Алгоритм построения деревьев регрессии на основе клонового отбора иммунных клеток

Мельников Т.А., Мельников Г.А., Сташевский П.С.
Новосибирский государственный технический университет, Новосибирск, Россия

Аннотация. **Деревья регрессии являются одним из инструментов регрессионного анализа. Они позволяют осуществить разделение входного пространства на сегменты с последующим построением для каждого из них собственной модели и представить кусочно-заданную функцию регрессии в интуитивно понятной и наглядной форме. В этой статье мы представляем новый алгоритм построения регрессии моделей на основе моделирования клонового отбора иммунных клеток. Алгоритм сравнивается как с жадными алгоритмами, так и с эвристическими алгоритмами глобального поиска. Результаты экспериментов показывают, что предложенный алгоритм превосходит аналоги по среднеквадратичной адекватности идентификации и приводит к менее сложным моделям.**

***Ключевые слова*: деревья регрессии, эволюционные алгоритмы, алгоритм клонового отбора иммунных клеток.**

I. ВВЕДЕНИЕ

В настоящее время можно собирать и хранить огромные объемы данных, не прикладывая практически никаких усилий, и по впечатляюще низкой стоимости. Всё больше и больше компаний, исследовательских центров и правительственных учреждений создают огромные архивы таблиц, документов, изображений и звуков в электронной форме. Такой рост количества и многообразия доступных данных порождает необходимость в эффективных, робастных и гибких методах их анализа, с целью извлечения полезной информации из данных и использовании её при принятии решений.

Деревья регрессии являются одним из важных классов регрессионных моделей, позволяющим осуществить разделение входного пространства на сегменты с последующим построением для каждого из них собственной модели и представить кусочно-заданную функцию регрессии в интуитивно понятной и наглядной форме. В таком дереве внутренние узлы содержат правила разделения пространства объясняющих переменных; дуги – условия перехода по ним; а листья – локальные регрессионные модели. Для простоты интерпретации в большинстве случаев используются одномерные разделения вида $x_i \in B$, если x_i – категориальная переменная, и $x_i \leq c$, в противном случае.

В анализе данных деревья регрессии могут быть использованы для решения следующих задач:

1. Описание данных: описание данных в компактной форме;
2. Кластеризация: объединение объектов в группы (кластеры) на основе схожести признаков для объектов одной группы и их отличий между группами;
3. Регрессия: определение характера и формы зависимости между исследуемыми переменными (включая прогнозирование).

Многообразие решаемых задач не единственная привлекательная сторона деревьев регрессии. Они:

- практически не требуют никаких априорных знаний и\или допущений об исходных данных;
- обладают высоким уровнем автоматизации;
- способны работать с различными типами шкал, как с метрическими, так и с неметрическими;
- способны обрабатывать пропуски;
- интуитивно понятны и просты в интерпретации;
- прозрачны для анализа.

Несмотря на то, что возможность их применения в регрессионном анализе данных была продемонстрирована ещё в [1] (1984 год), алгоритмам данной группы было уделено сравнительно мало внимания. Целью данной работы является изложение сути нового метода построения деревьев регрессии на базе муравьиных алгоритмов.

II. ПРЕДЫДУЩИЕ РАБОТЫ

Большинство алгоритмов построения деревьев регрессии являются жадными. Это означает, что дерево строится путем рекурсивного разделения множества данных и принятия оптимального решения на конкретном шаге разделения. Этот процесс можно описать следующим образом:

1. Выбор локально оптимального разделения данных C по некоторому критерию R. На этом этапе выбирается объясняющая переменная, относительно которой будут делиться данные, и выбирается точка разделения для выбранной объясняющей переменной;
2. Разделение множества данных на подмножества;
3. Рекурсивное применение данного алгоритма к выделенным подмножествам.

Отличаются жадные алгоритмы построения деревьев регрессии главным образом правилом выбора лучшего разделения данных. Ряд альтернативных правил выбора разделений был предложен для построения деревьев регрессии.

