Бустинг биоинспирированных алгоритмов для решения задачи кластеризации

Ю. А. Кравченко[1], А. Н. Нацкевич[2], И. О. Курситыс[3]
Южный федеральный университет, Таганрог, Россия
[1] krav-jura@yandex.ru, [2]natskevich.a.n@gmail.com

***Аннотация*. В статье рассмотрена разработка модели бустинга для решения задачи кластеризации. Приведена постановка задачи. Представлен краткий литературный обзор существующих алгоритмов бустинга, отмечены их достоинства и недостатки. В качестве примера решения задачи кластеризации данных представляется новая модель решения задач оптимизации, базирующаяся на использовании взвешенного множества слабых алгоритмов кластеризации и их бустинга основанного на идеях биоинспирированных алгоритмов. Предложен новый механизм решения задачи кластеризации. Эвристика предложенного алгоритма бустинга заключается в моделировании поведения колонии муравьев, что позволяет проводить взвешенную оценку качества результата слабых обучающихся алгоритмов для получения наиболее высокого качества решения задачи кластеризации. Проведенные исследования показали, что решения, полученные при помощи использования подхода бустинга алгоритмов, позволяют получать решения, не уступающие или превосходящие по качеству решения, полученные современными алгоритмами.**

Ключевые слова: бустинг; кластеризация; обучение с учителем; обучение без учителя; машинное обучения; биоинспирированные алгоритмы

I. Введение

Одной из наиболее сильно выраженных тенденций развития общества в настоящий момент является постоянный рост данных [1]. В качестве примера можно привести статистику компании IBM, согласно которой 2.5 эксабайта данных генерируется каждый год [2]. Такое большое количество данных делает затруднительным возможность их обработки имеющимися методами.

Данная проблема обосновывает актуальность создания новых масштабируемых алгоритмов анализа данных. Одним их наиболее часто используемых методов анализа данных является кластеризация, что обосновывается необходимостью деления огромного количество постоянно растущего объема данных на кластеры [1, 3] для последующего упрощения их обработки с целью выделения информации. Изначально имеется некоторое число объектов и число кластеров. Число кластеров может задаваться заранее или определяться алгоритмом. Любой объект может быть отнесен к любому кластеру. Основная цель кластеризации – разделение данных на группы (кластеры), состоящие из наиболее идентичных элементов. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятие решений, применяя к каждому кластеру свой метод анализа.

Кластеризация, рассматриваемая как самый важный и перспективный в плане изучения подход неконтролируемого обучения [3]. Также стоит отметить, что задача кластеризации данных относится к классу NP-полных задач, создание эффективных методов решения этой задачи является актуальной проблемой. Для решения данной задачи было разработано большое количество алгоритмов, которые отличаются друг от друга сложностью, временными затратами и эксплуатационными свойствами.

Многие методы кластеризации такими учеными, как Mayr A, Binder H, Gefeller O, Schmid M. Они провели классификацию этих методов и выделили основные две группы: классические и современные методы. Их анализ представлен в работе [5].

Одними из достаточно эффективных методов решения задачи кластеризации являются методы ядра (kernel method). Детальная информация об этих алгоритмах приведена в [6, 7]. Одни из наиболее часто используемых методов – Approximate kernel k-means [8]. Основная идея этого алгоритма – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Идея алгоритмов кластеризации, базирующихся на ярде – использование метода опорных векторов. Такой подход позволяет частично уменьшить влияние шума на результат кластеризации, но, как показывают эксперименты [5], время исполнения алгоритма по сравнению с традиционным k-means существенно повышается, что делает алгоритм менее применимым к данным большой размерности.

Также достаточно эффективными являются Алгоритмы, основанные на использовании Ансамблей. Одни из представителей – алгоритмы, базирующиеся на использовании генетического подхода и алгоритмы, базирующиеся на применении теории нечетких множеств [11]. Основная идея данных алгоритмов заключается в генерации набора исходных результатов кластеризации по определенному методу. Итоговый результат кластеризации получается путем интеграции исходных результатов кластеризации различными алгоритмами. Преимущество

Работа выполнена при финансовой поддержке РФФИ, проект №17-07-00446

такого подхода заключается в возможности распараллеливания используемых алгоритмов. Среди минусов можно выделить недостаточное понимание разницы между первичными результатами кластеризации. Также можно отметить сложность разработки общей целевой функции (concensus function) [5].

