МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

«Новосибирский национальный исследовательский государственный университет»

(Новосибирский государственный университет)

Структурное подразделение Новосибирского государственного университета –

Высший колледж информатики НГУ

КАФЕДРА ЕСТЕСТВЕННОНАУЧНЫХ ДИСЦИПЛИН

**СОЗДАНИЕ БИБЛИОТЕКИ АЛГОРИТМОВ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ КЛИНИЧЕСКИХ ИССЛЕДОВАНИЙ В ПАРАЛЛЕЛЬНЫХ ГРУППАХ**

Дипломный проект

на квалификацию техник

Научный руководитель

к.ф.-м.н. ИВМиМГ СО РАН

Лукинов В.Л.

«\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_2017 г.

Студент IV курса,

гр. 14214

Кошкарева С. В.

«\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_2017 г.

Новосибирск

2017

СОДЕРЖАНИЕ

[ПЕРЕЧЕНЬ ТЕРМИНОВ И СОКРАЩЕНИЙ 3](#_Toc492846105)

[ВВЕДЕНИЕ 4](#_Toc492846106)

[1. ПОСТАНОВКА ЗАДАЧИ 7](#_Toc492846107)

[1.1. Описание предметной области 7](#_Toc492846108)

[1.2. Формулировка задачи 7](#_Toc492846109)

[1.3. Функциональные требования 8](#_Toc492846110)

[1.4. Нефункциональные требования 8](#_Toc492846111)

[1.5. Характеристики выбранных технических средств 9](#_Toc492846112)

[2. АЛГОРИТМЫ РЕШЕНИЯ ЗАДАЧИ 10](#_Toc492846113)

[2.1.1. ООП модель 10](#_Toc492846114)

[3. РАЗРАБОТКА БИБЛИОТЕКИ 11](#_Toc492846115)

[4. ОТЛАДКА И ТЕСТИРОВАНИЕ 12](#_Toc492846116)

[ЗАКЛЮЧЕНИЕ 13](#_Toc492846117)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 14](#_Toc492846118)

ПЕРЕЧЕНЬ ТЕРМИНОВ И СОКРАЩЕНИЙ

ВВЕДЕНИЕ

Целью данной работы является разработка программного комплекса алгоритмов, предназначенного для проведения статистического анализа данных, полученных в результате клинических исследований в параллельных группах.

На протяжении своего развития медицина всегда старалась найти более эффективные способы лечения и диагностики. Пройдя путь от интуитивных обобщений, метода проб и ошибок, объединяя разрозненный эмпирический опыт, она перешла к доказательности. Доказательная медицина подразумевает такой подход к медицинской практике, при котором каждое решение, относящееся к выбору метода лечения, должно иметь научное обоснование, которое основывается на данных, полученных в ходе четко спланированного исследования, использующего адекватные методы статистического анализа.

Статистический анализ является неотъемлемой частью практически любого исследования, и только с его помощью можно объективно судить о результатах исследования пополнить доказательную базу. Расчет статистических показателей, которые позволяют оценить достоверность различия, корреляцию и взаимное влияние анализируемых факторов происходит по определенной технологии с использованием математических функций и создания моделей.

В большинстве случаев клинические исследования проводятся в параллельных группах. При проведении исследований в параллельных группах испытуемые двух или более групп получают различные курсы лечения или различные дозы ЛС. Для достижения статистической достоверности испытуемые распределяются по группам методом случайной выборки. Модели исследований в параллельных группах считаются оптимальными для определения эффектов лечения и формулирования выводов на основе полученных результатов.

Ввиду того, что объемы данных и размеры групп (выборок) могут сильно варьировать, а данные быть весьма разнообразными, возникает необходимость в более гибком подходе и выборе методов статистического анализа, подходящих конкретной задаче.

