Разработка ансамбля алгоритмов кластеризации на основе изменяющихся метрик расстояний

© П. В. Бочкарёв © В. С. Киреев

Национальный исследовательский ядерный университет «МИФИ»,
Москва, Российская Федерация
pvbochkarev@mephi.ru vskireev@mephi.ru

Аннотация

В настоящее время происходит активное накопление данных большого объёма в различных информационных средах, таких как социальные, корпоративные, научные и другие. Интенсивное использование больших данных в различных областях стимулирует повышенный интерес исследователей к развитию методов и средств обработки и анализа массивных данных огромных объёмов и значительного многообразия. Одним из перспективных направлений в аналитике интенсивных данных является кластерный анализ, который позволяет решить такие задачи как, сокращение размерности исходного набора данных, выявление паттернов и т.д. В данной статье авторами предлагается ансамбль алгоритмов кластеризации, состоящий из базовых алгоритмов K-means, отличающихся по одному параметру - метрике расстояния между объектами. Для оценки работы разработанного ансамбля использованы открытые данные архива UCI.

1 Введение

Данные большого объёма (BigData), используются в различных процессах, таких как извлечение информации из веб-ресурсов, выявления общих закономерностей в областях с интенсивным использованием данных и т.д. Эти данные необходимо структурировать, классифицировать, подвергать тщательному анализу. В этом случае кластерный анализ является основой многих научных исследований [14]. Кластеризация (от англ. cluster – скопление), это сегментация через выделение определённых объединений однородных элементов, которые рассматриваются как самостоятельные единицы, обладающие определёнными свойствами [15].

В результате процедуры кластеризации образуются «кластеры», то есть группы очень похожих объектов [16].

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Под критерием качества кластеризации обычно понимается некоторый функционал, зависящий от разброса объектов внутри группы и расстояний между ними [9].

Кластеризация отличается от классификации тем, что изначально неизвестны ни количество, ни свойства классов (кластеров). К особенностям кластеризации можно отнести следующее:

- возможность определения заранее неизвестного класса объектов по начальным характеристикам;
- возможность обработки сколь угодно большого количества объектов в достаточно короткие сроки.

Устойчивость решений в задачах кластеризации может быть повышена благодаря формированию ансамбля алгоритмов [13] и построению с его помощью коллективного решения на основе мнений участников ансамбля, где под мнением алгоритма подразумевается его вариант разбиения данных на кластеры.

Данные свойства кластерного анализа особо актуальны при работе в областях с интенсивным использованием данных, когда предметная область слабо формализована, например, для анализа текстовых документов, изображений и т.д.

Основное внимание в данной работе уделяется построению ансамбля алгоритмов кластеризации на основе изменяющихся метрик расстояний для анализа данных большого объема.

2 Современные подходы к решению проблемы

Выбор метода кластеризации зависит от количества данных и от того, требуется ли обрабатывать и анализировать несколько типов данных одновременно [6, 10].

На практике чаще всего используются гибридные подходы, в которых шлифование кластеров выполняется методом K-средних (см. форм.1), а начальное разбиение – одним из более универсальных и мощных методов.

$$V = \sum_i^k \sum_{x_j \epsilon S_i} \left(x_j - \mu_i\right)^2, \qquad (1)$$

где k – число кластеров, S_i – полученные кластеры, $i=1, 2,\ldots, k$ и μ_i – центры масс векторов.

Данные о сравнении алгоритмов представлены в таб.1 [7].

Таблица 1 Сравнительная таблица алгоритмов

Алгоритм кластеризации	**Входные данные**	**Результаты**
иерархическая	число кластеров или порог расстояния для усечения иерархии	бинарное древо кластеров
K-средних	число кластеров	центры кластеров
C-средних	число кластеров, степень нечеткости	центры кластеров, матрица принадлежности
выделение связных компонент	порог расстояния R	древовидная структура кластеров

Для определения расстояния между объектами в кластерном анализе используются различные метрики расстояний между объектами *x* и *x'*. Наиболее востребованными в кластерном анализе являются следующие метрики:

