Сокращение числа виртуальных экспериментов с помощью оценки корреляций параметров взаимодействующих гипотез

© Е. А. Тарасов

Московский государственный университет им. М.В. Ломоносова
Москва, Россия

Аннотация

В данной работе представлен подход, позволяющий исследователю сократить число виртуальных экспериментов, уменьшив количество наборов тестовых сценариев. Рассматриваемый подход основывается на вычислении корреляций между параметрами различных гипотез. Для решения данной задачи был выполнен обзор и сравнительный анализ существующих систем: Hephaestus, FCCE, ϒ-DB, реализующих схожий функционал. Далее, был произведен обзор алгоритмов отбора признаков, позволяющих уменьшить исследуемое пространство параметров и выявить взаимосвязь между ними. Были сформулированы функциональные требования к проектируемой системе. Рассмотрена практическая задача, которая может быть решена в рамках реализации данной платформы и описан ее частный случай, а именно, оценка корреляции параметров гипотез в астрономической задаче, которая будет использоваться в качестве тестового задания на этапе отладки системы.

Работа поддержана РФФИ (гранты 14-07-00548, 16-07-01028).

1 Введение

В современном мире исследования всё более зависимы от данных, которые становятся ключевым источником для получения новых знаний в той или иной области человеческой деятельности [4]. Такой подход получил название исследований с интенсивным использованием данных (ИИИД) [10] и развивается в соответствии с 4-й парадигмой научного развития [8]. Одним из ключевых элементов ИИИД является явное использование *гипотез* в определении *виртуального эксперимента* [3]. Гипотезам соответствует некоторая формальная спецификация свойств исследуемого явления, которая чаще всего имеет математическое представление. Сформулированные гипотезы нуждаются в тщательной проверке.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

В процессе выполнения виртуального эксперимента происходит манипулирование параметрами гипотез, т.е. набором переменных, которые в некоторых случаях могут быть коррелированы между собой, а также с параметрами других гипотез [10].

Так как число потенциальных гипотез в виртуальном эксперименте может быть огромным, а их взаимодействие нетривиальным, то в результате образуется пространство с большим числом виртуальных экспериментов, часть из которых плохо описывает наблюдения и нуждается в отсеивании ещё до выполнения эксперимента. Как следствие, исследователю необходимо средство, позволяющее заранее выявить и отсеять виртуальные эксперименты с прогнозируемо плохим результатом. В тоже время наличие сложных зависимостей в данных затрудняет их понимание исследователем и не позволяет делать это вручную [3]. Машинные средства оценки корреляции позволяют автоматически разделить виртуальные эксперименты на группы с заранее прогнозируемо хорошим и плохим результатом эксперимента [4].

Разрабатываемая архитектура платформы рассматривается в рамках более широкого проекта лаборатории, следовательно, будет интегрирована в него отдельным модулем [10].

Статья организована следующим образом. В разделе 2 приводится обзор существующих систем, позволяющих решать схожую задачу поиска корреляций и причинно-следственных связей. В разделе 3 приводится обзор алгоритмов отбора признаков. В разделе 4 формулируются требования к проектируемой системе. В разделе 5 формулируется тестовая задача поиска корреляции между двумя гипотезами. В разделе 6 формулируются дальнейшие шаги по развитию данной работы.

2 Обзор платформ для поиска корреляций над большими массивами данных

2.1 Hephaestus

В основе системы Hephaestus [3] лежит работа с виртуальными экспериментами над данными. Система помогает исследователю искать

корреляционные зависимости между большим числом переменных, предоставляет возможность сформировать гипотезы по наиболее перспективным связям, а затем с помощью тщательного тестирования перейти к причинно-следственной зависимости. Центральным блоком системы является SQL-подобный декларативный язык для описания виртуального эксперимента, с помощью которого возможно проектировать дизайн эксперимента, специфицировать основные гипотезы и их параметры, тестировать их, исполнять выбранные эксперименты и публиковать исследования. Так как применение только статистических методов и машинного обучения может приводить к ошибочным результатам в поиске причинно-следственных связей, то системы ориентирована на симбиоз между человеком и машиной. Используя корреляционные связи, которые были проверены экспертами в предметной области, Hephaestus собирает вероятностно причинные графы для спецификации семантики предметной области. Причинный граф поддерживает большое количество связей, обнаруженных в процессе выполнения виртуального эксперимента, а также позволяет исследовать последовательность выявления вероятностных причин и аномалий.