Различные статистические методы предполагают нормальность распределения данных. Такие методы называют параметрическими. Они строятся на основе параметров выборочной совокупности и представляют функции этих параметров, непараметрические - функции от вариант данной совокупности с их частотами. Непараметрические критерии применимы к распределениям самых различных форм. Последние имеют определенные преимущества по сравнению с параметрическими, благодаря меньшим требованиям к их применению, большему диапазону возможностей и, часто, большей простоте реализации. Однако нужно учитывать более низкую точность этих критериев по сравнению с параметрическими.

К сожалению, на практике в ходе сбора данных далеко не всегда получаются полностью и правильно укомплектованные наборы. Опечатки и пропуски отдельных значений являются повсеместным явлением и поэтому, прежде чем начать применять статистические методы, обрабатываемые данные следует привести к «каноническому» виду. Для этого необходимо идентифицировать такие проблемы, а в дальнейшем, либо удалить фрагменты объектов с недостающими элементами, либо заменить имеющиеся пропуски и опечатки разумными значениями. Также еще одной проблемой «сырых» данных является наличие выбросов в исходных исследовательских данных. Под «выбросом» понимается наблюдение, которое слишком велико или слишком мало по сравнению с большинством других имеющихся наблюдений. Чувствительность разных статистических методов к наличию выбросов в данных неодинакова. Так, наличие выбросов может сделать использование некоторых статистических моделей невозможным и в то же время никак не сказаться на результатах другой.

Именно по вышеописанным причинам перед проведением статистического анализа необходимо выполнить проверку начальных данных на валидность, т.к. согласно практике, предварительное исследование данных может занимать до 50% времени, затраченного на анализ.

Валидизация данных является процессом обнаружения и исправления ошибок, таких как: пропущенные данные, выход величин за определенные пределы, последовательность дат, лабораторные показатели относительно нормы, взаимосвязь данных. На данный момент валидизация данных зачастую осуществляется при помощи визуальной «ручной» проверки. При визуальной проверке происходит проверка на полноту присланной документации, содержащей информацию, подлежащую обработке (индивидуальные регистрационные карты, бланки с результатами анализов). Производится визуальная оценка качества указанной информации (некорректные исправления, опечатки, нечитаемые данные. При визуальной проверке может потребоваться медицинская экспертиза (осуществление проверки информации на валидность, требующая специальных медицинских знаний и не реализуемая программными средствами). Программная проверка существенно ускорит дальнейший анализ, а так же окажется более эффективной по сравнению с проверкой, выполняемой визуально. Во время программной валидизации данных выявляются пропущенные данные, опечатки, величины, выходящие за определенные пределы, определяется последовательность всех дат, лабораторные показатели сверяются с референтными пределами. Так же программа должна решать такую специфическую задачу, как определение нормальности распределения данных.

В ходе выполнения дипломной работы была изучена предметная область, выбрано и изучено программное средство, а именно язык R, разработаны и описаны архитектура библиотеки и алгоритмы подпрограмм, написан программный код библиотеки, которая идентифицирует потенциальные проблемы исследования данных, выполнены анализ, тестирование и отладка библиотеки, рассмотрен пример применения на реальных задачах.

1. ПОСТАНОВКА ЗАДАЧИ
   1. Оп**исание предметной области**

При статистическом анализе данных клинических исследований необходимо выполнять рутинные процедуры, поддающиеся автоматизации. Создание единого промышленного программного комплекса алгоритмов статистического анализа данных клинических исследований в параллельных группах, реализованных в виде R-библиотеки с сопутствующей документацией, позволит существенно сократить время проведения статистического анализа и поможет проводить более качественные исследования в сжатые сроки.

* 1. Формулировка задачи

Целью работы являлась реализация системы автоматизированного управления установкой БНЗТ, а именно: разработка консоли оператора, сопряжение периферии с консолью и отображение данных оператору с периодом не менее раза в секунду.

Контроль управляющих узлов источниками питания должен осуществляться при помощи заложенных в них протоколах связи (например, SCMP). Диагностика состояния ускорителя осуществляется через измерение токов, напряжений и температур.