Первые алгоритмы построения деревьев регрессии *CART* [1] и *M*5 [2] использовали правило минимизации дисперсии. В алгоритме *CART* минимизируется взвешенная сумма дисперсии целевой переменной после разделения, а в качестве локальных моделей используются средние значения целевой переменной, т.е. константы. Алгоритм M5 стал следующим шагом в развитии алгоритмов построения деревьев регрессии. Этот алгоритм использует правило выбора разделений схожее с *CART*, но строит линейные регрессионные модели в листовых узлах.

Правила выбора разделений на основе дисперсии не берут во внимание тип локальных моделей, поэтому в алгоритме *RETIS* [3] используется правило минимизации суммы квадратов остатков локальных моделей при выборе разделения. Однако, самый большой набор данных, на котором был протестирован алгоритм, содержал лишь 300 наблюдений, поскольку вычислительные затраты для этого правила очень велики. С тех пор многие авторы [4] – [6] пытались уменьшить вычислительную сложность данного правила.

Другой подход используется в алгоритмах *GUIDE* [7] и *SECRET* [8]. Они преобразуют исходную задачу регрессии в задачу классификации и затем используют методы, применяемые для построения деревьев классификации.

Жадные алгоритмы построения деревьев регрессии быстры и зачастую эффективны, но обычно приводят к локально оптимальным решениям. Существует несколько алгоритмов выходящих за рамки жадных алгоритмов: *M*5*Opt*, *GMT* и *AntMT*.

Алгоритм *M*5*Opt* [9]. В его основе лежит достаточно простая идея, направленная на избежание перебора всех возможных деревьев регрессии. На верхних уровнях используется полный перебор всех возможных вариантов разделения данных по атрибутам, а для оставшейся части используется жадный поиск. Это обеспечивает баланс между исследованием пространства поиска и временем выполнения алгоритма.

Алгоритм *GMT* [10] относится к алгоритмам генетического программирования и наследует основные черты алгоритмов данной группы. Алгоритм работает с совокупностью деревьев регрессии (популяцией особей) и для улучшения качества моделей использует аналоги механизмов генетического наследования, генетической изменчивости и естественного отбора. Все операции выполняются непосредственно над деревьями, например, в качестве оператора скрещивания может служить обмен поддеревьями, начиная с выбранных узлов.

Алгоритм *AntMT* [11] основан на моделировании поведения колонии муравьев. В основе алгоритма лежит идея моделирования непрямого обмена информацией через наблюдение особого вещества в окружающей среде – феромона, оставляемого муравьями и используемого ими при поиске кратчайшего пути от муравейника до источника пищи. Он побуждает муравьев следовать пути (выбирать разделения), который ведет к хорошему решению анализируемой задачи. Количество откладываемого на пути феромона пропорционально качеству полученного решения. Со временем феромон испаряется, что позволяет забыть неудачные разделения.

III. ПРЕДЛАГАЕМЫЙ АЛГОРИТМ

Биологические основы

Алгоритм базируется на гипотезе о том, что клетки, которые способны распознавать чужеродный антиген, размножаются в соответствии со степенью распознавания, т.е. чем лучше клетка распознает антиген, тем больше рождается ее клонов. При этом в процессе воспроизводства потомков они подвергаются мутациям, которые способствуют увеличению их способности распознавать антигены. Мутации проходят в соответствии с принципом чем лучше клетка, тем меньшим мутациям подвергается клетка и наоборот. Таким образом обеспечивается обучение иммунной системы. Клетки с плохой способностью к распознаванию удаляются, а количество клеток с большей способностью к распознаванию увеличивается.

