А. А. Сычёв

ИМПЛЕМЕНТАЦИЯ МЕТОДА ЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ

Рассматривается имплементация метода логического анализа данных, способа генерации паттернов, с помощью которых можно предсказывать результаты, базируясь на входных данных.

Введение

Логический Анализ Данных - это метод, основанный на комбинаторных и оптимизационных подходах. В основе Логического Анализа Данных лежит базовая концепция дифференциации и интеграции содержащихся в наборе данных. Данные могут либо содержать позитивный, негативный результат, либо быть новыми, неклассифицированными.

I. Описание процесса бинаризации и минимизации атрибутов

В работе данного метода используются бинарные атрибуты. Процесс преобразования небинарных данных в бинарные называется бинаризацией. Согласно концепции бинаризации значения атрибутов могут быть преобразованы в набор бинарных атрибутов, основываясь на пороговых значениях [1]. Для определения пороговых значений требуется отсортировать значения каждого атрибута по убыванию, чередуя при равенстве значений атрибуты из положительных и отрицательных записей. Новообразованный бинарный атрибут получает значение 1, если значение оригинального атрибута было больше порогового значения, и 0, если меньше. Способ определения порогового значения указан в формуле 1.

$$CutPointValue = \frac{(A_i + A_{i-1})}{2}. \qquad (1)$$

В результате получится новый бинарный атрибут A_1. Данный атрибут будет иметь значение 1, если соответствующие значения A превысят пороговое значение, и 0 в обратном случае.

Для того, чтобы определить, какие атрибуты следует убрать, проводится корреляционный анализ. Само понятие корреляции представляет собой отношение между двумя или более объектами. Корреляция между двумя атрибутами определяется тем, насколько атрибуты схожи друг с другом. Корреляция атрибутов может быть рассчитана как корреляционный коэффициент, который может быть как положительный, так и отрицательный. Корреляционный коэффициент рассчитывается по формуле 2.

$$CC = \frac{n(\sum AB) - (\sum A)(\sum B)}{\sqrt{(n\sum A^2 - (\sum A)^2)(n\sum B^2 - (\sum B)^2)}}. \qquad (2)$$

Если значение корреляционного коэффициента между атрибутом и результатом приближается к 1 или -1, то это значит, что атрибут и результат линейно взаимосвязаны. Для независимых атрибута и результата корреляционный коэффициент равен 0.

II. Генерация шаблонов

С помощью полученных атрибутов строятся шаблоны, которые будут проводить классификацию. Для этого использовался алгоритм Frequent Pattern-Growth. Данный алгоритм строит дерево, в качестве ветвей которого используются значения бинарных атрибутов, а в качестве листьев – результат обучающей записи. Во время построения дерева в случае встречи ветви с тем же значением на той же позиции, индекс данной ветви увеличивается на 1, а сам повторяющийся узел не добавляется. Затем для каждого из результата строится дерево пути, которое включает в себя вышеупомянутые ветви. Если предмет встречается два или более раза, то его индексы, т.е. частоты появлений в условном базисе, суммируются. Затем выбираются узлы с индексом больше некоего значения. В результате получается набор шаблонов, которые могут предсказывать значения.

III. Выводы

Данный метод логического анализа данных используется для валидации результатов уже имеющихся алгоритмов классификации или отличающихся способов реализации данного алгоритма. Это используется в сфере автоматизированного анализа данных телеметрии с целью улучшения надежности обработки подобных данных.

Список литературы

1. Yadav, A. Logical analysis of data. Report / A. Yadav, A. Singh, P. H. Sheth, S. Gangopadhyay // IIT Roorkee. – 2016. – P. 19-45.

Сычёв Алексей Анатольевич, магистрант кафедры программного обеспечения инфомационных технологий БГУИР, aliakseisychou@gmail.com.

Научный руководитель: Скобцов Вадим Юрьевич, доцент кафедры программного обеспечения инфомационных технологий, кандидат технических наук, доцент, vasko_vasko@mail.ru.