1. Евклидово расстояние
$$p(x,x') = \sqrt{\sum_i^n (x_i - x_i')^2}; \quad (2)$$
2. Манхэттенское расстояние
$$p(x,x') = \sum_i^n |x_i - x_i'|; \quad (3)$$
3. Расстояние Чебышева
$$p(x,x') = max(|x_i - x_i'|); \quad (4)$$
4. Коэффициент Жаккара
$$K(x,x') = \frac{\sum_i^n x_i x_i'}{\sum_i^n x_i^2 + \sum_i^n x_i'^2 - \sum_i^n x_i x_i'}; \quad (5)$$
5. Динамическая трансформация временной шкалы (dynamic time wrapping, DTW)
$$DTW(x,x') = \frac{\min\{\sum_{k=1}^K d(\omega_k)\}}{K}, \quad (6)$$
где *K* – длина пути между *x* и *x'*, который вычисляется по специальной матрице трансформаций [11].

Выбор метрики существенно влияет на качество кластеризации.

В настоящее время в кластерном анализе проявляется тенденция к применению коллективных методов [1]. Ранее было отмечено, что алгоритмы кластерного анализа не являются универсальными: каждый алгоритм имеет свою особую область применения (таблица 1). В том случае, если рассматриваемая область содержит различные типы данных, для выделения кластеров необходимо применять не один определённый алгоритм, а набор различных алгоритмов.

Ансамблевый (коллективный) подход позволяет снизить зависимость конечного решения от выбранных параметров исходных алгоритмов и получить более устойчивое решение даже при большом количестве шумов и выбросов в данных [9].

Существуют следующие основные методики получения ансамбля алгоритмов (см. Рис. 1)[8]:

1. нахождение консенсусного разбиения, т.е. согласованного разбиения при имеющихся нескольких решениях, оптимального по некоторому критерию;
2. вычисление согласованной матрицы сходства/различий (co-occurrence matrix).

Рисунок 1 Ансамбли алгоритмов кластеризации

При формировании окончательного решения используются результаты, полученные различными алгоритмами, либо одним алгоритмом с различными значениями параметров, по разным подсистемам переменных и т.д. В настоящее время ансамблевый подход является одним из наиболее перспективных направлений в кластерном анализе.

Примерами использования ансамбля алгоритмов кластеризации могут служить следующие. Созданный на основе непараметрического алгоритма MeanSC, ансамбль позволил улучшить показатели кластеризации многоканальных изображений [13]. А также, используя ансамбль алгоритмов кластеризации на основе K-средних и алгоритма SVM (Support Vector Machines), удалось повысить точность обнаружения сердечных аномалий, что позволило сократить время установления диагноза [2].

Таким образом, применяя ансамбль с различными наборами алгоритмов, в соответствии с их преимуществами и особенностями, можно создать наиболее подходящую схему кластеризации для определённой предметной области. Ранее также указывалось, что важным фактором, влияющим на результат кластеризации, является выбор конкретной метрики расстояний между объектами. Объединяя эти два подхода, можно существенно повысить эффективность кластерного анализа.

3 Предлагаемый подход

3.1 Ансамбль алгоритмов кластеризации

Предлагаемый авторами ансамбль алгоритмов представляет собой сочетание последовательных алгоритмов K-средних, каждый из которых предлагает свое разбиение, и иерархического агломеративного алгоритма, объединяющего

полученные решения с помощью особого механизма. В отличие от ансамбля, использующего алгоритм MeansSC [13], предложенный ансамбль опирается на результаты предварительного исследования исходных данных, которые представляют собой небольшой набор размеченных экспертами объектов. Минимально необходимый процент объема исходной выборки, гарантирующий заданную точность, подлежит дальнейшему изучению. Для определенности, в данной работе используется 0.5%, что в случае увеличения объема данных, очевидно, должно подлежать пересмотру.

На первом шаге каждый алгоритм K-средних, разбивает данные на кластеры, используя свою метрику расстояния. Затем, рассчитывается точность и вес мнения алгоритма в ансамбле по формуле 7:

$$\omega_l = \frac{Acc_l}{\sum_{l=1}^{L} Acc_l}, \quad (7)$$

где Acc_l – точность алгоритма l, т.е. отношение количество правильно кластеризованных объектов к объему всей выборки, а L – количество алгоритмов в ансамбле.