Hephaestus – это мета-система для исполнения виртуального эксперимента над существующими базами данных, которые могут уже выполнять некоторую аналитику локально. Она ориентирована на работу с очищенными и размеченными данными. Система состоит из следующих модулей:

- Получения набора данных. Авторы стремятся создать некую поисковую систему, которая принимает на вход строку, описывающую параметры гипотезы, возвращает ранжированный список потенциальных причинно-следственных зависимостей.
- Тестирования гипотез. Является основным рабочим модулем системы. Движок составляет запрос для оценки каждого возможного взаимодействия, разбивает образцы на контрольные блоки и высчитывает заданную метрику точности для каждого из них. После расчета статистики для блока, движок объединяет результаты, получая взвешенную оценку гипотезы.
- Ранжирования результатов. Гипотезы объединяются и сортируются по некоторой вероятностной оценке.

Вероятностно причинный граф – направленный ацикличный граф, содержащий коллекции причинно-следственных связей. Этот символьный язык позволяет исследователю интегрировать новые полученные знания для предметной области со своими ранее доступными знаниями в определенных областях науки.

2.2 FCCE

Платформа FCCE [16] разрабатывается для поиска корреляций в разнородных наборах данных, охватывающих большие временные диапазоны. Она ориентирована на работу с минимальными задержками и доступом к не пред-обработанным исходным данным.

Данная система рассматривалась на примере 2-х задач из области безопасности: обнаружения доменных имен потенциальных сетей зараженных рабочих станций и пост-инцидентное расследование проникновения.

Для минимизации задержек обработки данных решено было отказаться от использования традиционных реляционных баз данных для хранения информации и перейти на NoSQL.

Ключевым компонентом модели данных является концепция признаков. Признаки определяют связь между парой ключ-значение, каждый элемент из которых может содержать несколько атрибутов. FCCE представляет упрощенную реляционную модель данных для пользователя, где каждая таблица хранит один тип признаков. Каждая строка идентифицируема ключом и может содержать несколько атрибутов.

FCCE обеспечивает API для хранения, получения и вычисления корреляции над признаками. Разработчики предлагают оригинальный подход интеграции модуля оценки критерия корреляции в движок исполнения запросов над хранилищем, позволяющий ускорить время ответа на запрос и снизить накладные расходы на вычисление и ввод-вывод. FCCE использует два отличительных механизма для поддержки эффективности операций нахождения корреляций между признаками: канал запросов и модификатор запросов. Канал – механизм, позволяющий передавать признаки, извлеченные из одного запроса в другой запрос в качестве входных данных, т.е. последовательно можно объединять несколько GET функций, тем самым создавай пересечения нескольких признаков. Модификатор – над GET запросом предполагает использование широкого набора опций для более тонкого контроля его поведения.

Архитектура платформы состоит из следующих модулей:

- Извлечение. Для каждого источника эксперты в области определяют метод извлечения признаков из сырых данных.
- Агрегация. Данные собираются из различных локальных экстракторов в так называемые коллекторы, которые выполняют функцию дедупликации, отказоустойчивости, балансировки нагрузки.
- Хранение. Централизованное хранилище, над которым выполняются запросы к признакам.
- Получение. Модуль обеспечивает интерфейс запросов над хранилищем признаков. Доступ к данным организован с помощью 3 компонент. Первый состоит из Сервиса регистрации, осуществляет поиск корневого коллектора, и Протокола запросов, посылает запросы к соответствующему хранилищу,

используя тип признаков и ключи в качестве предикатов запросов. Второй, используя специальный протокол, может подписаться на определенный экстрактор или коллектор, так что при появлении интересующей пары ключ-значение они сразу попадут в него. Третий реализует интерфейс поиска корреляции признаков и позволяет настроить различные функции корреляции для получения знаний из различных типов признаков.

Более подробно механизм поиска корреляции признаков авторы статьи собираются раскрыть в будущих работах. В перспективе они так же планируют перенести функционал поиска корреляций с уровня доступа к данным на уровень хранилища признаков для уменьшения задержки при обработке сложных запросов.

2.3 Ƴ-DB

Разработки системы Ƴ-DB [4, 5] ведутся с целью поддержки процесса проведения научных исследований, обеспечивая возможность управления гипотезами и их анализа. Предиктивная аналитика строится над вероятностной базой данных.