В качестве измерителей параметров были выбраны дистанционные устройства ввода-вывода, управляемые консолью оператора через Ethernet (протокол Modbus).

Так же требуется ведение журнала всех значений с периодом не менее 1 раза в секунду.

Были поставлены и решены следующие задачи:

* Изучена текущая конфигурация установки;
* Произведен анализ требуемых устройств;
* Изучены принципы работы с периферийными устройствами;
* Разработана консоль оператора;
* Налажена связь консоли оператора с устройствами управления и диагностики;
* Разработана база данных для хранения данных об эксперименте;

Программирование осуществлялось на языке R, с использованием следующих пакетов:

* base
* methods
* utils
* grDevices
* xlsx

Использовалась система контроля версий git и репозиторий на сервере GitHub.

* 1. Функциональные требования

В рамках дипломной работы были поставлены следующие задачи:

* Ввод данных должен осуществляется в виде Excel-таблиц или внутренней структуры языка R - data.frame.
* Поиск пропущенных значений (незаполненных полей).
* Поиск опечаток.
* Поиск «выдающихся значений» или «выбросов».
* Исследование нормальности распределений методами Шапиро-Уилка, Андерсона-Дарлинга, Крамера-фон Мизеса, Лиллиефорса, Шапиро-Франчиа, вывод результатов в виде таблицы в .doc-файле или .csv-файле.
* Сравнение числовых показателей в 2 группах U-критерием Манна-Уитни или t-критерием Стьюдента в зависимости от типа распределений показателя или по выбору пользователя, вывод результатов в виде таблицы в .doc файле или .csv файле, построение графиков boxplot.
* Сравнение категориальных показателей в двух группах точным двусторонним критерием Фишера или критерием хи-квадрат в зависимости от типа распределений показателя, или по выбору пользователя, вывод результатов в виде таблицы в .doc-файле или .csv-файле, построение графиков в виде столбчатых диаграмм.
  1. Нефункциональные требования
* Реализация методами ООП, используя объектную модель S4 в языке R.
* Использование системы контроля версий Git в связке с сервером GitHub или Bitbucket.
* Тестирование созданной библиотеки.
* Использование встроенной графической системы ggplot2.
  1. Характеристики выбранных технических средств

Для написания проекта был выбран язык R и среда разработки RStudio.

R – язык программирования высокого уровня, предназначенный для статистической обработки данных с упором на визуализацию и воспроизводимость. Так же является свободной кроссплатформенной программной средой вычислений с открытым исходным кодом в рамках проекта GNU. R – интерпретируемый язык с интерфейсом командной строки. R сочетает в себе функциональное, процедурное, объектно-ориентированное и рефлексивное программирование, т.е. является мультипарадигменным языком.

Был разработан на основе языка программирования S. С 2011 года поддерживается и развивается организацией R Foundation. Сегодня R является безусловным лидером среди свободно распространяемых систем статистического анализа. R обладает хорошей расширяемостью с помощью пакетов. Каждый такой пакет представляет собой библиотеку, содержащую набор специфических функций. В среде R реализованы многие статистические методы: линейные и нелинейные модели, проверка статистических гипотез, анализ временных рядов, классификация, кластеризация, графическая визуализация. Так же, одной из особенностей языка является поддержка графических возможностей, которая позволяет визуализировать данные в виде различных графиков и диаграмм.

R имеет обширное сообщество и развитую систему поддержки, включающую обновление компонентов среды, интерактивную помощь и различные образовательные ресурсы.

1. АЛГОРИТМЫ РЕШЕНИЯ ЗАДАЧИ
   1. ООП модель

Т.к. ключ и значение являются переменными типа list, то они имеют вложенную структуру, представляющую из себя массив массивов.