Алгоритм построения деревьев моделей на основе моделирования клонового отбора иммунных клеток

Алгоритм выглядит следующим образом:

1. Инициализация. Начальная популяция из *N* деревьев генерируется случайным образом.
2. Рассчитывается значение целевой функции для каждого дерева.
3. Процесс обучения. Выполняются шаги 3.1 – 3.5.
   - 3.1. Отбор. Отбирается *S* наилучших деревьев.
   - 3.2. Репродукция. Создаются клоны ДР по правилу «чем лучше каждое дерево, тем больше клонов такого дерева создается». В разработанном алгоритме упорядоченная выборка отобранных деревьев делится на 3 части, для деревьев первой части делается 3 клона, для второй – 2, а для третьей – 1.
   - 3.3. Осуществление мутаций. Для каждого дерева произвести мутации по инверсивно-пропорциональному правилу: чем лучше дерево, тем меньше уровень мутации. Все упорядоченное по значение целевой функции множество клонов делится на три части, к лучшей части применяется только 1 случайная мутация, ко второй – 2, к третьей – 3.
   - 3.4. Создание новой популяции. Из популяции убирается *M* худших деревьев и добавляется до *N* случайно сгенерированных деревьев.
   - 3.5. Проверка критерия останова алгоритма. В качестве критерия останова в алгоритме используется достижение заданного числа итераций без изменения лучшего решения.

4. Завершение работы, отбор и оформление результата решения.

В качестве целевой функции в алгоритме используется расширенный байесовский информационный критерий из [12]. В нем осуществляется минимизация статистики:

$$EBIC = n \cdot \ln \frac{SSE}{n} + J \cdot (\ln n + \ln p), \quad (1)$$

где *SSE* – сумма квадратов остатков модели на обучающих данных; *J* – количество настраиваемых параметров модели; *n* – количество примеров в обучающей выборке; *p* – величина, характеризующая сложность пространства моделей (в нашем случае она равна произведению размера дерева на количество объясняющих переменных). В выражении (1) первое слагаемое – это максимальное значение логарифмической функции правдоподобия модели, а второе слагаемое представляет собой штраф за сложность модели. Данный критерий учитывает не только точность дерева, но и оценивает его сложность, благодаря чему алгоритм не должен переобучаться.

Виды мутаций

Алгоритм выглядит следующим образом:

1. Превращение листа в узел. Выбранный случайный лист заменяется на случайное дерево.
2. Превращение узла в лист. Узел превращается в лист, строится линейная регрессия от многих переменных для этого листа.
3. Обмен правилами между узлами. Выбираются два случайных узла, которые меняются правилами разделения данных.
4. Изменение переменной, выбранной для разделения в узле. Выбирается случайная объясняющая переменная, со случайной точкой разделения.
5. Изменение точки разделения при той же объясняющей переменной в узле. Объясняющая переменная остается той же, но случайным образом выбирается новая точка разделения.
6. Удаление переменной из локальной регрессионной модели.

IV. ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

Алгоритм был протестирован на 6 наборах данных из UC Irvine Machine Learning Repository [13] и KEEL-dataset repository [14]. Для алгоритма были выбраны следующие параметры: N – 120, S – 50, M – 60, количество узлов в деревьях начальной популяции варьировалось от 5 до 25. Для оценки выбраны два показателя – корень из среднего квадрата ошибки (RMSE) и размер деревьев регрессии. Все результаты получены с помощью 10-слойной перекрестной проверки и усреднены по 20 запускам. Результаты приведены в таблице 1.

На всех наборах данных предлагаемый алгоритм CSMT имеет адекватность идентификации выше, чем M5, RETIS и GMT на 1% – 48%. Однако на половине наборов данных он уступает алгоритму AntMT на 1% – 13%. Таким образом по адекватности идентификации разработанный алгоритм превосходит жадные алгоритмы, а также схожий с ним алгоритм GMT.

Если говорить о сложности моделей, то размер деревьев у алгоритма клонового отбора обычно меньше, чем у остальных алгоритмов. Стоит отметить, что на больших наборах данных, таких как Ailerons, конкуренты строят сравнительно большие модели, анализ которых достаточно труден. В то время как разработанный алгоритм составляет модели на этом же наборе данных, в среднем из 4,84 узлов, что на 12 – 121 узел меньше.

