Измерение расстояния между объектами с пропусками при помощи метода попарного сравнения

Бишук Антон Юрьевич

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Москва
2021 г

Задача построения метода сравнения объектов

Цель

Предложить метод, способный определять расстояние между объектами путем попарного сравнения объектов между собой

Решаемая проблема

Описания объектов могут содержать пропущенные значений некоторых показателей. Это означает то, что мы не знаем значение признака для текущего объекта, однако всё же необходимо объекты сравнить между собой.

Идеи Решения

Методы решения

Представим ряд подходов, применение которых, позволяет сравнивать объекты между собой, даже если некоторые значения неизвестны.

1. Заполнить все пропуски одним специальным символом
2. Убрать те признаки из сравнения, значение которых неизвестно.
3. Заполнить значение средним по признаку (среднее арифметическое/медиана/мода) / средним по объекту.

Методы сравнения

1. Косинусное расстояние,
2. Поэлементное сравнение.

Специальный символ

Данный подходит под узкий спектр задач. Он принесет точность если специальный символ будет вписывать в логику задачи. Примером удачного применения может служить заполнение информации о просмотренных фильмах на сайте. Если фильм не просмотрен на сайте, это не значит, что пользователь его не видел, однако такой смело можно предположить и заполнить пропуск нулём.

1. Заменяем все пропуски подобранным специальным символом.
2. Считаем косинусное расстояние между текущим вектором и всеми остальными и выбирает ту пару, расстояние между которыми минимально.
3. При поэлементном сравнении: для каждой пары если элементы совпадают, то увеличиваем счетчик на 1 и в конце нормируем это на длину объекта.

Убрать признаки из сравнения

Данный подход хорошо подходит если матрица объектов слабо разрежена, поскольку в противном случае может так получиться, что не найдется объектов, с которыми мы смогли сравнить текущий.

1. Убираем из рассмотрения все столбцы-признаки, значение которых неизвестно для текущего объекта
2. а) Сравниваем наш объект только с теми объектами, у которых в оставшихся признаков не осталось пропуска.
   б) Каждый раз при сравнении с другим объектом, также вычеркиваются признаки, о которых не известно в сравниваемом объекте.
3. Если применение подхода a) позволяет сравнить почти все элементы друг с другом, то косинусное расстояние и поэлементное сравнение применяется точно так же, как в прошлом пункте.
4. Если применяется подход b, то после подсчета косинусного расстояния и поэлементного расстояния, полученный результат домножается на коэффициент доверия, пропорциональный числу элементов, по которому проводилось сравнение.

Заполнить значение

Данный подход имеет широкое применение и множество вариаций, основанных на теории информации, однако тут будет представлен наивный подход. Заполнение пропуска средним по признаку происходит в случае, когда все признаки объекта имеют разную природу(так например время подъема лучше заполнить средним среди времени подъема всех людей, чем средним среди возраста, времени года и весом), а заполнение пропуска средним по объекту возможно, когда признаки представляют из себя признаки одной природы(например просмотр фильмов на сайте. Если человек посмотрел много фильмов, то скорее всего он посмотрит и тот, который еще не посмотрел)

1. Заполняем пропуск согласно выбранной политике
2. Применяем метод с косинусным расстоянием и поэлементным сравнением как это делалось в первом пункте.