совместно используют оперативную память этой машины.

Подготовка студентов к анализу Big Data потребует знакомства с вычислительной инфраструктурой, которая позволяет решать реальные проблемы, связанные с большими наборами данных. Наличие данных, программного обеспечения и вычислительных ресурсов может помочь в подготоке следующего поколения специалистов, имеющих навыки работы с Big Data. В процессе работы с такими данными будут появляться существенно новые идеи их обработки, которые инициируют и новые исследования.

Следует сказать, что анализ Big Data это не одна проблема или одна методология. Данные часто неоднородны, и лучшая атака на проблемы может включать в себя выявление частных подзадач, для каждой из которых может быть найден лучший вычислительный алгоритм или наиболее точная интерпретация результатов. Обнаружение таких подзадач само по себе может быть считать проблемой. С другой стороны данные часто обеспечивают лишь частичный вид на проблему, и тогда решение может потребовать объединения (слияния) нескольких источников данных. Эти перспективы сегментации или слияния не будет вступать в конфликт, если учесть особенности предметной области, которые позволят раскрыть соответствующую комбинацию данных.

Можно было бы надеяться, что опыт работы с Big Data позволит выявить некоторые общие, стандартные процедуры, которые будут полезными по умолчанию для любого набора Big Data. Ну, как например, быстрое преобразование Фурье FFT оказывается общеупотребительным инструментом в классической обработке сигналов. Тем не менее, Комитет настроен пессимисточно относительно идеи, что такие процедуры существуют в целом. Это не противоречит тому, что некоторые полезные общие процедуры обработки не будут появляться. На самом деле предложенный перечень алгоритмических проблем нацелен на то, что для каждой из проблем будут предложены общие подходы к их решению и далее будут разработаны общеупотребительные алгоритмы. Но важно подчеркнуть необходимость гибкости в реализации этих алгоритмов для обеспечения широты их применения. И здесь же подчеркнем, что наличие общеупотребительных процедур не означает их реализации ‘под ключ’ для наивных пользователей.

Специалисты Комитета по анализу больших данных полагают, что построение системы для обработки Big Data потребует инженерного мастерства и обоснованных научных суждений. Развертывание такой системы потребует моделирования решений, умения обращаться с приближенными решениями, внимания к диагностике и надежности.

Комитет ожидает появления новых программных и аппаратных платформ, ориентированных на анализ Big Data. Для управления такими платформами в контексте решения реальные проблем необходимо появление нового класса инженеров. Это потребует введения соответствующих предметов в программы университетов по подготовке бакалавров и магистров по компьютерным специальностям

В заключение представим таблицу научных и технических приложений, на которые оказали воздействие технологии Big Data.

Таблица 1. Предметные области и Big Data

Области в 1995 г	**Области в 2014 г**	**Частные случаи**
Физ. Науки	Физ. науки	Астрономия, физика частиц
Климатология	Климатология	
Обр-ка сигналов	Обр-ка сигналов	
Медицина	Медицина	Снимки, истории болезней
Искуств. интеллект	Искуств. интеллект	Естеств. яз., комп. видение
Marketing	Marketing	Интернет реклама, корп. лояльность
N/A	Политология	Модел-ие изменений режима
N/A	Юриспруденция	Обнар-ие злоупотр., наркозав-ть
N/A	Изучение культуры	География культуры, землячество
N/A	Социология	Сравн. соц-ия, соц. сети
N/A	Биология	Геном, протеомы, экология
N/A	Нейронауки	fMRI, многоэлектрод. Запись
N/A	Психология	Соц. психология

1.4 Выводы по главе 1

1. Ключевые решения

Анализ литературы по Big Data показывает следующие ключевые решения по 3-м аспектам:

Архитектура

- Локальная обработка с параллельными и конвейерными вычислениями
- Распределенная обработка на многомашинных комплексах

Модели

- Применение непрерывных моделей на основе физических представлений, вместо множества дискретных
- Применение моделей статистической физики для описания характеристик данных
- Применение принципов естественного отбора при пошаговом поиске оптимальных решений