Среди разработанных методов стоит выделить бустинг, что обусловлено его достаточно активным развитием. Бустинг – это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов [4].

Как показывают аналитические обзоры, выполненные такими учеными, как Dongkuan XuYingjie Tian [1] и Ka-Chun Wong [3], не все современные разработанные алгоритмы бустинга способны дать оптимальное решение задач сферы машинного обучения при приемлемых временных затратах. Таким образом, проблема разработки алгоритма, сочетающего в себе полиномиальную сложность решения проблемы при приемлемых временных затратах является актуальной.

Также стоит отметить, что алгоритмы бустинга изначально разрабатывались для решения задач обучения с учителем, например, задачи классификации и для решения задачи кластеризации потребуется некоторая адаптация данного алгоритма.

II. Основы бустинга и различия в его приминении для решения задач обучения с учителем и обучения без учителя

Наиболее широкое распространение метод бустинга получил в сфере машинного обучения для решения задач обучения с учителем таких, как задача классификации. Основная идея решения подобных задач – обучение определенного алгоритма на уже классифицированных (labeled) данных с целью создания достоверных прогнозов для неклассифицированных данных [5]. Обучение с учителем – частная дисциплина, входящая в состав машинного обучения, которая также включает в себя обучение без учителя, которое базируется исключительно на анализе неклассифицированных данных. Также на основе этих двух подходов базируется полу-контролируемое обучение с учителем, которое включает как элементы из обоих дисциплин [9].

В случае решения задач обучения с учителем алгоритмы, как правило, получают некую обобщающую функцию *h(.)*, которая содержит решение задачи классификации. При этом основная цель решения задачи классификации – провести категоризацию объектов в предопределенный набор классов.

Также отметим, что в самом общем случае выходные данные решения задачи классификации представлены в качестве *Y*, содержащей информацию о двух классах, которые кодированы, как *{-1, 1}*.

При этом основная задача машины – обучиться на наборе тренировочных данных $(y_1, x_1),\ldots,(y_n, x_n)$, которые уже классифицированы для последующего прогнозирования классификации новых объектов x_{new}. При этом прогнозы принадлежности для данных $(x_1,...,x_n)$ являются реализациями из *X*, а *n* – размер набора тренировочных данных. Задача машины – разработать правило прогнозирования *h(.)* для корректной классификации новых поступающих данных.

$$(y_1, x_1),\ldots,(y_n, x_n) - (\text{sup}ervised\_learning) \rightarrow h(x_{new}) = y_{new}$$

В случае решения задач обучения без учителя, набор тренировочных данных отсутствует, а в качестве некой обобщающей функции выступает целевая функция оценки качества полученного решения. При этом основная задача машины – разработать изначальное правило распределения множества объектов на кластеры. Формула выглядит следующим образом:

$$(x_1, x_n) - (unsupervised\_learning) \rightarrow (y_1, x_1),\ldots,(y_n, x_n)$$

Например, в случае решения задачи кластеризации может оцениваться среднее внутрикластерное расстояние или среднее межкластерное расстояние.

Решением задачи кластеризации является множество $V' = \left\{ Y_i' \middle| l = 1,2,...,k \right\}$. Запланированным вариантом решения V' является разбиение множества объектов по множеству кластеров.

В качестве оценки решения V’ рассматривается целевая функция, имеющая следующий вид:

$$F = \frac{p^0}{p^i} \rightarrow \max, \quad (3)$$

где p^0 – среднее межкластерное расстояние, p^i – среднее внутрикластерное расстояние.

Рассмотрим подробнее механизм организации бустинга с помощью использования биоинспирированного алгоритма.

