СПОСОБЫ АВТОМАТИЗИРОВАННОГО АНАЛИЗА РАСХОЖДЕНИЙ ЗНАЧЕНИЙ ПАРАМЕТРОВ НА БОЛЬШИХ СОВОКУПНОСТЯХ ДАННЫХ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Мыц С. И.

Волорова Н. А. – канд. техн. наук, доцент

В условиях наличия больших объемов данных возникает необходимость в автоматизированном анализе их параметров. Ни одному человеку или группе людей не будет под силу вручную обработать весь объем информации поступающей от достаточно крупной современной информационной системы. Для решения такой задачи применяются инструменты, созданные с использованием алгоритмов и математических методов.

Взглянем на некоторый частный случай и на его примере рассмотрим возможные подходы к решению проблемы. Допустим, у нас есть некоторый веб-сайт, к которому пользователи задают запросы, а затем как-то взаимодействуют с результатами этих запросов. Чтобы исследовать это взаимодействие в глобальном плане, вводятся параметры, считаемые по некоторому множеству запросов. Эти параметры назовём метриками.

В контексте текущей задачи метрикой будем называть некоторую числовую величину, которая считается на основе данных из множества запросов и является показателем некоторого интересного нам свойства этого множества. В качестве примеров простых метрик можно привести следующие:

- общее количество действий пользователя со всеми результатами запроса
- отношение количества действий пользователя с опрделённым результатом к количеству показов этого результата
- среднее время до первого взаимодействия пользователя с результатами в рамках запроса
- количество пользователей, взаимодействующих с системой

Могут возникать ситуации, когда значения метрик изменяются и нам нужно понять, по каким причинам это произошло и что больше всего повлияло на это изменение. Это можно делать с помощью анализа значений метрик на срезах данных. Срезом данных назовём подмножество исходного множества запросов, взятое по какому-то критерию.

Теперь можно обобщить постановку задачи. Пусть у нас имеется множество объектов (см. множество запросов), имеется числовая функция от произвольного подмножества объектов (см. метрика) и есть набор предикатов, позволяющих выделять подмножества исходного множества (см. срезы данных).

Поставленную выше задачу можно решать различными способами ввиду того, что нет чёткого определения “наибольшего влияния на изменения”, но мы пока остановимся на трёх подходах:

- Использование анализа чувствительности
- Деревья решений и коэффициент влияния факторов
- Сравнение среза с другой подвыборкой

Воспользоваться анализом чувствительности можно следующим образом:

1. Возьмём в качестве аргументов некоторой функции все имеющиеся срезы (0 или 1, принадлежит запрос этому срезу или нет), а затем с помощью регрессии построим такую искусственную функцию, которая будет приближать значение метрики на совокупности срезов.
2. Подсчитаем коэффициенты Соболя (Sobol Indices) для этой функции и на основе их в качестве результата вернём список, отсортированный по их значениям.

Деревья решений являются одним из подходов в машинном обучении, позволяющим решать, среди прочих, задачу регрессии. Полезным побочным эффектом их применения является возможность подсчёта “полезности” отдельного аргумента в деле предсказания значения целевой функции. Можно воспользоваться этими коэффициентами для выделения менее и более “важных” срезов, по аналогии с предыдущим пунктом.

Ещё одним способом решения задачи является следующий:

1. Рассмотрим каждый из срезов данных.
2. Выделим ему срез для сравнения: случайный срез аналогичного размера, всё непопавшее в срез и т.п.
3. Сравним с использованием статистических критериев значения на исходном и парном срезе.
4. Используя результаты сравнения отранжируем срезы.

На основе перечисленных выше методов строится обобщённый инструмент, предоставляющий возможности конфигурирования, предподсчёта данных, запроса списка наиболее повлиявших срезов и визуализации результатов.

Подход, описанный в этом докладе, позволяет автоматизировать анализ и помогает в исследовании изменений большой системы. Благодаря этому можно оперативнее реагировать на проблемы, а также лучше и точнее анализировать результаты вносимых в систему правок.

Список использованных источников:

1. C.D. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval, Cambridge UP, 2008