Обработка

- Редукция объема данных - выбор репрезентативного подмножества
- Редукция размерности данных - переход к сокращенному описанию
- Масштабируемость обработки - регулирование компромисса между скоростью и точностью

Подходы, связанные с выбором моделей и способом обработки, не связаны необходимостью перехода к параллельным или распределенным вычислениям. Они могут быть реализованы на машинах традиционной архитектуры

2. Примеры ключевых решений

Приведем примеры реализаций предложенных решений

1) Локальная обработка с параллельными и конвейерными вычислениями
   - Параллельные вычисления выполняются на машинах SIMD архитектуры,
   - Конвейерные вычисления выполняются на машинах MISD архитектуры
2) Распределенная обработка на многомашинных комплексах
   - Такая обработка условно соответствует архитектуре MIMD

3) Применение непрерывных моделей на основе физических представлений, вместо множества дискретных

- Модель распределения тепла по графу связей слов текста
- Гравитационная модель для отражения взаимного притяжения слов текста

4) Применение моделей статистической физики для описания характеристик данных

- Случайное распределение слов как распределение частиц по энергиям

5) Применение принципов естественного отбора при пошаговом поиске оптимальных решений

- Генетические алгоритмы при решении оптимизационных задач
- Алгоритм MIA в методе МГУА

6) Редукция объема данных - выбор репрезентативного подмножества1

- Проверка стат. гипотез о законах распределения выборки и подвыборки

7) Редукция размерности данных - переход к сокращенному описанию

- Метод Главных Компонент, кластер-анализ

8) Масштабируемость обработки - регулирование компромисса между скоростью и точностью

- Обработка данных на подпространствах
- Применение стохастического квазиградиента

Литература к главе 1

1. National Research Council of USA: Massive Data Sets. In: Proceedings of a Workshop. National Academy Press, 1996
2. National Research Council of USA: Frontiers in Massive Data Analysis, Report of the Committee on the Analysis of Massive Data, National Academy Press, DOI 10.17226/18374, 2013
3. Halevy, A., Norvig P., Pereira F.: The unreasonable effectiveness of data. IEEE Intelligent Systems, N2, 1996, pp.8-12

Глава 2. Отбор репрезентативной выборки из больших данных

2.1 Репрезентативность

1. Состояние вопроса

Отбор представительного подмножества объектов является одним из эффективных путей в обработке больших наборов данных. Это касается как автоматических время-затратных алгоритмов, так и ручной обработки материала экспертами. Репрезентативность или представительность мы рассматриваем здесь в узком смысле как равенство статистических распределений параметров объектов для подмножества и для полного набора объектов. Мы предлагаем простой метод для выбора такого подмножества на основе тестирования сложных статистических гипотез, включая искусственную гипотезу, чтобы избежать неоднозначности. Функциональность метода демонстрируется на двух наборах данных, где один относится к деятельность компаний по мобильной связи в России, а другой к междугородним автобусным перевозкам в Крыму.

Тема отбора представительного подмножества объектов из большого набора данных становится очень популярной в эру так называемых больших данных. Можно было найти много публикаций по этой теме, связанных с определенными проблемами машинного обучения и определенными предметными областями. Эту тему в литературе часто называют 'sampling'. Известный отчет [6] представляет варианты сэмплинга. Следующие два основных типа случайного сэмплинга являются предметно-независимыми: выборка, основанная на примерах – будем называть ее индивидуальным сэмплингом, и выборка, основанная на группах объектов - будем называть ее групповым сэмплингом.

Типичная задача индивидуального сэмплинга состоит в поиске минимального количества объектов для оценки количества объектов, подобных данному, в полном наборе объектов. Алгоритмы, основанной на индивидуальном сэмплинге, состоят в вычислении близости между объектами и данным образцом, оценке распределения образцов в полной коллекции данных, и применении формул статистики, учитывающих количество объектов в подмножестве и в полерм множествеобъектов. Обзор методов индивидуального сэмплинга представлен в [2].