При этом формула для подсчета внутрикластерного расстояния имеет следующий вид:

$$\mathrm{P}^i = \frac{1}{X}\sum_{j=1}^{n}\sum_{i=1}^{n} p\left(x_i, c_j\right) \rightarrow \min. \quad (4)$$

где *p* – расстояние, которое вычисляется по формуле выбранной метрики, $x \in X$ – текущий элемент, $c \in C$ – центроид данного кластера, *k*– общее количество элементов, *l* – количество элементов в конкретном *j* кластере.

Среднее межкластерное расстояние описывает расстояние между объектами, входящими в состав различных кластеров и определяется по следующей формуле:

$$p^0 = \frac{1}{U}\sum_{u \in U} p\left(u_i, u\right) \rightarrow \max, \quad (5)$$

где p – расстояние с учетом выбранной метрики, u_i – рассматриваемый центроид, u – центроид, относительно которого вычисляется среднее межкластерное расстояние, n – общее количество кластеров.

III. Бустинг биоинспирированных алгоритмов для решения задачи кластеризации

В данной работе для решения задачи кластеризации используется модель бустинга биоинспирированных алгоритмов. Основная идея разработанного алгоритма заключается в использовании взвешенной версии определенного набора алгоритмов и множество вероятностей, определяющее вхождения каждого отдельного объекта в конкретный кластер. Данный набор алгоритмов используется многократно, что позволяет подобрать алгоритм, наиболее хорошо подходящий для кластеризации каждого конкретного набора данных. Схема работы алгоритма представлена на рисунке.

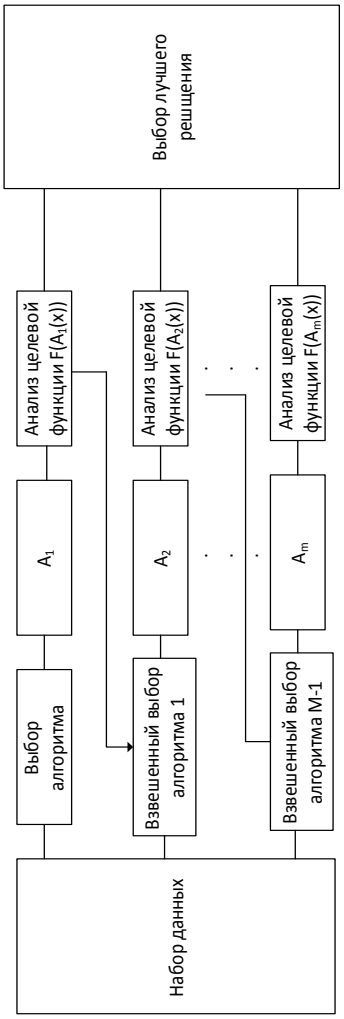

Fig. 1. Схема работы алгоритма бустинга

Рассмотрим работу алгоритма более подробно. На первом шаге происходит случайный выбор алгоритма, производящего кластеризацию, после чего происходит оценка его решения в соответствии с целевой функцией.

Входные параметры: Набор данных, состоящих из множества объектов, которые необходимо кластеризовать $X = \{x_i \mid i = 1, 2, ..., n\}$, где n – общие число объектов для кластеризации. Множество алгоритмов, которые могут быть применены для кластеризации объектов $A = \{a_j \mid j = 1, 2, ..., \mathrm{m}\}$, где m – общее число алгоритмов. Количество итераций T.

Стоит отметить одну из особенностей алгоритма: поскольку в наборе алгоритмов, которые могут быть применены в процессе бустинга используются биоинспирированные алгоритмы, появляется возможность использования множества, определяющего вероятность попадания каждого отдельного элемента в кластер, что может быть использовано, например, муравьиным алгоритмом.

В процессе работы алгоритма выбор каждого отдельного биоинспирированного алгоритма кластеризации осуществляется по следующей формуле:

$$f_j = \left(\alpha \frac{V_b}{V_t} + \beta u_i \right), \quad (6)$$

где V_b – результат вычисления целевой функции лучшего полученного решения, V_t – результат вычисления целевой функции текущим алгоритмом на текущей итерации, $u \in U$ – вероятность выбора i-го алгоритма из множества алгоритмов, которые могут быть использованы для кластеризации, α – коэффициент, определяющий значимость критерия улучшения показателя целевой функции, β – коэффициент, определяющий вес критерия вероятности выбора каждого конкретного алгоритма.

