DOI: 10.17117/na.2016.12.02.142

<http://ucom.ru/doc/na.2016.12.02.142.pdf>

Поступила (Received): 24.12.2016

Турыгина В.Ф., Матвеевнина А.И. Метрические алгоритмы кластеризации

Turygina V.F., Matvevnina A.I.
Metric clustering algorithms

В современных научных публикациях и специализированных журналах зачастую встречается такой термин как «Data Mining». Data Mining – это новое направление в исследовании данных, которое представляет собой объединение большого числа математических инструментов анализа и современных достижений в области информационных технологий

Ключевые слова: метод, анализ данных, кластеризация

Data Mining – a new direction in research data, which is an amalgamation of a large number of mathematical tools of analysis and the latest achievements in the field of information technology

Key words: method, data analysis, clustering

Турыгина Виктория Федоровна

Старший преподаватель

Уральский федеральный университет им. первого Президента России Б.Н. Ельцина
г. Екатеринбург, ул. Мира, 19

Turygina Victoria Fedorovna

Senior Teacher

Ural federal university named the first President of Russia B.N. Yeltsin
Yekaterinburg, Mira st., 19

Матвеевнина Арина Ильинична

Студент

Уральский федеральный университет им. первого Президента России Б.Н. Ельцина
г. Екатеринбург, ул. Мира, 19

Matvevnina Arina Ilinichna

Student

Ural federal university named the first President of Russia B.N. Yeltsin
Yekaterinburg, Mira st., 19

В современных научных публикациях и специализированных журналах зачастую встречается такой термин как «Data Mining». Data Mining (на рус. «интеллектуальный анализ данных») – это новое направление в исследовании данных, которое представляет собой объединение большого числа математических инструментов анализа и современных достижений в области информационных технологий. Новая технология анализа данных гармонично объединяет в себе строгие математические законы и неформальные методы.

Data Mining позволяет обнаружить в данных ранее неизвестные закономерности, интерпретация которых будет практически полезна для принятия как тактических, так и стратегических решений в различных сферах деятельности. Впервые этот термин был введен Григорием Пятецким-Шапиро в 1989 году.

Data Mining использует алгоритмы, требующие большого количества сложных вычислений. И поэтому только благодаря росту производительности

процессоров новое направление получило широкое практическое применение в аналитике. Сегодня с помощью алгоритмов Data Mining решается целый ряд задач:

- классификация объектов;
- регрессия и прогнозирование;
- кластеризация;
- построение ассоциаций;
- последовательные шаблоны;
- анализ отклонений.

Data Mining предлагает большое количество различных методов для решения задачи кластеризации. Задача кластеризации данных заключается в сборе сведений об изучаемой выборке, после чего объекты выборки должны быть упорядочены в однородные группы. Проблема кластеризации отлична от задачи классификации исходных данных, поскольку при решении задачи кластеризации количество кластеров, на которые должна быть поделена исследуемая выборка, не известно и определяется лишь в процессе реализации алгоритма.

Задачу кластеризации позволяет решить метрический алгоритм кластеризации, оценивающий сходство элементов выборки между собой и образующий кластеры из тех элементов выборки, которые наибольшим образом схожи между собой. При этом также важно, чтобы объекты разных кластеров были как можно более отличны друг от друга.

Проблема кластеризации данных в общем виде решается с помощью последовательного выполнения таких работ, как: определение выборки для последующей кластеризации; отбор набора переменных, с помощью которых будут оцениваться наблюдаемые значения выборки; расчет значений мер сходства объектов выборки; применение алгоритма кластерного анализа для разбиения выборки на кластеры, содержащие объекты со сходными признаками; интерпретация результатов проведенного анализа.

Наиболее популярным методом, использующимся для решения задачи кластеризации исходных данных, является метод k-средних. Суть метода заключается в разделении имеющихся n эмпирических наблюдений, принадлежащих пространству R^m , на k кластеров таким образом, чтобы каждое наблюдение было отнесено к тому кластеру, центроид которого расположен к нему наиболее близко. Центроид – это представитель кластера. При этом расстояние между наблюдаемым значением исследуемой переменной и центроидом определяется с помощью меры Евклида (Евклидово расстояние):

$$\rho(x, y) = \sqrt{\sum_{i=1}^n (x_i - c_i)^2}, \quad (1)$$

где $\rho(x, y)$ – величина Евклидова расстояния;

$x = (x_1, x_2, \dots, x_n)$ – точка в пространстве R^m ;

$c = (c_1, c_2, \dots, c_k)$ – центроиды для каждого из k кластеров.

Пусть имеются эмпирические значения исследуемого параметра: $(x_1, x_2, x_3, \dots, x_n)$, $x_i \in R^m$. В результате применения метода k-средних выборка из n наблюдений делится на k кластеров (групп), причем $k \leq n$, таким образом, чтобы минимизировать сумму квадратов отклонений точек каждого кластера от центроидов кластеров, которым эти точки принадлежат.

Целевая функция метода k-средних имеет вид:

$$\sum_{i=1}^k (x_i - c_j)^2 \rightarrow \min, \quad (2)$$

где x_i – точка в пространстве R^m ;

c_j – центроид j -го кластера.

Как только определен метод, с помощью которого будет оцениваться степень близости объектов и центроидов кластеров, дальнейшее решение задачи сводится к нахождению центроидов кластеров.

Число групп, на которые будет разделена исследуемая выборка, определяется самим исследователем. Выбор точек, которые будут представлять каждый кластер в отдельности и с помощью которых будет осуществляться дальнейшее разбиение всей выборки, является много итерационным процессом.

Первоначально центроиды для каждого из кластеров выбираются либо случайным образом, либо так, чтобы расстояния между кластерами были максимальны. Затем вычисляется расстояние от каждого отдельного наблюдения до всех центроидов кластеров, после чего наблюдения будут отнесены к тому кластеру, центроид которого расположен наиболее близко. Причем каждое наблюдение может принадлежать лишь одной группе, даже если расстояния от этой точки до центроидов нескольких кластеров одинаково. После того, как определены центроиды кластеров, и выборка разделена на k кластеров, центроид каждого кластера перемещается к центру (среднему арифметическому) всех точек, принадлежащих кластеру. Этот шаг является обоснованным, поскольку лучшим образом каждый кластер будет описан с помощью точки, сумма расстояний от которой до каждого представителя кластера минимально. Именно по этой причине метод получил название « k -средних», поскольку найденные центроиды фактически представляют собой среднее значение каждого кластера.

Каждая следующая итерация алгоритма осуществляет поиск наиболее лучшего центроида для каждого кластера. Работа алгоритма завершается, когда значения кластеров при каждой следующей итерации не изменяются.

В результате применения метода k -средних все возможные значения зависимой переменной будут представлять собой конечное счетное множество, численно равное количеству выделенных кластеров (бинарная классификация, классификация по временам года и т.д.). Таким образом, зависимая переменная будет объяснена с помощью качественных переменных – кластеров, принимающих значения 0 и 1 соответственно в каждом случае. Качественные переменные, множество допустимых значений которых конечно и счётно, называются фиктивными переменными (дамми-переменными). В данном случае фиктивная переменная служит индикатором принадлежности наблюдаемого значения зависимой переменной к определенному кластеру.

Главный недостаток – трудоемкость вычислений, но при современных вычислительных возможностях и навыках программирования это проблема является не актуальной.