Групповой сэмплинг является самым популярным типом выборки. Типичная задача группового сэмплинга больших данных состоит в поиске минимального количества объектов, структура которых близка (подобна) структуре полного набора данных. Один из первых алгоритмов группового сэмплинга описан в [3]. Здесь число центров групп зафиксировано, и только эти центры рассматривают как репрезентативные объекты. Алгоритм поиска определяет местоположение этих центров, принимая во внимание плотность распределения объектов в данном пространстве параметров. Тот же самый подход использовался в нашей работе [5]. Здесь весь набор объекта был сгруппирован с методом MajorClust [7]. В отличие от предыдущего случая, этот метод определяет количество групп автоматически, и это обстоятельство делает этот метод одним из самых эффективных для кластер-анализа. Каждый кластер представлен 3 объектами (семантические описатели), отражая отношение к другим объектам внутри этого кластера и вне его. Эти описатели, взятые от всех групп, рассматриваются как репрезентативное подмножество. Наконец, мы рассмотрим недавно изданную статью [9]. Здесь, кластеры формируются методом k-средних на n объектах. Затем находят m объектов около каждого центра. После этого алгоритма случайно выбирает d x k других объектов. Предполагается, что значения k, m и d назначены пользователем, где $m<<n$ и $d<<n$. Эти $(m+d)$ x k объектов есть так называемое T-представительное подмножество, если качество кластеризации превышает некоторый заданный порог T. В противном случае m и d увеличивается и расчеты повторяются.

Главный недостаток группового сэмплинга в отличие от индивидуального сэмплинга состоит в необходимости вычислять попарные расстояния, для чего требуются время и память при работе с большими данными. Действительно, для объектов $N=10^3$ объектов мы имеем $\sim 10^6/2$ вычислений, для $N=10^6$ объектов мы уже имеем $\sim 10^{12}/2$ вычислений и т.д.

Сэмплинг, представленного в этой работе, направлен ина выбор минимального количества объектов, распределения параметров которых должны быть близко (подобно) их распределениям во всем наборе данных. С такой выборкой было бы возможным решать проблемы и индивидуального и группового сэмплинга, то есть оценить полное число образцов во всем наборе и оценить структуру данных всего набора. Эта проблема сводится к хорошо-известной проверке гипотезы об однородности двух наборов данных [4]. Соответствующий алгоритм формирует растущее подмножество объектов, пока принятая гипотеза однородности не будет

принята для всех параметров объектов. Однако, такой простой алгоритм не учитывает требование независимости от данных и неоднозначность, связанную с принятием гипотезы. Предложенный нами метод пытается избежать этих недостатков.

2. Постановка проблемы

Согласно цели, сформулированной выше, мы должны сформировать представительное подмножество объектов при следующих условиях:

1) Подмножество должно отразить распределения параметров всего набора данных;
2) Подмножество не должно зависеть от местоположения объектов в наборе данных;
3) Подмножество должно уменьшить неоднозначность, связанную с принятием статистической гипотезы;
4) У подмножества должен быть минимальный объем.

Решение проблемы состоит в тестировании нескольких статистических гипотез, касающихся требований (1), (2) и (3). Пошаговая процедура с растущим числом объектов позволяет удовлетворять требование (4). Предложенный метод проверен на двух наборах данных, связанных с компаниями мобильной связи в России и к междугородними пассажирскими автобусными перевозками в Перу.

2.2 Проверка сложных гипотез

1. Два одномерных множества

Процедура проверки подобия двух наборов случайных величин известна в статистике, и мы только напомним некоторые положения [4]. Пусть имеем два набора объектов, каждый представлен одним параметром: $X=\{x_1,x_2,..x_{i..}\}$, $i=1,\ldots n_x$, и $Y=\{y_1,y_2,..y_{j...}\}$, $j=1,\ldots n_y$.. Чтобы сравнить их распределения, мы проверяем следующую нулевую гипотезу: ‘Распределения значений равны’. Тогда мы должны рассмотреть считать следующие два типичных и в некотором смысле противоположных случая [4]:

