Предобработка данных (Data Preprocessing)

> Разделы: Бизнес-задачи

В процессе предобработки данных производится их подготовка к анализу, в результате которой они приводятся в соответствие с требованиями, определяемыми спецификой решаемой задачи.

Предобработка является важнейшим этапом Data Mining, и если она не будет выполнена, то дальнейший анализ в большинстве случаев невозможен из-за того, что аналитические алгоритмы просто не смогут работать или результаты их работы будут некорректными. Иными словами, реализуется принцип GIGO — garbage in, garbage out (мусор на входе, мусор на выходе).

Предобработка данных включает два направления: очистку и оптимизацию.

Очистка производится с целью исключения различного рода факторов, снижающих качество данных и мешающих работе аналитических алгоритмов. Она включает обработку дубликатов, противоречий и фиктивных значений, восстановление и заполнение пропусков, сглаживание, подавление шума и редактирование аномальных значений. Кроме этого, в процессе очистки восстанавливаются нарушения структуры, полноты и целостности данных, преобразуются некорректные форматы.

Оптимизация данных как элемент предобработки включает снижение размерности, выявление и исключение незначащих признаков. Основное отличие оптимизации от очистки в том, что факторы, устраняемые в процессе очистки, существенно снижают точность решения задачи или делают работу аналитических алгоритмов невозможной. Проблемы, решаемые при оптимизации, адаптируют данные к конкретной задаче и повышают эффективность их анализа.

Предобработка данных выполняется на протяжении всего процесса Data Mining: при выгрузке данных из первичных источников и OLTP-систем, в хранилище данных и в аналитической платформе.

В Loginom существует группа специализированных обработчиков, относящихся к разделу трансформация и осуществляющих все виды предобработки данных. Подробнее о качестве собранных данных и важности их очистки в статье «Очистка данных перед загрузкой в хранилище».