ТАБЛИЦА 1
СРАВНЕНИЕ ПРОИЗВОДИТЕЛЬНОСТИ АЛГОРИТМОВ ПОСТРОЕНИЯ ДЕРЕВЬЕВ РЕГРЕССИИ

	CSMT		AntMT		M5		RETIS’		GMT	
	RMSE	Size	RMSE	Size	RMSE	Size	RMSE	Size	RMSE	Size
Abalone	2.15	**4.0**	**2.14**	10.8	2.24	34.9	2.16	18.4	2.24	6.7
Ailerons	**0.000163**	**4.84**	0.000164	17.2	0.000192	126.1	0.000186	38.6	0.000200	24
Auto-mpg	**2.84**	7.36	3.06	11.4	3.34	11.0	3.18	12.8	3.23	**4.7**
CPU	**33.95**	**4.92**	52.14	5.5	74.2	7.6	57.08	7.1	63.4	6.1
Housing	4.23	**5.96**	**3.91**	12.1	4.35	23.8	4.31	10.2	4.21	6.6
Stock	1.17	**8.3**	**1.02**	22.4	1.08	65.2	1.03	32.2	1.22	18

V. ЗАКЛЮЧЕНИЕ

В данной работе представлен новый алгоритм построения деревьев регрессии на основе клонового отбора иммунных клеток. Эксперименты показывают, что алгоритм CSMT строит более простые модели, которые не уступают по адекватности идентификации эвристическим алгоритмам глобального поиска (AntMT, GMT) и превосходят модели, построенные жадными алгоритмами.

В качестве дальнейших направлений исследований можно выделить следующие:

- исследование и разработка различных компонент иммунных алгоритмов для решения задачи построения деревьев регрессии;
- исследование применимости нелинейных локальных моделей;
- введение внутренних регрессионных узлов для выделения общих факторов, одинаково влияющих на несколько сегментов.

ЛИТЕРАТУРА

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. “Classification and Regression Trees”, Wadsworth International Group, Belmont, 1984.

[2] J.R. Quinlan “Learning with continuous classes”, Proc. AI’92, 5th Australian Joint Conference on Artificial Intelligence, Singapore, 1992, pp. 343-348.

[3] Karalic “Employing linear regression in regression tree leaves”, Technical Report IJS DP-6450, Ljubljana, Slovenia: Jozef Stefan Institute, 1992.

[4] W.P. Alexander, S.D. Grimshaw “Treed regression”, Journal of Computational and Graphical Statistics, 1996, №5, pp. 156-175.

[5] D. Malerba, F. Esposito, M. Ceci, A. Appice “Top-down induction of model trees with regression and splitting nodes”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, №26, pp. 612-625.

[6] D. Vogel, O. Asparouhov and T. Scheffer “Scalable look-ahead linear regression trees”, In: Proc. of 13th ACM SIGKDD, New York, ACM Press, 2007, pp. 757–764.

[7] W.-Y. Loh “Regression trees with unbiased variable selection and interaction detection”, Statistica Sinica, vol. 12, 2002, pp. 361-386.

[8] Dobra, J. Gehrke “SECRET: A scalable linear regression tree algorithm”, In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 481-487.

[9] D.P. Solomatine, L. A. Siek “Semi-optimal Hierarchical Regression Models and ANNs”, Proc. Intern. Joint Conference on Neural Networks, Budapest, Hungary, 2004, pp. 1173-1177.

[10] M. Czajkowski, M. Kretowski “An Evolutionary Algorithm for Global Induction of Regression and Model Trees”, International Journal of Data Mining, Modelling and Management, in press.

[11] Мельников, Г.А. Применение методов искусственного интеллекта для исследования инфекционных заболеваний: магистерская дис. ... «Магистр техники и технологии»: 230100 / Мельников Григорий Андреевич. – г. Новосибирск, 2012. – 141 с.