Значения принадлежат нормальному закону с неизвестным средним и дисперсией

В этом случае мы формируем так называемую z-статистику и сравниваем ее с критическим значением z_{crit}. Здесь:

$$z = |m_x - m_y| / \sqrt{(s^2_x/n_x + s^2_y/n_y)}$$

где: m_x и m_y - оценки средних значений, s^2_x, и s^2_y оценки дисперсии, n_x и n_y объемы множеств. Мы предполагаем, что наборы данных являются столь большими, что оценки практически равны реальным средним значениям и дисперсии. Переменная $z_{crit} = z_{crit}(\alpha)$ является табулированной функцией для различных уровней α. Последняя определяет доступную статистическую ошибку 1-го рода. Правило звучит так:

- если $|z| < z_{crit}(\alpha)$, то гипотеза принимается;
- иначе гипотеза отвергается

Типичные значения α 0.1, 0.05, 0,01

Значения принадлежат произвольному закону

В этом случае нужно использовать любой непараметрический тест. В нашей работе мы применяем тест Колмогорова-Смирнова. Чтобы использовать его, вычисляется следующая статистика:

$$\lambda = \sqrt{[(n_x \times n_y)/(n_x+n_y)]} \times D,\ D = max\ |F_{emp}(x_i) - F_{emp}(y_j)|$$

где: $F_{emp}(x_i)$ эмпирическая функция распределения для 1-го набора, и $F_{emp}(y_j)$ эмпирическая функция распределения для 2-го набора. Функция λ табулирована, так что имея вычисленное $\lambda_{crit}(\alpha)$ мы можем принять или отвергнуть гипотезу.

2. Два многомерных множества

Пусть у нас есть два набора объектов, представленных *n* параметрами каждый. Мы можем рассмотреть два режима проверки: твердый режим и мягкий режим.

Твердый режим

Твердый режим рассматривает два многомерных набора как подобные, если все n параметров 1-го набора имеют те же самые распределения, какое имеют соответствующие параметры 2-го набора. Подобие каждого параметра может быть проверено, используя правила, представленные в р. 2.2-1. Поэтому нулевая гипотеза отклоняется, если подобие отсутствует хотя бы для одного параметра. Легко подсчитать, что вероятность ошибки I рода $\mu = 1-(1-\alpha)^n$. Когда $\alpha \sim 0$, то $\mu \sim n\alpha$.

Мягкий режим

Мягкий режим способ рассматривает два многомерных набора как подобные, если, по крайней мере, каких-либо $k < n$ параметров 1-го набора имеют те же самые распределения, как и соответствующие параметры 2-го= набора. Поэтому, нулевая гипотеза отклоняется, если в наборах имеется менее, чем k подобных параметров. Формула для вероятности ошибки 1 рода подобна предыдущей: $\mu = 1-(1-\alpha)^k$.

2.3 Нулевая гипотеза и противоположная гипотеза

Классическая статистика говорит, что принятие любой нулевой гипотезы означает только отсутствие значительного противоречия между данной гипотезой и существующими данными. Но это не означает, что гипотеза правильна [4].

В нашей практике такая ситуация произошла впервые почти 15 лет назад, когда объекты были классифицированы при помощи статистической меры близости между объектами и данными классами [1]. Эта мера состояла в проверке нулевой гипотезы, и при этом оказывалось, что объекты часто одновременно принадлежали нескольким классам.

В нашем случае эксперименты показывают следующее: при ограниченном объеме данных любая гипотеза, даже самая неожиданная, относительно распределения данных даже может быть принята. Но с увеличением объема выборок несоответствие между данными и гипотезой становятся все более очевидными, и, в конце концов, гипотеза может быть отклонена. Поэтому в начальной стадии выбора с очень небольшим количеством объектов есть высокая вероятность принять любую гипотезу, которая приводит к статистической ошибке II рода. Чтобы уменьшить этот эффект, мы предлагаем использовать противоположную искусственную гипотезу: в

том случае, когда эта гипотеза принимается, то принятие основной нулевой гипотезы должно быть отклонено.

Предложенная искусственная гипотеза касается равномерного распределения рассматриваемых параметров. Такое решение может быть поддержано предположением, что такое распределение редко используется для практически полезных приложений, и обычно исключается из рассмотрения в алгоритмах, связанных с отбором объектов. Необходимо подчеркнуть, что равномерное распределение может быть не лучшим для нашего случая. Но, по крайней мере, это служит фильтром для ошибок II рода для нулевой гипотезы. Чтобы проверить гипотезу о равномерном распределении, мы используем тест Колмогорова-Смирнова.

