ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

Отчет о программном проекте

на тему «Разработка системы предсказания успешного завершения учебной дисциплины»
(промежуточный, этап 1)

Выполнил:
студент группы БПМИ185 _____ Александра Игоревна Лежанкина
Подпись И.О. Фамилия

_____
Дата

Принял:
руководитель проекта Андрей Андреевич Паринов
Имя, Отчество, Фамилия

_____
Должность

_____
Место работы

Дата _____2020 _____ _____
Оценка (по 10-тибалльной шкале) Подпись

Москва 2020

Содержание

Задачи:

Изучение и сравнение алгоритмов кластеризации

Ward hierarchical clustering

Теория

Данный вид кластеризациии является восходящим. Изначально каждый студент – это отдельный кластер, затем объекты сливаются во всё более крупные кластеры: на каждом шаге выбирается пара кластеров с минимальным расстоянием друг от друга и объединяется. При этом данное расстояние хранится для последующего подсчёта оптимального числа кластеров. Объединение происходит до тех пор, пока все студенты не сольются в один кластер.

Структура данных

Для хранения и обработки данных использую структуру Непересекающихся множеств, которая позволяет легко объединять множества (кластеры), хранит информацию в виде списка родителей, где каждый i-й индекс соответствует i-ому студенты. Реализую функцию как класс с методами

- «find» – поиск родителя;
- «union_sets» – объединение множеств;
- «get_item» – возвращени элемента, соответствующего определённому индексу;
- «count_sets» – поиск числа кластеров;
- «count_elems» – поиск числа элементов в кластере;
- «result» – кластеры.

Формула расстояния

Формула для подсчёта расстояния между кластерам U и V. ρ – расстояние в Евклидовой метрике

- **Метод Уорда** (англ. *Ward's method*)

$$\mathrm{R}_{\mathrm{ward}}(U,V) = \frac{|U|\cdot|V|}{|U|+|V|}\rho^2\left(\sum_{u\in U}\frac{u}{|U|},\sum_{v\in V}\frac{v}{|V|}\right)$$

Оптимальное число кластеров

Изначально запускается функция кластеризации с единичным параметром на месте числа желаемых кластеров, при этом минимальное расстояние, найденное на каждом шаге алгоритма, записывается в список.

Далее мы проходим по образованному списку в поисках максимальной разницы между соседними расстояниями. Шаг, на котором будет найдено это расстояние, будет являться переменной I в формуле num = N – i + 1, которая отвечает за число кластеров, где N – число студентов.

Затем функция кластеризации запускается повторно, но уже с параметром num. Итоговое разбиение является ответом.

Входные данные

Хранятся вводимые данные в виде списка списков, где каждому студенту соответствует определённый индекс.

Среди характеристик полезны:

- Оценки за дисциплины;
- Количество часов, затраченных на освоение предмета;
- Идентификаторы посещения различных курсов или другие полезные скиллы (1 при наличии, 0 иначе).

Выходные данные

В итоге мы имеем набор кластеров оптимального размера. Данные выводятся в виде ключа – номера кластера и набора студентов, которые вошли в кластер. Пока в реализации это просто индекс, но в зависимости от ввода это могут быть имена студентов или набор характеристик.

Стандартизация

Так как данные могут иметь различный масштаб: например, оценки колеблются от 1 до 10, а число часов может доходить до сотни и более, будет полезно стандартизировать данные так, что их среднее значение будет равно 0, а отклонение 1. Для этого буде использовать StandartScaler и функцию fit_transform.

Equivalence Classification Algorithm (Eclat)

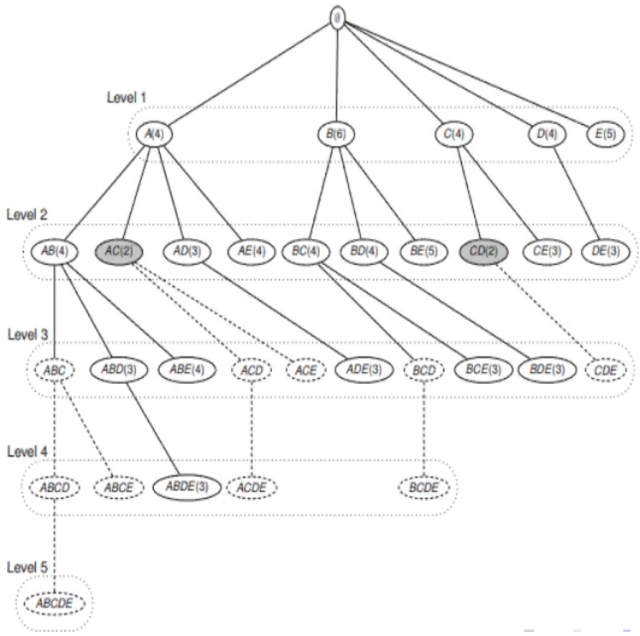
Теория

В отличие от других алгоритмов поиска ассоциативных правил данный использует поиск в глубину. Также в данном алгоритме база данных сканируется только единожды.