[12] Chen J. Extended Bayesian information criteria for model selection with large model spaces / J. Chen, Z. Chen // Biometrika. – 2008. – P. 759–771

[13] Frank, A. Asuncion “UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]”, Irvine, CA: University of California, School of Information and Computer Science, 2010.

[14] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera “KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework”, Journal of Multiple-Valued Logic and Soft Computing, 2011, №17, pp. 255-287.

Тимофей Андреевич Мельников – магистрант кафедры вычислительной техники Новосибирского государственного технического университета.
Email: temmelnik@gmail.com

Григорий Андреевич Мельников – аспирант кафедры вычислительной техники Новосибирского государственного технического университета.
Email: grmel89@gmail.com

Сташевский Павел Сергеевич – к.т.н., доцент кафедры вычислительной техники Новосибирского государственного технического университета. Автор более 20 научных работ. Сфера научных интересов: методы машинного обучения, интеллектуальные системы.
Email: stashpavel@gmail.com

Algorithm of constructing the regression trees based on the clonal selection of the immune cells

TA. MELNIKOV, G.A. MELNILOV, S.P. STASHEVSKIY

Abstract. The regression tree is the instrument for regression analysis. They enable to division of the input space into segments with construction for each own model and present piecewise regression in clear and obvious form. This article presents a new algorithm of building regression models based on the modelling clonal selection of the immune cells. The algorithm is compared with greedy and heuristic global search algorithm. Results of the experiments show that the proposed algorithm is superior for

analogs RMS accuracy and leads to a less complicated models.

Key words: regression trees, evolutionary algorithms, the algorithm clonal selection of the immune cells

REFERENCES

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. "Classification and Regression Trees", Wadsworth International Group, Belmont, 1984.

[2] J.R. Quinlan "Learning with continuous classes", Proc. AI'92, 5th Australian Joint Conference on Artificial Intelligence, Singapore, 1992, pp. 343-348.

[3] Karalic "Employing linear regression in regression tree leaves", Technical Report IJS DP-6450, Ljubljana, Slovenia: Jozef Stefan Institute, 1992.

[4] W.P. Alexander, S.D. Grimshaw "Treed regression", Journal of Computational and Graphical Statistics, 1996, №5, pp. 156-175.

[5] D. Malerba, F. Esposito, M. Ceci, A. Appice "Top-down induction of model trees with regression and splitting nodes", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, №26, pp. 612-625.

[6] D. Vogel, O. Asparouhov and T. Scheffer "Scalable look-ahead linear regression trees", In: Proc. of 13th ACM SIGKDD, New York, ACM Press, 2007, pp. 757–764.

[7] W.-Y. Loh "Regression trees with unbiased variable selection and interaction detection", Statistica Sinica, vol. 12, 2002, pp. 361-386.

[8] Dobra, J. Gehrke "SECRET: A scalable linear regression tree algorithm", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 481-487.

[9] D.P. Solomatine, L. A. Siek "Semi-optimal Hierarchical Regression Models and ANNs", Proc. Intern. Joint Conference on Neural Networks, Budapest, Hungary, 2004, pp. 1173-1177.

[10] M. Czajkowski, M. Kretowski "An Evolutionary Algorithm for Global Induction of Regression and Model Trees", International Journal of Data Mining, Modelling and Management, in press.

[11] Мельников, Г.А. Применение методов искусственного интеллекта для исследования инфекционных заболеваний: магистерская дис. ... «Магистр техники и технологии»: 230100 / Мельников Григорий Андреевич. – г. Новосибирск, 2012. – 141 с.

[12] Chen J. Extended Bayesian information criteria for model selection with large model spaces / J. Chen, Z. Chen // Biometrika. – 2008. – P. 759–771

[13] Frank, A. Asuncion "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]", Irvine, CA: University of California, School of Information and Computer Science, 2010.

[14] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework", Journal of Multiple-Valued Logic and Soft Computing, 2011, №17, pp. 255-287.