Структура данных таблицы, с которой работает библиотека, может состоять из четырех типов значений: дата, непрерывные, дискретные и категориальные (номинальные). Непрерывные переменные (continuous variables) могут принимать любые численные значения, которые естественным образом упорядоченны на числовой оси (например, рост, вес, АД, РОЭ). Дискретные переменные (discrete variables) могут принимать счётное множество упорядоченных значений, которые могут просто обозначать целочисленные данные или ранжировать данные по степени проявления на упорядоченной ранговой шкале (клиническая стадия опухоли, тяжесть состояния пациента). Категориальные переменные (categorial variables) являются неупорядоченными и используются для качественной классификации (пол, цвет глаз, место жительства); в частности, они могут быть бинарными (дихотомическими) и иметь категорические значения: 1/0, да/нет, имеется/отсутствует. Поэтому для каждого из четырех типов реализован свой метод поиска ошибок.

* + 1. Описание алгоритма поиска опечаток

Ниже описан метод для категориальных значений таблицы.

Получаем доступ к элементам определенной колонки в таблице данных по индексу, который был передан при вызове метода. Присваиваем значения этих элементов новой переменной, с которой будем работать. В цикле, последовательно проверяется каждый ее элемент.

В первую очередь осуществляется поиск пропущенных значений. Если проверка выявила существование незаполненной ячейки в колонке, то индекс такого элемента передается в переменную-слот объекта класса missingValue для дальнейшей раскраски. Вызывается метод, который производит запись сообщения о найденной ошибке и ее координатах в пользовательский текстовый файл-отчет.

Если результат поверки на заполнение оказался успешным и элемент действительно существует, начинается проверка на совпадение этого элемента с одним из значений словаря. Для корректности сравнения, значение словаря и значение элемента колонки приводятся к верхнему регистру, во избежание ошибок.

Если значение колонки не совпало ни с одним значение словаря, проверяется совпадение элемента с одним из ключей. Оба значения так же приводятся к верхнему регистру. Если совпадение найдено, то по нужному индексу в выходной таблице производится замена ключа словаря на соответствующее ему значение. Индексы элемента передаются в переменную-слот объекта класса Misprint для раскраски выходной таблицы и происходит вызов метода

Если же элемент колонки не совпал ни с одним ключом, то индексы этого элемента в таблице передаются в переменную-слот объекта класса UnsolvedMisprint, и производится запись сообщения об ошибке в файл.

Метод для непрерывных и дискретных значений таблицы.

Так же, как и в методе, реализованном для категориальных переменных, производится доступ к определенной колонке и проверка ее элементов на пропуски.

Значение элемента колонки подставляется под шаблон, заданный регулярным выражением, которое имеет тип: начало строки число, а дальше либо конец строки, либо (разделитель точка или запятая одно и более чисел). Если элемент не подходит под шаблон, значит, что в нем содержатся буквы, или любые другие символы. Следовательно, индексы элемента передаются слоту объекта класса UnsolvedMisprint, который, как было сказано ранее, содержит индексы элементов таблицы, которые будут раскрашены как опечатки, которые не удалось исправить. Производится запись сообщения оданной ошибки в файл.

1. РАЗРАБОТКА БИБЛИОТЕКИ
2. ОТЛАДКА И ТЕСТИРОВАНИЕ

ЗАКЛЮЧЕНИЕ

В ходе работы были проведен обзор средств работы с большими данными и на их основе разработана система в рамках лямбда архитектуры. Система удовлетворяет всем поставленным требованиям:

• Легко горизонтально масштабируема

• Обладает высокой отказоустойчивостью

• Обладает высоким быстродействием

• Легко расширяема на разные источники данных

• Реализует требуемый функционал

В рамках работы была доказана целесообразность применения технологий работы с большими данными, применительно к области улучшения безопасности дорожного движения. На основании результатов, полученных с помощью этой системы, могут быть предприняты мероприятия по снижению аварийности дорог. Также система эффективна как мониторинговая.

Система обладает большими возможностями и потенциалом к дальнейшему развитию и улучшению. Таким образом, все поставленные задачи в ходе работы были выполнены, а цель работы можно считать достигнутой.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