Для каждого полученного разбиения составляется предварительная бинарная матрица различий размера n x n, где n-количество объектов, необходимая для определения, занесены ли объекты разбиения в один класс. Затем рассчитывается согласованная матрица различий, каждый элемент которой представляет собой взвешенную (с использованием веса из формулы 7) сумму элементов предварительных матриц. Полученная матрица используется в качестве входных данных для алгоритма иерархической агломеративной кластеризации. Затем с помощью обычных приемов, таких как определение скачка расстояния агломерации, можно выбрать наиболее подходящее кластерное решение. Процедура создания ансамбля алгоритмов представлена на рисунке 2.

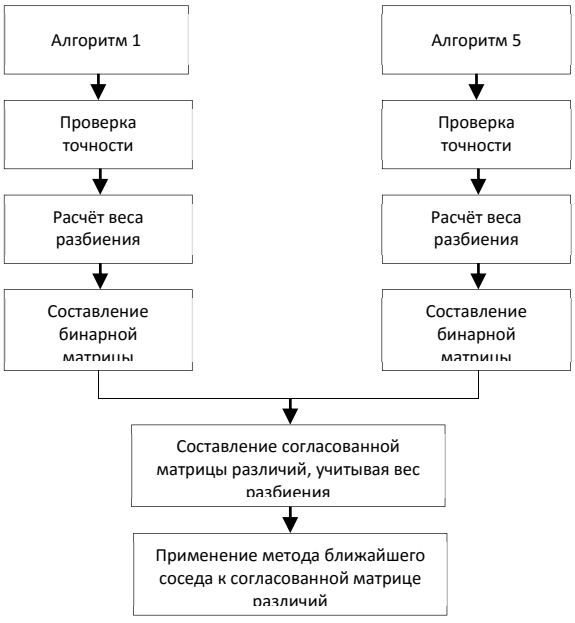

Рисунок 2 Ансамбль алгоритмов кластеризации

3.2 Алгоритмы кластеризации

В данном ансамбле алгоритмов кластеризации были использованы пять K-средних (см. форм.1), как один из наиболее востребованных алгоритмов кластеризации больших данных [12]. Для данных алгоритмов были использованы такие метрики, как:

- Евклидово расстояние (см. форм. 2)
- Манхэттенское расстояние (см. форм. 3)
- Расстояние Чебышева (см. форм. 4)
- Коэффициент Жаккара (см. форм. 5)
- DTW расстояние (см. форм. 6)

3.3 Наилучшее разбиение на кластеры

Для получения наилучшего разбиения на кластеры необходимо, как было упомянуто выше, составить бинарную матрицу сходства\различий на каждое L разбиение в ансамбле:

$$H_i = \{h_i(i,j)\}, \quad (8)$$

где $h_i(i, j)$ равен нулю, если элемент i и элемент j попали в один кластер, и 1 если нет.

Следующим шагом в составлении ансамбля алгоритмов кластеризации является составление согласованной матрицы бинарных разбиений.

$$H^* = \{h^*(i,j)\}, \quad (9)$$

$$h^*(i,j) = \sum_{l=1}^{L} w_l h_l(i,j), \quad (10)$$

где w_l – вес алгоритма.

Для формирования наилучшего разбиения по согласованной матрице был выбран алгоритм ближайшего соседа.

4 Валидация предлагаемого ансамбля

Для тестирования и оценки ансамбля алгоритмов кластеризации использовалось программное средство RapidMiner [3]. С помощью RapidMiner можно решать, как исследовательские (модельные), так и прикладные (реальные) задачи интеллектуального анализа данных, включая анализ текста, анализ мультимедиа, анализ потоков данных, что подходит для тестирования ансамбля алгоритмов кластеризации. В качестве данных для кластеризации использовались открытые данные с web-сайта UCI [4]. Данный пример содержит информацию о платежах клиентов с помощью пластиковых карт, всего 30 тыс. записей и 24 атрибута. Данные были размечены экспертным способом на 2 кластера, и эти результаты были взяты в качестве корректного решения.

