ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ ДЛЯ ОПРЕДЕЛЕНИЯ ГРУППЫ РИСКА СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

П.А. Зяблецев
Томский политехнический университет
paz4@tpu.ru

Введение

В настоящее время существует несколько способов определения риска возникновения сердечно сосудистых заболеваний. Первая группа методов - это обследование в медицинских учреждениях, но стоит отметить что мало кто из нас обратится в поликлинику при отсутствии боли или явного дискомфорта, а ишемическая болезнь сердца может проходить без каких-либо признаков. Другие методы анализа риска основаны на эмпирических данных проспективных исследований за большими группами людей. Благодаря таким исследованиям было обосновано понятие суммарного риска и разработана таблица оценки риска сердечно-сосудистых заболеваний SCORE. Однако на данный момент данная таблица имеет много недостатков и важной задачей является создание системы оценки группы риска сердечно-сосудистых заболеваний с высокой точностью. Для решения данной задачи было решено применить методы интеллектуального анализа данных.

Интеллектуальный анализ данных

Интеллектуальный анализ данных – это совокупность математических моделей, численных методов, программных средств и информационных технологий, обеспечивающих обнаружение в эмпирических данных доступной для интерпретации информации и синтез на основе этой информации ранее неизвестных, нетривиальных и практически полезных для достижения определенных целей знаний. [1]

Процесс интеллектуального анализа данных состоит из следующих этапов:

1. Сбор данных
2. Подготовка данных (фильтрация, дополнение, кодирование)
3. Использование методов анализ данных
4. Верификация результата (проверка полученных закономерностей и моделей)
5. Применение полученной модели для прогнозирования.

Рассмотрим каждый этап более подробно:

Первый этап подразумевает собой сбор данных различного формата из различных источников: структурированных (базы данных) и не структурированных (социальные сети, текстовые файлы, вебсайты). Эти данные объединяются в витрины данных или хранилища данных, с которыми в дальнейшем мы уже можем работать. Даже если у нас есть реляционная база данных и данные из других источников нас не интересуют рекомендуется перенести информацию из БД в хранилище данных, так как аналитические запросы к оперативной БД будут занимать слишком много серверного времени, блокирую таблицы.

Этап подготовки данных для анализа является крайне важным, и большая ошибка относиться к нему небрежно. Так как если мы имеем данные низкого качества на входе, то, вероятнее всего, и результат будет такого же качества. Под качеством данных подразумевается их полнота, точность, своевременность и возможность интерпретации. На данном этапе необходимо изучить данные на предмет пропуска значений, дублирования, шумов и выбросов. Для улучшения качества данных следует исключить объекты с пропущенными значениями из обработки, убрать дублирование, исключить данные с крайними значениями. Этот этап занимает много времени и сил, но его необходимость неоспорима.

На третьем этапе мы используем методы анализа данных. Это может быть машинное обучение, применение деревьев решений, искусственных нейронных сетей, корреляционный и регрессионный анализ, многомерный анализ данных (технология OLAP) и многие другие. После этого этапа мы получаем модель, которая прогнозирует значение на основе исходных данных или распределяет данные на категории, классы или кластеры.

На этапе проверки мы определяем достаточную ли прогнозную точность имеет наша модель и удовлетворяет ли нашим ожиданиям.

И на последнем этапе мы используем нашу модель уже непосредственно для наших целей, будь то улучшение бизнес процесса, прогнозирование заболевание или предсказание природных катастроф.

Применение методов анализа для выявления риска возникновения сердечно-сосудистых заболеваний.

На первом этапе мы имеем данные из Томского регистра острого инфаркта миокарда, которые ведется с 1984 года (всего учтено около 50 000 случаев, в том числе более 25 000 подтвержденных случаев острого инфаркта миокарда). Данные представляют собой файлы Excel с данными о больных. Всего имеется более 50 тысяч записей с более чем 100 различными параметрами.

Наиболее важный и трудоемкий этап – это подготовка данных. Наш исходный файл Excel представляет собой список пациентов и большое количество их параметров (209 столбцов в таблице). Для построения хранилища данных нам нужны не все параметры, этой таблицы, а лишь те, которые

позволят вычислять суммарный риск возникновения серьезных сердечно-сосудистых случаев, а также смерти от сердечно-сосудистых заболеваний. Некоторые подмножества столбцов требуют свертки в отдельные атрибуты. Полезные данные из таблицы Excel будут перенесены в таблицу реляционной БД, так как в дальнейшем с ней проще работать с помощью запросов SQL и далее на основе этой таблицы создавать хранилище данных.