Опишем шаги алгоритма более подробно:

1. Инициализировать множества вероятностей выбора каждого отдельного алгоритма кластеризации $U = \{u_j \mid j = 1, 2, ..., \mathrm{m}\}$.
2. Пока $t < T$.
   a) запустить один из алгоритмов кластеризации. Выбор алгоритма осуществить по формуле (6);
   b) произвести кластеризацию элементов с помощью выбранного алгоритма, после чего оценить проведенную кластеризацию по формуле (3), внутрикластерное и межкластерное расстояние оценить по формулам (4), (5);
   c) для конкретного выбранного алгоритма обновить вероятность выбора;
   d) инкрементировать номер итерации $t \leftarrow t + 1$;
3. Закончить цикл.
4. Выбрать лучшее решение в соответствии с целевой функцией.

Рассмотрим более подробно некоторые из пунктов алгоритма. В пункте (a) приведена формула для оценки

качества работы определенного алгоритма кластеризации. Поскольку на первой итерации данных о предыдущей оценке целевой функции нет, данный критерий не учитывается, и формула выглядит следующим образом.

$$f_j = \left(\beta u_j\right) \tag{7}$$

IV. Заключение

В работе предлагается новая модель, базирующаяся на бустинге биоинспирированных алгоритмов для кластеризации данных. Предложены новые механизмы решения задач кластеризации. В отличие от канонической модели бустинга было добавлено множество взвешенное множество алгоритмов для улучшения качества кластеризации каждого конкретного набора данных. Такой подход является эффективным способом поиска рациональных решений для задач оптимизации. Алгоритм оптимизации может быть успешно применен для решения сложных комплексных задач оптимизации.

Источником усовершенствования может стать более корректный подбор параметров при осуществлении бустинга.

Другим источником усовершенствования может быть попытка использования множество вероятностей попадания каждого конкретного элемента из набора данных в каждый конкретный класс. Использования подобного множества становится возможным в случае использования в качестве алгоритма бустинга исключительно биоинспирированных алгоритмов кластеризации.

Также стоит отметить тот факт, что скорость работы алгоритмов может быть увеличена путем использования параллельной парадигмы программирования [10]. В качестве метода декомпозиции можно выбрать как декомпозицию по данным, так и декомпозицию по функциональным особенностям алгоритма.

Список литературы

[1] Ka-Chun Wong, "A Short Survey on Data Clustering Algorithms", IEEE Second International Conference on Soft Computing and Machine Intelligence, 2015.

[2] IBM Consumer products industry blog. Industry insights. Electronyc resource https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/

[3] Mayr A, Binder H, Gefeller O, Schmid M. The Evolution of Boosting Algorithms – From Machine Learning to Statistical Modelling. Methods Inf Med 2014; 53: 419–427.

[4] Бустинг. Особенности применения в области машинного обучения. Электронный ресурс URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%91%D1%83%D1%81%D1%82%D0%B8%D0%BD%D0%B3.

[5] Donkuan, X. Yingjie T. A comprehensive survey of clastering algorithms // Annals of Data Science, Volume2, Issue 2, 2015, pp 165–193.

[6] Müller K, Mika S, Rätsch G. An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw 12: 2001, pp 181–201.

[7] Filippone M, Camastra F, Masulli F, A survey of kernel and spectral methods for clustering. Pattern Recognit 41: 2008, pp 176–190.

[8] Radha C, Rong J, Timothy C.H, Anil K.J. Scalable Kernel Clustering: Approximate Kernel k-means. Computer Vision and Pattern Recognition, 2014.

[9] Praveena M & Jaiganesh V, “A Literature Review on Supervised Machine Learning Algorithms and Boosting Process”, International Journal of Computer Applications, Vol.169, No.8, (2017), pp. 32–35.

[10] Карпов В.Е. Введение в распараллеливание алгоритмов и программ. Компьютерные исследование и моделирование. 2010, т. 2. № 3. с. 231–272.

[11] Punera K, Ghosh J. Consensus-based ensembles of soft clusterings. ApplArtifIntell 22: 2008, pp 780–810.