Ниже представлены элементы схемы эксперимента, разработанной в RapidMiner. Для снижения размерности исходных данных был выбран метод главных компонентов (Principal component analysis, PCA) (см. Рис. 3). В качестве критерия выбора количества компонент был выбран критерий Кайзера (собственное значение компоненты больше единицы).

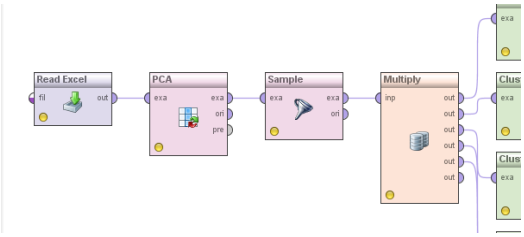

Рисунок 3 Снижение размерности данных

На следующем шаге была выявлена точность каждого алгоритма путём сравнения полученного разбиения на два кластера каждым алгоритмом с кластерами, размеченными экспертным способом. После получения значения точности каждого алгоритма, по формуле 7 был рассчитан вес мнения алгоритма (см. Рис. 4). Так, из графика видно, что наибольшим весом обладает алгоритм, использовавший расстояние Чебышева, а наименьшим весом - алгоритм с метрикой Жаккара.

Рисунок 4 Диаграмма веса алгоритмов

Далее была проведена кластеризация данных каждым алгоритмом. На следующем шаге, используя возможности RapidMiner, по формуле 8 были получены бинарные матрицы разбиения

На основе полученных результатов можно определить значение индекса качества группировки (вес разбиения). Используя вес каждого разбиения и сумму значений бинарных матриц сходства\различий (см. форм. 9), была составлена согласованная матрица различий для ансамбля алгоритмов кластеризации (см. форм. 10).

Применяя алгоритм ближайшего соседа к рассчитанной матрице, с помощью возможностей Rapidminer, было определено наилучшее разбиение.

На рисунке 5 представлена часть результатов работы алгоритма – дендрограммы, полученной на последнем этапе его работы.

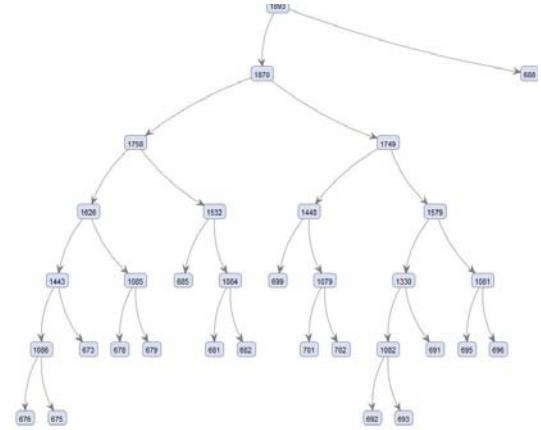

Рисунок 5 Работа алгоритма иерархической кластеризации

В результате применения предложенного подхода было получено окончательное решение, состоящее из двух кластеров, характеризующих поведение клиентов при осуществлении платежей, обладающее достаточно высокой точностью, согласующееся с экспертным мнением (см. Рис. 6). Из рисунка видно, что точность предложенного ансамбля превышает точность стандартного алгоритма К-средних, с различными метриками.

Рисунок 6 Сравнение точности алгоритмов

5 Заключение

Задача интеллектуального анализа и обработки Больших Данных последние несколько лет является предметом изучения множества специалистов и важной составляющей этого анализа указывается кластеризация этих данных, позволяющая приблизиться к решению проблемы трех V (объема данных для хранения - Volume, скорости обработки - Velocity и разнообразия исходных типов данных - Variety) [5]. Таким образом, кластерный анализ становится одним из ключевых в сферах обработки интенсивных данных, так как это один из

эффективных методов, который существует на сегодняшний день. Применяя ансамбль алгоритмов кластеризации, можно повысить достоверность разбиения данных на группы. Существенным является то, что данный метод может применяться в различных областях. Рассмотренный в данной статье ансамбль алгоритмов кластеризации нивелирует недостатки метрик расстояний для алгоритмов K-средних, тем самым повышая достоверность разбиения. Дальнейшее исследование предложенного ансамбля алгоритмов на основе меняющейся метрики расстояний планируется в рамках гранта РФФИ № 15-07-08742.