Из 209 проанализированных столбцов были выбраны 12, а именно: ID пациента, ID случая, социальное положение, возраст, стенокардия, артериальная гипертензия, сахарный диабет, курение, приступ произошел вовремя, индекс массы тела, фермент сыворотки крови, дата смерти.

Для переноса полезных данных из таблицы Excel в БД необходимо создать пакет переноса данных. С помощью SQL Server Data Tools for Visual Studio 2012 создается проект пакета переноса данных SSIS (SQL Server Integration Services).

Данные являются достаточно сырыми после переноса. Например, в файле Excel нет как такого столбца как возраст, но есть год рождения и дата приема, на основании которых мы может определить возраст. Помимо этого, нужно провести работу по форматированию всех параметров к нужным типам данных (в исходном файле типы самые разные и не соответствуют нашим задачам).

После этого можно создавать хранилище данных по схеме звезда, это необходимо для дальнейшего анализа данных с помощью технологии OLAP. Структура хранилища представлена на рисунке 1.

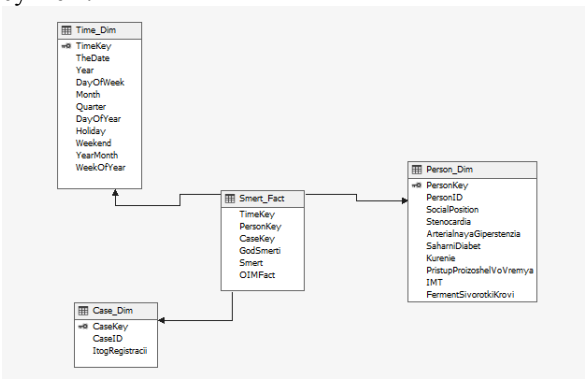

Рис. 1. Структура хранилища данных

После создания хранилища данных в SQL Server нужно заполнить его данными из нашей БД. Для этого будем использовать пакет по переносу данных из нашей БД в хранилище данных.

Таким образом после запуска пакета наше хранилище данных заполнится. Можно приступать к созданию “куба данных”.

Создание куба будет происходить с помощью ПО «SQL Server Data Tools», с помощью которого мы создавали пакеты SSIS. Только теперь в создании проекта нам нужно выбрать “Проект интеллектуального анализа данных и многомерных данных служб Analysis Services”.

Создание такого куба данных позволяет получать быстрые отчеты по различным выборкам, это позволит понять и проанализировать зависимости между различными параметрами.

На данном этапе интеллектуальный анализ не завершен, планируется применить методы машинного обучения для предсказания сердечно-сосудистых заболеваний. Тем не менее применение технологии OLAP позволяет произвести предварительный анализ данных и сравнение всех возможных вариантов.

Заключение

В результате выполнения данной работы было спроектировано хранилище данных для последующей обработки этих данных технологией OLAP. Данные из громоздких и неудобных для обработки файлов Excel были перенесены в хранилище данных со всеми необходимыми конвертациями. Был создан «OLAP-куб», который позволяет получать отчеты о всех зависимостях между атрибутами в режиме реального времени. Данная работа является одним из важнейших этапов по созданию системы оценки группы риска сердечно-сосудистых заболеваний. Данный этап имеет практическую пользу для аналитиков. Эксперт может проверять различные теории в режиме реального времени, все данные можно вращать и получать необходимые «срезы» и «разрезы». Благодаря созданию «куба» появляется возможность проанализировать данные РОИМ Томской области и на основании этих данных сделать вывод о наиболее опасных факторах риска и их сочетаниях.

Список использованных источников

1. Дюк В., Самойленко А. Data Mining: учебный курс (+CD-ROM). 2001 г. Издательство: Питер. Серия: Учебный курс. – 368 с.
2. Введение в многомерный анализ. [Электронный ресурс]. – URL: https://habrahabr.ru/post/126810/ (Дата обращения 11.10.18г.)
3. Difference Between Data Mining and OLAP. [Электронный ресурс]. – URL: http://www.differencebetween.com/difference-between-data-mining-and-vs-olap (Дата обращения 15.09.18г