В работе делается упор на управление параметрами гипотез. Ключевые особенности такого подхода и их отличия от управления экспериментальными данными заключаются в следующем:

- Работа ведется не со всеми данными, полученными в результате эксперимента, а только лишь с некоторым отобранным подмножеством. Тем самым уменьшается объем, но увеличивается структурированность данных.
- Если при работе с обычными данными модель доступа к ним ориентирована на работу с измерениями (денормализованный вид), то модель хранения параметров гипотез определяется из её структуры, т.е. происходит нормализация по факторам неопределенности.
- К неопределенностям на уровне данных, источниками которой являются их неполнота и несогласованность, так же добавляется неопределенность, порожденная существованием множества конкурирующих гипотез.

В качестве примера авторами был разобран сценарий расчета физиологических гипотез, а именно тестирование трёх различных теоретических моделей насыщенности гемоглобина кислородом.

Первый этап работы с гипотезами – это их кодирование. Для вычисления предсказаний гипотезы используют асимметричные функции, которые выполняют оценку над входными переменными (параметры) для вычисления значений выходных переменных (предсказаний). Техника кодирования гипотез базируется на наличие структуры гипотезы в машиночитаемом формате W3C MathML. Платформа Ƴ-DB имеет XML адаптер для извлечения моделей зашифрованных в формате MathML и вывода причинных зависимостей.

Ключевым компонентом архитектуры платформы является канал синтеза, который представляет собой последовательный процесс обработки данных. На вход поступает структура гипотез и их данные. Из структуры извлекается функциональные зависимости. Данные помещаются в большую таблицу, содержащую все переменные как реляционные атрибуты в таблице. Затем включается компонент синтеза и трансформирует данные из большой таблицы в вероятностную базу данных, где каждая гипотеза декомпозируется в таблицы претендентов. Авторами был предложен алгоритм трансформации каждой гипотезы в вероятностную таблицу. Базовый принцип проектирования неопределенного моделирования состоит в том, чтобы определить только одну случайную переменную для каждого действительного фактора неопределенности (u-фактор). Модель гипотезы сама по себе это теоретический u-фактор, чья неопределенность исходит из множества моделей, ориентированных на объяснение того же явления. Множество испытаний каждой гипотезы нацеленных на один и тот же феномен порождает множество эмпирических u-факторов. Для поддержки тестирования гипотезы вероятностное распределение феномена должно учитывать оба вида u-фактора.

Предиктивная аналитика выполняется над вероятностной базой. Ƴ-DB не предлагает каких либо новых инструментов для тестирования гипотез. Насколько можно понять, это статические методы на основе Байеса.

Прототип системы разработан как Web-приложение, написанное на Java, с компонентами канала, реализованными на стороне сервера поверх MayBMS. Где MayBMS – это расширение PostgreSQL. Как отмечают авторы управление данными гипотез является перспективным новым полем исследовательской деятельности, позволяющим получить больше пользы из эксперементальных данных, открытых исследовательскими лабораториями. В планы дальнейшего развития входит улучшить: статистические способности и масштабируемость системы для тестирования выборок большого объема.

2.4 Отличия от существующих подходов

В рамка данного подхода исследователь работает с уже существующими гипотезами, моделирующими свойство какого-либо явления в природе, экономике, бизнесе и т.д. Гипотеза является ключевым элементом рассматриваемого метода. Все параметры гипотез являются ценными и несущими информацию и поэтому от них нельзя избавляться.

Из-за присутствия априорных знаний, накопленных в виде гипотез, выбор параметров не

является полностью черным ящиком – в отличие от методов машинного обучения. Исследователю также заранее известна некоторая часть взаимосвязей между гипотезами, т.е. какие из них зависимы друг от друга.

Для части гипотез могут быть доступны локальные наблюдения, соответствующие их параметрам и выступающих в качестве ограничений на совокупность этих гипотез. Кроме наблюдений может быть доступна некоторое теоретическое распределение параметров гипотез.

Гипотезы могут быть сформулированы как набор правил, система математических уравнений и пр. Разрабатываемая в рамках данного подхода система должна уметь работать с разнообразными входными данными. При добавлении новых значений параметров гипотез или включений новых гипотез в существующий набор система должна обновлять набор сокращенных экспериментов.

Предлагаемый метод работает с симуляциями и наблюдениями. Сопоставляя экспериментальные и фактические данные, мы восстанавливаем полную модель по частично доступной.

Рассматриваемый подход нацелен на исключение заведомо «плохих» экспериментов, т.е. тех, которые производят симуляции с большой ошибкой. Установление причинно-следственных связей не является целью данной работы. Возможность поддержки данного механизма планируется в дальнейшем.

