УДК 004.021
DOI: 10.18101/978-5-9793-1397-9-101-106

CI-CLOPE: РАСШИРЕНИЕ АЛГОРИТМА CLOPE

© **Шабанов Владимир Юрьевич**
аспирант,
Новосибирский государственный университет
Россия, 630090, г. Новосибирск, ул. Пирогова, 1
E-mail: vavanrrrage@gmail.com

© **Михайлов Александр Сергеевич**
аспирант,
Новосибирский государственный университет
Россия, 630090, г. Новосибирск, ул. Пирогова, 1
E-mail: vavanrrrage@gmail.com

В настоящее время идет активное развитие такой отрасли информационных технологий, как машинное обучение. Кластеризация является одним из основных его пунктов. Особое внимание уделяется транзакционным данным. В данной статье описывается алгоритм кластеризации CLOPE для работы с транзакционными данными, представлено его расширение для заранее заданного набора кластеров и продемонстрированы результаты.

Ключевые слова: машинное обучение; анализ данных; категориальные данные; транзакционные данные; CLOPE; кластеризация; LargeItem.

Введение

Кластеризация является одной из самых важных техник машинного обучения. Ее суть состоит в группировке объектов между собой таким образом, чтоб объекты одной группы были максимально схожими между собой, а объекты разных групп максимально отличались [1]. Наличие шума в данных делает обнаружение кластеров более сложным процессом. Хотя люди отлично справляются с нахождением кластеров в двух и, возможно, даже трех измерениях, необходимы автоматические алгоритмы для данных больших размеров.

В последнее время возрос интерес к категориальным данным, т. е. данным с неколичественными характеристиками. Такие данные отличаются тем, что сравнивать их характеристики можно только в значении равенства [2].

Набор транзакционных данных состоит из нескольких транзакций, каждая из которых содержит различное количество элементов. Цель алгоритмов транзакционной кластеризации — разделить исходную коллекцию транзакций на набор непустых кластеров так, чтобы каждый кластер содержал однородное подмножество транзакции.

Под однородными транзакциями понимается набор транзакций, которые имеют много общих элементов. Другими словами, кластер — это набор

транзакций, в которых определенные элементы происходят с большей частотой, чем где-либо еще. Транзакции, сгруппированные в одних и тех же кластерах, имеют высокую степень перекрытия.

Быстрая и эффективная кластеризация транзакционных баз данных является довольно трудоемким процессом ввиду их высокой размерности и большого объема. Подходы, основанные на вычислении попарного расстояния между объектами (например, k-means) хорошо себя показывают на данных низкой размерности, но их эффективность часто бывает неудовлетворительной на номинальных данных.

Кластеризация номинальных данных

Иерархические методы кластеризации, такие как ROCK, seed, достаточно эффективны, но имеют плохую производительность на большом наборе данных. Одним из самых популярных алгоритмов для работы с большим количеством категориальных данных является LargeItem, который работает итеративно, оптимизируя глобальный критерий [3]. Оценочные функции могут быть определены как глобально, так и локально. Локальные функции широко распространены в кластеризации количественных данных. Однако при обработке больших объемов данных применение локального критерия может привести к высокой стоимости вычислений. Применение глобального критерия позволяет алгоритму работать быстрее, и поэтому этот алгоритм хорошо подходит для кластеризации больших массивов номинальных данных.

Алгоритм CLOPE

Алгоритм CLOPE использует такую глобальную оценочную функцию, которая пытается увеличить внутрикластерное перекрытие элементов транзакций путем увеличения отношения высоты и ширины кластерной гистограммы [4]. В данном алгоритме введен параметр, означающий «тесность» кластера, в зависимости от которого может варьироваться число полученных кластеров в разбиении. По сравнению с предыдущим и другими иерархическими алгоритмами, CLOPE показывает более высокую производительность и качество кластеризации. Достигается это за счет внедрения более простой, но при этом эффективной глобальной метрики для кластеризации категориальных данных. Опишем работу алгоритма подробнее.

Пусть D — транзакционная база данных, представляющая собой набор транзакций $\{t_1, \ldots, t_n\}$, где каждая транзакция состоит из какого-то количества m элементов $\{i_1, \ldots, i_m\}$, а результатом кластеризации является набор элементов $\{C_1, \ldots, C_k\}$, объединение которых совпадает с объединением всех транзакций без пересечений. Для каждого кластера C вводятся и считаются следующие характеристики:

D(C) — набор уникальных объектов;

Occ (i, C) — количество появлений в кластере C объекта i;

$S(C) = \sum Occ(i, C)$ — размер (площадь) кластера, равный сумме элементов всех транзакций;

W(C) = |D(C)| — ширина кластера;

H(C) = S(C) / W(C) — высота кластера.

Несложно понять, что чем больше высота кластера, тем более похожи транзакции, принадлежащие кластеру. Однако одной лишь высоты недостаточно для определения оценочной функции.

Ввод градиента $G(C) = H(C) / W(C) = S(C) / W(C)^2$ может дать более высокое качество кластеризации. Также необходимо принимать во внимание форму кластеров и количество транзакций к ним. Таким образом, вводится функция стоимости:

$$Profit(C) = \frac{\sum_{i=1}^{k} G(C_i) * |C_i|}{\sum_{i=1}^{k} |C_i|} = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^2} * |C_i|}{\sum_{i=1}^{k} |C_i|}.$$

Обобщение этой функции выглядит следующим образом:

$$Profit(C) = \frac{\sum_{i=1}^{k} G(C_i) * |C_i|}{\sum_{i=1}^{k} |C_i|} = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} * |C_i|}{\sum_{i=1}^{k} |C_i|},$$

где вводится параметр $r > 0$, называемый коэффициентом отталкивания и используемый для вариативности уровня внутрикластерной схожести. И чем больше значение данного параметра, тем большее количество кластеров может быть сгенерировано.