Литература

[1] J. Ghosh, A. Acharya Cluster ensembles. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. V. 1(4). P.305−315.

[2] Kausar Noreen;Abdullah Azween; Samir Brahim Belhaouari;Palaniappan Sellapan;AlGhamdi Bandar Saeed;Dey Nilanjan. Ensemble Clustering Algorithm with Supervised Classification of Clinical Data for Early Diagnosis of Coronary Artery Disease.// Journal of Medical Imaging and Health Informatics, V. 6, Number 1, February 2016, P. 78-87.

[3] Predictive Analytics Platform | RapidMiner https://rapidminer.com/

[4] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (26.06.2016)

[5] А. К. Горшенин, С.Я. Шоргин. Разработка информационной технологии интеллектуального анализа больших данных // Современные проблемы прикладной математики, информатики, автоматизации, управления: Материалы 3-го Международного научно-технического семинара (Севастополь, 9–13 сентября 2013). –М.:ИПИ РАН, 2013. С. 104–114.

[6] А.П. Кулаичев. Методы и средства комплексного анализа данных. – М.: Форум — Инфра-М, 2006. — 512 с.

[7] Б.О. Лялька, Антонова-Рафи Ю.В. Оценка эффективности кластеризационных алгоритмов. // Научные труды SWORLD, 2015, №2 (39), С. 25-29.

[8] В.Б. Бериков. Классификация данных с применением коллектива алгоритмов кластерного анализа // Знания-Онтологии-Теории (ЗОНТ-2015), 2015, С. 29-38

[9] В. Б. Бериков. Коллектив алгоритмов с весами в кластерном анализе разнородных данных // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика, 2013, №2 (23), с 22-31.

[10] В.С. Киреев. Оценка результатов кластеризации при использовании различных критериев качества// Программные продукты и системы, 2009, №3, С. 36-39.

[11] Д.Е. Мозохин, В.А. Калягин. Сравнительный анализ алгоритмов кластеризации в сетях фондовых рынков // Алгоритмы, методы и системы обработки данных. 2015, 4(33), С. 73-90

[12] И. Демин. Концепция кластера в технологиях интеллектуального анализа данных// Риск: Ресурсы, Информация, Снабжение, Конкуренция, 2012, 1, С. 260-263.

[13] И. А. Пестунов, В. Б. Бериков, Ю. Н. Синявский. Сегментация многоспектральных изображений на основе ансамбля непараметрических алгоритмов кластеризации // Вестник СибГАУ, 2010, №5(31), С. 56-64.

[14] С. А. Суслов. Кластерный анализ: сущность, преимущества и недостатки // Вестник НГИЭИ. 2010. №1, С. 51-57.

[15] С.А. Батуркин, Е.Ю. Батуркина, В.А. Зименко, И.В. Сигинов. Статистические алгоритмы кластеризации данных в адаптивных обучающих системах // Вестник РГРТУ, 2010, № 1 (31), С. 82-85.

[16] С. Л. Подвальный, А. В. Плотников, А. М. Белянин. Сравнение алгоритмов кластерного анализа на случайном наборе данных // Вестник ВГТУ, 2012, Т.8 №5, С. 4-6.

Development of Ensemble of Clustering Algorithms Based on Varying Distances Metrics

Pyotr V. Bochkaryov, Vasiliy S. Kireev

Currently there is an active accumulation of big data in various information environments, such as social, corporate, scientific and other domains. Intensive use of big data in various fields stimulates the increased interest of researchers to the development of methods and means of processing and analyzing massive data volumes with significant variety. One of the promising areas in data intensive analytics is cluster analysis, which allows to solve such problems as: reducing the dimension of the original dataset, identifying patterns, etc. In this article, the authors propose an ensemble of clustering algorithms, consisting of the basic algorithm K-means, characterized by one parameter - the distance metric between objects. For the evaluation of performance of the designed ensemble the open data archive of UCI was used.