Понижение размерности (Data reduction)

> Синонимы: Сокращение размерности, Снижение размерности

В аналитических технологиях под понижением размерности данных понимается процесс их преобразования в форму, наиболее удобную для анализа и интерпретации. Обычно оно достигается за счет уменьшения их объема, сокращения количества используемых признаков и разнообразия их значений.

Часто анализируемые данные являются неполными, когда они плохо отображают зависимости и закономерности исследуемых бизнес-процесов. Причинами этого могут быть недостаточное количество наблюдений, отсутствие признаков, которые отражают существенные свойства объектов. В этом случае применяется обогащение данных.

Понижение размерности применяется в противоположном случае, когда данные избыточны. Избыточность возникает тогда, когда задачу анализа можно решить с тем же уровнем эффективности и точности, но используя меньшую размерность данных. Это позволяет сократить время и вычислительные затраты на решение задачи, сделать данные и результаты их анализа более интерпретируемыми и понятными для пользователя.

Сокращение числа наблюдений данных применяется, если решение сравнимого качества можно получить на выборке меньшего объема, сократив, тем самым, вычислительные и временные затраты. Особенно это актуально для алгоритмов, не являющихся масштабируемыми, когда даже небольшое сокращение числа записей приводит к существенному выигрышу в вычислительных временных затратах.

Сокращение числа признаков имеет смысл проводить тогда, когда информация, необходимая для качественного решения задачи, содержится в некотором подмножестве признаков и необязательно использовать их все. Особенно это актуально для коррелирующих признаков. Например, признаки «Возраст» и «Стаж работы», по сути, несут одну и ту же информацию, поэтому один из них можно исключить.

Наиболее эффективным средством сокращения числа признаков являются факторный анализ и метод главных компонент.

Сокращение разнообразия значений признаков имеет смысл, например, если точность представления данных избыточна и вместо вещественных значений можно использовать целые без ухудшения качества модели. Но при этом уменьшится занимаемый данными объем памяти и вычислительные затраты.

Подмножество данных, полученное в результате сокращения размерности, должно унаследовать от исходного множества столько информации, сколько необходимо для решения задачи с заданной точностью, а вычислительные и временные затраты на сокращение данных не должны обесценивать полученные от него преимущества.

Аналитическая модель, построенная на основе сокращенного множества данных, должна стать проще для обработки, реализации и понимания, чем модель, построенная на исходном множестве.

Решение о выборе метода сокращения размерности основывается на априорном знании об особенностях решаемой задачи и ожидаемых результатах, а также ограниченности временных и вычислительных ресурсов.

В Loginom понижение размерности данных осуществляется с помощью обработчиков факторный анализ и корреляционный анализ. Они применяются для оценки предполагаемой зависимости факторов с целью понижения размерности их пространства.