Таким образом, применение алгоритма CLOPE сводится к решению следующей задачи:

$$Profit(C, r) \rightarrow max$$

Для сравнения с другими алгоритмами CLOPE был опробован на знаменитой задаче о грибах, где уверенно превзошел такие алгоритмы, как LargeItem и ROCK.

Расширение CLOPE: алгоритм CI-CLOPE

На практике нередко встречаются задачи, когда исследователь заранее знает точное число групп, на которые необходимо разбить исходное множество. К ним можно отнести многочисленные социологические исследования [5].

Алгоритм CLOPE, как известно, не позволяет исследователю задать предпочтительное число кластеров. Он строит такое число кластеров, которое обеспечивает максимизацию глобальной функции стоимости.

Приведем фрагмент кода данного алгоритма:

```
/* Step 1 - Initialization */
while not end of the database file
   read the next transaction 〈t, unknown〉;
   put t in an existing cluster or a new cluster Ci
   that maximize profit;
   write 〈t, i〉 back to database;
/* Step 2 - Iteration */
repeat
   rewind the database file;
   moved = false;
```

```
    while not end of the database file
       read 〈t, i〉;
       move t to an existing cluster or new cluster Cj
          that maximize profit;
       if Ci ≠ Cj then write 〈t, j〉 ; moved = true;
 until not moved

 /* Step 3 - Count correction */
 k = cluster_count
 if (k > K)
    /* reduction */
    while (k > K)
    merge pair of clusters
    which will maximize profit
    k = k - 1
 else if (k < K)
    /* extension */
    while k < K
       take biggest cluster and
       split it with maximal profit
       k = k + 1
```

Основной целью при разработке расширения существующего алгоритма CLOPE являлось построение нового алгоритма, позволяющего задавать предпочтительное число кластеров. Таким образом, искомый алгоритм должен принимать на вход пару параметров <K, r>, где K — число кластеров, а r — коэффициент отталкивания. В настоящей работе предлагается расширение алгоритма CLOPE — CI-CLOPE (Concrete Integer CLOPE).

Работа алгоритма состоит из трех шагов: первые два — выполнение оригинального CLOPE, третий шаг осуществляет коррекцию числа кластеров. Пусть K — целевое количество кластеров, а k — количество кластеров, полученных в результате работы CLOPE. Тогда возможно три варианта:

- k = K. В этом случае работа алгоритма заканчивается.
- k > K. Число кластеров, полученных в результате первых двух шагов, больше целевого. Запускается процесс сокращения: среди множества кластеров ищется пара кластеров (C_n, C_m) такая, что при слиянии двух кластеров, значение глобальной функции стоимости является максимальным. Данная процедура производится до тех пор, пока не будет достигнуто целевое количество кластеров.
- k > K. Число кластеров, полученных в результате работы CLOPE, меньше целевого. Запускается процесс расширения: кластер с наибольшим числом элементов разбивается на два таким образом, чтобы достигалось максимальное значение функции стоимости. Этот шаг производится до тех пор, пока не будет достигнуто целевое количество кластеров.

Сходимость предлагаемого алгоритма очевидна по построению.

Заключение

В работе было представлено расширение алгоритма CLOPE. Зависимость от двух параметров позволяет исследователю более детально настроить алгоритм для решения поставленных задач. При этом, предлагаемый алгоритм обладает всеми известными плюсами оригинального CLOPE. В дальнейшие планы входит исследование зависимости результата кластеризации от порядка транзакций в исходной базе, а также усовершенствование и разработка новых подходов к разбиению и слиянию кластеров.

Литература

1. Duda R. O., Hart P. E., Stork D. G. Unsupervised Learning and Clustering // Pattern Classification. 2001. P. 517–601.

2. Gibson D., Kleinberg J., Raghavan P. Clustering categorical data: An Approach Based on Dynamical Systems // Databases. 1998. V. 1. P. 75.

3. Wang K., Xu C., Liu B. Clustering Transactions using Large Items // Proceedings of the Eighth International Conference on Information and Knowledge Management. 1999. P. 483–490.

4. Yang Y., Guan X., You J. CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data // Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002. P. 682–687.

5. Загоруйко Н. Г., Заславская Т. И. Применение методов распознавания образов в социологии. Новосибирск: Наука, 1968.

CI-CLOPE: CLOPE ALGORITHM EXTENSION

Vladimir Yu. Shabanov
Research Assistant,
Novosibirsk State University
1 Pirogova St., Novosibirsk 630090, Russia
E-mail: vavanrrrage@gmail.com

Aleksandr S. Mikhailov
Research Assistant,
Novosibirsk State University,
1 Pirogova St., Novosibirsk 630090, Russia
E-mail: vavanrrrage@gmail.com

In real time machine learning has rapid growth. Cluster analysis is one of the fundamental notion. Special case of clustering analysis is transactional data. In this paper there is a description of popular clustering algorithm CLOPE for transactional data. Extension of the algorithm for concrete number of clusters and results will be presented.

Keywords: machine learning; data mining; categorical data; transactional data; CLOPE; cluster analysis; LargeItem.