2.4 Метод и программа

1. Шаги алгоритма

Мы строим алгоритм согласно постановке проблемы, описанной в р.2.1. Алгоритм включает два этапа: стадия 1 (предварительная обработка) и стадия 2 (обработка). Предварительно мы фиксируем общее количество объектов N, начальное количество репрезентативных объектов M, шаг для увеличения числа репрезентативных объектов m, количеств параметров, которые будут проверяться k. Мы фиксируем также вероятность ошибки I рода для основной гипотезы α и вероятность ошибки I рода для искусственной гипотезы β. В каждом шаге мы используем 3 набора объектов: увеличивающееся репрезентативное подмножество, дополнительное случайное подмножество того же самого объема, и оригинальный набор объектов.

Стадия 1, предварительная обработка

1. Вычисление статистики для всего набора данных, такой как среднее, дисперсия или эмпирическая функция распределения для формул от п. 2.2
2. Набор данных перемешивается случайным образом, чтобы избежать любого влияния начального распределения объектов

Стадия 2, обработка

3. Выбор начального репрезентативного подмножества объектов M. Это могут быть первые объекты от всего набора объектов.
4. Тестирование сложной гипотезы подобия между репрезентативным подмножеством и всем набором объектов (гипотеза 1). Здесь используются формулы п. 2.2. Нулевая гипотеза принимается в твердом режиме.
5. Тестирование гипотезы подобия между репрезентативным подмножеством и дополнительным случайным подмножеством (гипотеза 2). Мы используем здесь формулы от п. 2.2. Объем обоих подмножеств должен быть равным. Нулевая гипотеза принимается в твердом режиме.
6. Тестирование искусственной гипотезы. Для этого проверяется соответствие между репрезентативным подмножеством и равномерным распределением. Мы используем здесь формулы от п. 2.2. Объем подмножества и теоретической выборки должен быть равным. Нулевая гипотеза принимается в мягком режиме.
7. Если гипотезы о подобии между репрезентативным подмножеством, дополнительным подмножеством и целым набором данных приняты, и искусственная гипотеза отклонена, то процесс останавливается: текущее репрезентативное подмножество можно рассмотреть как окончательное решение. Иначе к текущему репрезентативному подмножеству добавляется новая часть объектов M=M+m и алгоритм возвращают к шагу (2)

Рисунок 1 иллюстрирует процесс решения. Здесь: N общее число объектов, $M_{1\text{-}r}$, $M_{2\text{-}r}$ и $M_{3\text{-}r}$ количество объектов в *репрезентативном* (r) подмножестве объектов; $M_{1\text{-}a}$, $M_{2\text{-}a}$ и $M_{3\text{-}a}$ количество объектов в *дополнительном* (a) подмножестве объектов; $M_{1\text{-}u}$, $M_{2\text{-}u}$ и $M_{3\text{-}u}$ количество объектов в теоретической выборке с *равномерным* (u) распределением; индексы 1,2,3 означают номер итерации.

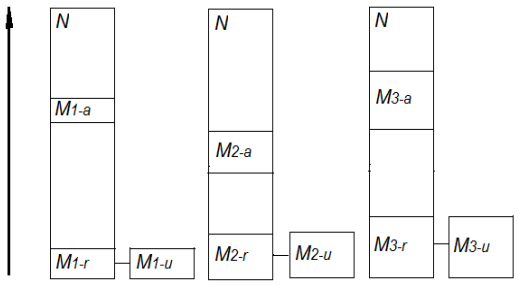

Рис 1. Схема алгоритма с увеличивающимся объемом репрезентативного подмножества

2. Функции программы

Предложенный метод реализован в программе, ориентированной на конечных пользователей. Интерфейс программы представлен на Рисунке 2 (это – стартовая версия). Интерфейс отражает функции программы. Программа записывает результаты в файл с фиксированным именем.