Для работы нам понадобится ввести несколько определений:

- Support (Поддержка) – идентификатор того, насколько часто встречается объект среди транзакций. Вычисляется, как число транзакций, содержащих те объекты, чью support мы ищем, делённое на количество транзакций. Измеряется в процентах (*100). Набор характеристик будет представлять интерес, если его Support выше, введённого ранее значения. Этот параметр (Минимальная поддержка) будет зависеть от того, с какой базой данных мы работаем, и сколько в ней студентов. На практике будет понятнее, какой размер минимальной поддержки делает работу эффективнее. Условно можно считать его равным 20%. Представленный тип хранения данных для алгоритма Eclat позволяет быстро рассчитывать поддержку. Суммарное число транзакций нам дано изначально, а число транзакций, в которых содержится наш объект item хранится как множество TID set. Чтобы посчитать число транзакций, в которых встречается несколько объектов, мы просто объединяем их и считаем мощность. Это можно сделать очень просто, если хранить данные в контейнере set.
- Confidence (Достоверность/Вероятность) – идентификатор корректности работы ассоциативного правила. Confidence(i | j) = Support(ij) / Support(i). Также необходимо задать какие-то ограничения максимальной достоверности, иначе правила окажутся слишком очевидными, и алгоритм будет не таким полезным. Мне кажется, можно зафиксировать этот параметр равным 60.

В ходе алгоритма мы поочередно просматриваем все возможные объединения объектов с некоторым ограничением. Сначала берем каждый объект сам по себе и считаем его поддержку, если она меньше заданной минимальной, отбрасываем его. Из оставшихся образовываем пары, считаем поддержку, отбрасываем. Затем тройки, четверки и так до тех пор, пока число рассматриваемых объектов в объединении не превзойдёт максимальное число характеристик для некоторого студента. Будет образовано подобное дерево при помощи поиска в глубину.



Входные данные

{item : TID set}, item – название объекта, TID set – множество номеров транзакций, содержащих данный объект item. В нашем случае объектами являются характеристики, относящиеся к студентам, а транзакциями, соответственно, наборы этих характеристик для каждого студента. Номер транзакции – номер студента в базе данных. Например, пусть исходная база данных содержит несколько студентов и набор характеристик для каждого. Характеристикой может быть оценка за дисциплину, затраченное на её прохождение число часов и т.д. Тогда для успешной работы алгоритма Eclat необходимо преобразовать каждую характеристику в объект item (Например, 10 по матанализу и 8 по программированию это два разных объекта-item). Номером транзакций будет являться номер студента, который имеет данную оценку среди своих характеристик.

Данные удобно будет хранить в виде дерева префиксов, а также отдельно словарь со значениями Support и Confidence для каждого ключа-узла из образованного дерева. Это обеспечит наглядный показатель того, какие отношения будут образовывать правила.

Выходные данные

Если выводить Confidence от введённых пользователем собственных характеристик, то это будет вероятность получения желаемой оценки. Иначе можно вывести вероятный результат работы нашего ассоциативного правила для заданных характеристик, тогда студент получит наиболее вероятный результат своей деятельности.

Использованные материалы

Ward hierarchical clustering

- https://www.researchgate.net/publication/51962445_Ward's_Hierarchical_Clustering_Method_Clustering_Criterion_andAgglomerative_Algorithm (Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm)
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
- https://scikit-learn.org/stable/modules/clustering#hierarchical-clustering
- https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/
- https://neerc.ifmo.ru/wiki/index.php?title=%D0%98%D0%B5%D1%80%D0%B0%D1%80%D1%85%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F

Equivalence Classification Algorithm (Eclat)

- http://oaji.net/articles/2015/1948-1434984559.pdf (International Journal of Computer Techniques -– Volume 2 Issue 3, May – June 2015 FREQUENT ITEMSET MINING USING ECLAT WITH RELATIVE PROFIT AND PRICE)
- https://www.geeksforgeeks.org/ml-eclat-algorithm/
- https://habr.com/ru/company/ods/blog/353502/