Таким образом сокращение числа экспериментов достигается за счет:

- Поиска корреляций – это позволяет объединить признаки исследуемого явления в некоторые группы, оказывающие влияние на него в некоторой совокупности.
- Анализ этих признаков – необходимо подобрать набор значений параметров в рамках выделенной группы, которые с определенными показателями точности описывали бы фактические данные наблюдений.
- Ранжирование гипотез по степени точности виртуального эксперимента. Это поможет исследователю обратить внимание на наиболее вероятные гипотезы без необходимости полного перебора гипотез.

3 Методы отбора признаков

Для уменьшения пространства параметров гипотез и виртуальных экспериментов используются методы отбора признаков [13]. Они позволяют увеличить скорость обработки данных и получения результат, не снижая показатели точности [18], путем выделения только тех информативных признаков, которые требуются для выполнения виртуального эксперимента.

Выделение набора признаков позволяет упростить понимание модели исследователем и, следовательно, использовать их в качестве входных данных для широко известных алгоритмов машинного обучения [6]. Так же данные методы позволяют уменьшить шум в данных и выявить взаимодействие между параметрами.

Методы отбора признаков возможно классифицировать следующим образом: Фильтры, Обертки, Встроенные [1, 13].

Фильтры. Опираются на общие характеристики обучающих данных и осуществляют процесс выборки признаков в качестве шага предварительной обработки независимо от индукционного алгоритма. Обладают низкой стоимостью вычислений. Фильтры используются в кластеризации для построения начального приближения. Не предназначены для выявления сложных связей между признаками, т.к. обладают низкой чувствительностью.

К таким методам можно отнести: **CFS** [6] – где выбор признаков на основе корреляций. Является простым многофакторным фильтрующим алгоритмом, который раскладывает подмножество признаков согласно эвристической функции оценки, основанной на корреляции. **INTERACT** [20] – двух этапный алгоритм, основанный на симметричной неопределенности и согласованности. **ReliefF** [11] – который является расширением алгоритма Relief, и работает путем случайной выборки экземпляра из данных, а затем находит его ближайшего соседа из того же или противоположного класса. **mRMR** [14] – выбирает признаки, которые имеют самое высокое значение информативности с целевым классом и обладающие минимальной избыточностью. Его разновидностью является **M_d фильтр** [17] – который использует меру монотонной зависимости для оценки информативности.

Обертки. Включают оптимизацию предиктора как часть процесса выбора. Позволяют выявлять зависимости признаков. Качество выборки зависит от индукционного алгоритма. Основным недостатком является вычислительная нагрузка, которая исходит от вызова алгоритма индукции для оценки каждого подмножества интересующих параметров.

К ним можно отнести: **WrapperSubsetEval** [19] – вычисляет наборы признаков с использованием схемы обучения. Для оценки точности схемы обучения для набора признаков используется перекрестная проверка. В качестве схемы обучения могут использоваться SVM и C4.5.

Встроенные. Выполняют функции выборки в процессе обучения. Как правило специфичны для алгоритмов машинного обучения. Применимость метода всегда зависит от типа решаемой задачи. Позволяют выявлять зависимости признаков. Обладают хорошей скоростью работы.

К ним относятся: **SVM-RFE** [15] – метод осуществляет выбор признаков итеративным обучением SVM классификатора с текущим набором признаков и удаляет наименее важный признак, указанный SVM. Существуют две версии этого метода: с линейным и нелинейным ядром. **FS-P** [12]

– основанный на перцептроне. Идея метода заключается в обучении перцептрона в контексте контролируемого обучения. Веса взаимосвязей используются как индикатор того, какие признаки могут быть наиболее информативными.

Другие методы. Так же к методам, позволяющим снизить размерность данных и оценить зависимость параметров можно отнести следующие техники. Анализ главных компонент (**PCA**) [7], которая включает в себя преобразование ряда коррелируемых переменных в меньшее число не коррелируемых. Анализ независимых компонент (**ICA**) [9], позволяющий не только декоррелировать параметры, но также уменьшает статистические зависимости более высокого порядка. Канонический корреляционный анализ (**CCA**) [7], устанавливающий соотношения линейных связей между двумя многомерными переменными. Неотрицательная факторизация (**NMF**) [9], позволяющая накладывать дополнительные ограничения на главные компоненты.

Применимость того или иного метода в разрабатываемой системе будет оценена на этапе отладки в рамках решения тестового сценария, описанного ниже.