Чтобы увеличить надежность результатов, критерий несмещенности проверяется с тремя дополнительными случайными подмножествами вместо одного (см. шаг 3 на стадии обработки). Но такой способ может быть отключен.

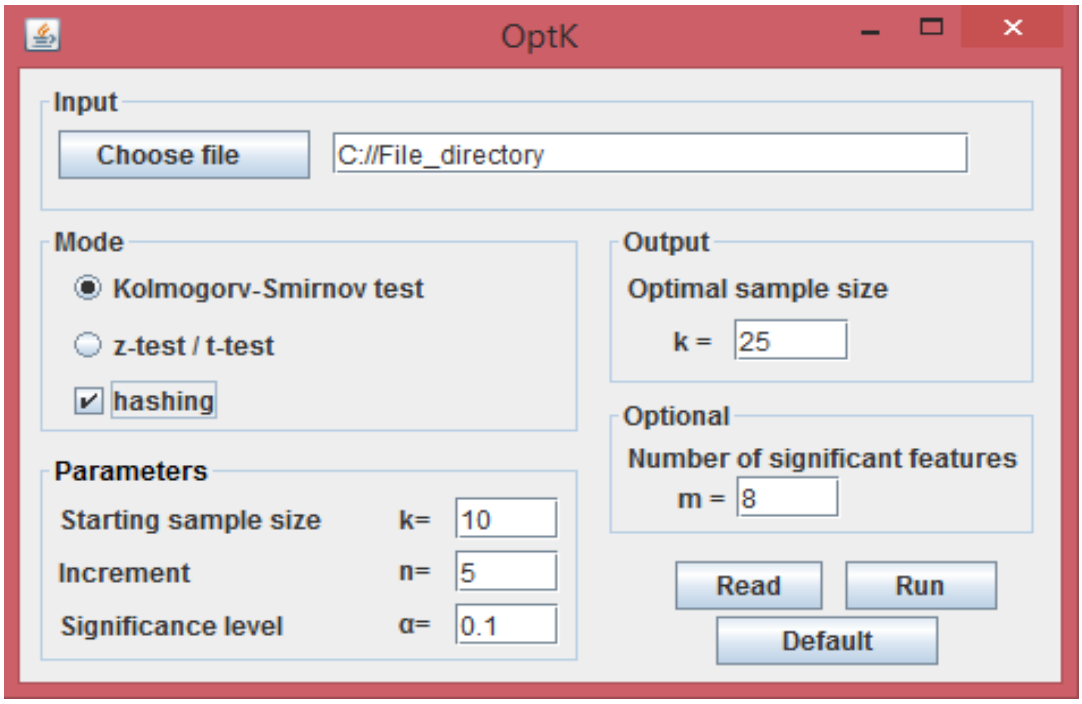

Рис.2. Интерфейс программы для отбора репрезентативного подмножества объектов

3. Модификация программы

Сейчас мы разрабатываем исследовательскую версию программы, чтобы изучить различные способы выбора объектов при изменении количества параметров, изменении уровня ошибок, и т.д.

2.5 Эксперименты

1. Оценка результатов

Мы используем подход, предложенный в [9]. Этот подход может быть изложен в следующей упрощенной форме. Пусть D_n и D_ν весь набор и выборка из этого набора, содержащая n и ν объектов соответственно, здесь $\nu < n$. Пусть P, $Q(P, D_\nu)$ и T есть проблема, качество решения проблемы с D_ν и порог качества соответственно.

Определение 1. Значение ν есть T-критический размер сэмплингаб если (а) для любого $k \geq \nu$ качество $Q(P, D_k) \geq T$, (б) имеется $k < \nu$ такое, что $Q(P, D_k) < T$.

Определение 2. Значение $Q_s = \nu/n$ есть качество сэмплинга D_n, где ν есть критический размер сэмплинга D_n

Поэтому, чтобы практически определить T-критический размер сэмплинга для данной проблемы P, нужно выполнить следующее: зафиксировать T и повторить несколько экспериментов с различными подмножествами D_k для различных k. Тогда нужно выбрать случаи $Q \geq T$ и взять максимальное значение k в этих случаях. Мы использовали именно этот прием в наших экспериментах.

2. Отбор компаний мобильной связи (Россия)

Российский рынок компаний мобильной связи включает приблизительно 600 компаний (данные 2015 г.). Проблема состоит в выборе репрезентативного подмножества из списка компаний, чтобы приблизительно оценить его привлекательность для инвестиций. Такую проблему рассматривали студенты бакалавриата РАНХиГС.

Для анализа были взяты следующие 4 параметра деятельности компании: доходность производства (*PP*), доходность активов (*PA*), доходность собственного

капитала (*PE*), коэффициент финансовой стабильности (*FS*). Нормализованные данные первых 3 компаний представлены в Таблице 1.

Таблица 1. Характеристики компаний (первые 3 из 600).

PP	***RA***	***RE***	***FS***
-0,32	-0,45	0,23	0,68
0,27	0,13	-0,37	-0,01
-0,38	-0,52	0,98	-0,09
…..	…..	…..	…..

В эксперименте мы меняли: (а) вероятность ошибки I рода α для гипотезы 1 и для гипотезы 2 и (б) вероятность ошибки I рода β для искусственной гипотезы. Начальное подмножество включало 5 объектов. Шаг также был равным 5 объектам. Мы выполнили 10 экспериментов для каждой комбинации (α,β) и зафиксировали максимальное количество отобранных объектов для этих комбинаций. Таким образом, мы представили так называемые (α,β) - критические размеры выборки с точки зрения р.2.3-1. Результаты представлены в Таблице 2. Знак '-' означает, что искусственная гипотеза не использовалась. Самая значимая комбинация здесь это (0,05, 0.01), для которой типовое качество решения равно Q_s=55/600~0.1.

Таблица 2. Критические размеры вывборки (компании)

$\alpha \backslash \beta$	-	0.01	0.05	0.10	0.50
0.05	5	***55***	55	15	10
0.10	5	65	65	15	10
0.50	260	260	260	260	260

Мы оценили качество выборки на основе результатов кластеризации. Для этого мы рассмотрели упомянутую (0.05,0.01) - комбинацию объектов и выполнили следующие процедуры:

1) Объединение в кластеры всего набора данных, эти кластеры рассматривали далее как классы в кластерной *F*-мере (см. ниже);
2) Группировка 55 случайных объектов, взятых из всего набора данных, эта процедура была повторена 10 раз;

3) Сравнение обоих результатов, используя кластерную *F*-меруб введеную в [8]. Эта мера отражает соответствие между классами и кластерами, полученными на шагах 1) и 2). Теоретически лучшее и худшее значения кластерной *F*-меры равны 1 и 0 соответственно.

В этом эксперименте мы использовали метод *k*-средних с *k*=3. Значение *k* было выбрано экспертами. Получающаяся *F*-мера принадлежала интервалу [0.45, 0.65].

3. Отбор режимов движения в межгородском автобусном сообщении (Перу)

Ежедневно в Перу выполняется приблизительно 6000 междугородних поездок на автобусе. Маршруты поездок соединяют приблизительно 100 мест назначения (данные 2015 г.). Все движения автобусов зарегистрированы через систему GPS в специальных базах данных. Частота регистрации составляет 1 минуту. Мы вводим понятие ‘объект движения’. Это - дистанция, которая преодолена транспортным средством за 10 минут. Поэтому 10 последовательных записей формируют один объект из движения. Мы собрали приблизительно 10000 записей, отражающих 24 поездки на автобусах в Перу, и расположили их последовательно - одну поездку на автобусе после другой. Эти записи можно рассмотреть как 10000/10=1000 объектов движения. Проблема состоит в выборе репрезентативного подмножества этого набора объектов для одной Перуанской комании, осуществляющей мониторинг межгородского автобусного сообщения. Работа проводилась совместно с кафедрой информатики католического Университета Сан-Пабло.

Для анализа мы использовали следующие 3 параметра для каждого объекта движения: средняя скорость (км/ч *AV*), изменение скорости (*VV*) и отклонение направления (градусы *DD*). Нормализованные данные первых 3 объектов на маршруте Лима-Арекипа представлены в Таблице 3.