4 Формализация требований

Разрабатываемая платформа должна удовлетворять следующим требованиям:

- Система должна быть модульной и функционально расширяемой.
- Должна поддерживать связность с другими компонентами глобальной системы, в рамках которой она реализуется.
- В качестве модуля хранения данных должна использоваться платформа, ориентированная на работу с большими объемами, а также поддерживающая современные средства аналитики.
- Ключевым компонентом системы является модуль по поиску корреляций параметров гипотез. В качестве используемых методов предлагается использовать различные подходы: байесовский, частотный, методы машинного обучения и сравнить полученными ими результаты между собой.
- Системы должна иметь возможность работать с уже сформулированными гипотезами. Гипотезы должны хранится в базе данных на ряду с экспериментальными данными и результатами проведения виртуального эксперимента.
- В зависимости от практической задачи методы машинного обучения могут отличаться от описанных в предыдущем разделе.

5 Сценарий для тестирования

В качестве сценария для тестирования будет рассмотрена частная задача Безансонской Модели Галактики [2] о нахождении корреляции параметров между двумя независимыми гипотезами, а именно Star Formation Rate и Initial Mass Function. Данные гипотезы описывают процесс зарождения звезды. Известно, что параметр γ в SFR коррелирует с моделями IMF. Как отмечают авторы:

“Однако, мы хотим подчеркнуть, что параметр γ коррелирует со значениями других параметров, используемых в модели, и особенно склонами (α) в IMF и возраста диска.” [2]

В настоящее время авторы фиксируют параметры остальных гипотез и изучают влияние α и γ на поведение модели вручную. Соответственно, в данном случае наша система значительно бы облегчила работу исследователей.

Гипотеза SFR представляет общую массу звезд, зарожденных в определенной области Галактики за некоторый интервал времени. В версии БМГ от 2014 года [2] функция представляется авторами в следующем виде:

$$SFR(i) = \exp(\gamma \times x_c(i)) \times d$$

Где γ – исследуемый параметр, x_c – возраст в i-ом интервале, d – размер возрастного интервала.

Гипотеза IMF представляет функцию распределения массы определенной популяции звезд. В общем виде может быть представлена:

$$\phi(m) = m^{-\alpha}$$

Где m – масса, α – параметр, характеризующий склон функции. Так как функция представлена на трех интервалах, то, таким образом, относительная масса внутри каждого интервала может рассчитываться как:

$$K_i \int_{m_i}^{m_{i+1}} m\phi(m)dm = MInt_i$$

Где K_i – коэффициент непрерывности, i – рассматриваемый интервал.

В данном примере представлены две гипотезы имеющие математическое представление и предположение о взаимной зависимости их параметров α и γ.

На первом шаге оценивается корреляция параметров данных двух гипотез. Значения оценки, полученные различными методами поиска корреляций, должны иметь сопоставимые между собой значения.

На втором шаге выполняется анализ данных признаков. Выполняя виртуальные эксперименты над ними и сравнивая результаты с реальными наблюдениями, накапливается информация об исследуемой модели.

Полученные значения дают возможность ранжировать гипотезы и выбрать наиболее вероятные, тем самым упрощая процесс выбора значения параметров гипотез модели. Результатом данной процедуры является набор значений параметров α и γ, при этом ожидается что описанные

ранее авторами параметры [2] будут включены в этот набор.

Заключение

В данной работе описан подход, позволяющий уменьшить пространство возможных гипотез в виртуальном эксперименте. Его идея базируется на поиске и оценке корреляций между параметрами выбранных гипотез. Дан обзор существующих и проектируемых систем, которые в той или иной мере реализуют функционал поиска корреляций над большими массивами данных. Представлены некоторые алгоритмы оценки взаимосвязи параметров, с учетом специфики работы с астрономическими данными [9]. Сформулированы требования, предъявляемые к проектируемой архитектуре.

В качестве дальнейших шагов развития подхода планируется реализация системы с учетом описанных ранее требований. В качестве практической задачи будут исследованы корреляции всех параметров гипотез Безансонской Модели Галактики [2], взаимосвязь двух гипотез которой рассматривалась в рамках тестовой задачи.

Благодарность

Автор статьи выражает благодарность Д. Ю. Ковалеву за предоставленную идею.

Литература

[1] Veronica Bolon-Canedo, Noelia Sanchez-Marono and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. Knowledge and Information Systems, 34(3), p. 483-519, 2013

[2] Maria A. Czekaj, Annie C. Robin, Francesca Figueras and Xavier Luri. Galaxy evolution: A new version of the Besancon Galaxy Model constrained with Tycho data. Barcelona: Universitet de Barcelona, PhD Thesis, 2012

[3] Jennie Duggan and Michael Brodie. Hephaestus: Data Reuse for Accelerating Scientific Discovery. In Proceedings of 7th Biennial Conference on Innovative Data Systems Research (CIDR'15), Asilomar, California, USA, 2015

[4] Bernardo Goncalves and Fabio Porto. Managing large-scale scientific hypotheses as uncertain and probabilistic data with support for predictive analytics. IEEE Computing in Science and Engineering, 17(5), p. 35-43, 2015

[5] Bernardo Goncalves, Frederico C. Silva and Fabio Porto. ϒ-DB: A system for data-driven hypothesis management and analytics, 2014. http://arxiv.org/abs/1411.7419

[6] Mark A. Hall. Correlation-based Feature Selection for Machine Learning. The University of Waikato, Hamilton, New Zeland, PhD Thesis, 1999

[7] David R. Hardoon, Sandor Szedmak and John Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. Neural Computation, 16(12), p. 2639-2664, 2004

[8] Tony Hey, Stewart Tansley and Kristin Tolle. The Fourth paradigm: Data-intensive scientific discovery. Redmond, Microsoft Research, 2009

[9] Zeljko Ivezic, Andrew J. Connolly, Jacob T. VanderPlas and Alexander Gray. Statistics, data mining, and machine learning in astronomy: A practical Python guide for the analysis of survey data. Princeton University Press, 2014

[10] Leonid Kalinichenko, Dmitry Kovalev, Dana Kovaleva and Oleg Malkov. Methods and tools for hypothesis-driven research support: a survey. Informatica and Appications, 9(1), p. 28-54, 2015

[11] Igor Kononenko. Estimating attributes: analysis and extensions of RELIEF. In Proceedings of the European conference on machine learning (ECML'94), Catania, Italy, p. 171–182, 1994

[12] Manuel Mejia-Lavalle, Enrique Sucar and Gustavo Arroyo. Feature selection with a perceptron neural net. In Proceedings of the international workshop on feature selection for data mining: Interfacing Machine Learning and Statistics, p. 131–135, 2006

[13] Luis C. Molina, Lluis Belanche and Angela Nebot. Feature Selection Algorithms: A Survey and Experimental Evaluation. Data Mining, 2002. ICDM 2003. In Proceedings of 2002 IEEE International Conference on Data Mining, p. 306-313, 2002

[14] Hanchuan Peng, Fuhui Long and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), p. 1226–1238, 2005

[15] Alain Rakotomamonjy. Variable selection using SVM-based criteria. The Journal of Machine Learning Research, 3, p. 1357-1370, 2003

[16] Douglas Schales, Xin Hu, Jiyong Jang, Reiner Sailer, Marc Stoecklin and Ting Wang. FCCE: Highly Scalable Distributed Feature Collection and Correlation Engine for Low Latency Big Data Analytics, In Proceeding of 2015 IEEE 31st International Conference on Data Engineering, p. 1316–1327, Seoul, IBM Research Report, 2014

[17] Sohan Seth and Jose C. Prıncipe. Variable selection: A statistical dependence perspective. In Proceedings of the international conference of machine learning and applications (ICMLA'10), p. 931–936, 2010

[18] Nigel Williams, Sebastian Zander and Grenville Armitage. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. ACM SIGCOMM Computer Communication Review, 36(5), p. 5-16, 2006

[19] Ian H. Witten and Eibe Frank. Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, San Francisco, 2005

[20] Zheng Zhao and Huan Liu. Searching for interacting features. In Proceedings of the international joint conference on artificial intelligence (IJCAI'07), p 1156–1161, Hyderabad, India, 2007

Reducing the number of virtual experiments by estimating the correlation parameters of interacting hypotheses

Evgeny Tarasov

This paper presents the approach that helps to researcher to reduce the number of virtual experiments through decrease the count of tested hypotheses. This approach is based on correlation search between parameters of different hypotheses. A review and analysis of modern platforms with similar functionality is done. Methods for reducing the number of virtual experiments are surveyed, including the features selection algorithm, which allows to reduce investigated parameters space and identify the interaction between them. Next, functional requirements of designed system are formulated. We consider the practical problem which can be solved in the framework of this system and consider its particular case – assessment of the correlation parameters in astronomical hypotheses problem, which will be used as the test task during system debugging.