Таблица 3. Характеристики объектов движения (первые 3 из 1000).

AV	***VV***	***DD***
-0.55	0.13	0.34
0.92	-0,30	-0.49
0.89	-0,95	2.42
…..	…..	…..

Мы повторили эксперимент относительно различных комбинаций (α,β) с объектами движения. Результаты представлены в Таблице 4. Самая значимая комбинация здесь также (0,05, 0.01), для которой типовое качество равняется Q_s=20/1000=0.02

Таблица 4. Критические размеры выборки (объекты движения)

$\alpha \backslash \beta$	-	0.01	0.05	0.10	0.50
0.05	5	***20***	15	15	10
0.10	5	25	15	15	10
0.50	445	445	345	345	345

Мы оценили качество выборки на основе результатов кластеризации также, как это мы сделали для компаний мобильной связи. В этой процедуре мы использовали 20 и 1000 объектов. Количество кластеров равно 4. Именно этому количеству кластеров соответствовало наилучшее значение индекса Дуна [8]. Получающаяся *F*-мера принадлежала интервалу [0.85, 1.00].

2.6 Выводы по главе 2

В нашем исследовании мы предлагаем метод отбора репрезентативного подмножества объектов из большого набора объектов, который обеспечивает:

- прогноз распределений параметров объектов всего множества на основе их распределений в подмножестве;
- независимость выбора объектов от положения этих объектов во всем множестве;
- повышенная точность как следствие разрешения неоднозначности, связанной с принятием нулевой гипотезы.

Эти преимущества достигнуты на основе применения сложных статистических гипотез. Метод имеет линейную сложность и может использоваться с любыми заранее неизвестными распределениями. Метод был проверен на двух реальных наборах данных и продемонстрировал обещанные результаты.

В будущем мы намереваемся рассмотреть следующие приложения метода: выбор представительных подмножеств университетов и колледжей в заданных странах или группах стран, выбор представительных подмножеств студентов в данных университетах и колледжах, другие приложения в области наук и образования.

Литература к главе 2

1) Alexandrov, M., Gelbukh, A., Lozovoi, G.: Chi-square Classifier for Document Categorization. In: Comp. Ling. and Intell. Text Processing. Springer, LNCS, vol. 204, pp. 457-459, (2001)
2) Au, T., Chin, M-L.i., Ma, G.: Mining rare events data by sampling and boosting: a case study. In: Proc. of Intern. Conf. on Inform. Systems, Technology and Management (ICISTM-2010), CCIS, vol.54, pp. 373-379, (2010)
3) Chaudhur, D., Murthy C., Chaudhur B.: Finding a subset of representative points in a data set. In: IEEE Trans. on Systems, Man, and Cybernetics, vol. 24(9), pp. 1416–1424, (1994)
4) Cramer, H.: Mathem. methods of statistics. Priceton Landark in Mathematics, (2016)
5) Gelbukh, A.,Alexandrov, M., Bourek, A., Makagonov,P.: Selection of representative documents for clusters in a document collection. In: Natural Lang. Proc. and Inform. Systems. GI-Edition, LNI, Germany, vol. 29, pp. 120-126, (2003)
6) National Research Council: Frontiers in Massive Data Analysis, Report of NRC, National Academies Press,USA, (2013)
7) Stein, B. , Niggemann, O.: On the Nature of Structure and its Identification. In: Graph-Theoretic Concepts in Computer Science, Springer, LNCS, vol. 1665, pp. 122-134, (1999)
8) Stein, B., Meyer zu Eissen, S., Wilssbrock, F.: On Cluster Validity and the Information Need of Users. In: Proc. of 3rd IASTED Intern. Conf. on AI and Applications (AIA-2003), Acta Press, pp. 216-221, (2003)
9) Sung, A., Ribeiro, B., Liu, Q.: Sampling and evaluating the Big Data for Knowledge Discovery. In: Proc. of Intern. Conf. on Internet of Things and Big Data (IoTBD 2016), Science and Techn. Publ., pp. 378-382, (2016)