№ графа	t	I	t_a	I^*	$I^* - I$	t/t_a
1	31,922	15	0,066	16	1	483,6667
2	27,653	16	0,638	16	0	43,34326
3	27,634	15	3,636	15	0	7,60011
4	25,34	16	16,881	16	0	1,501096
5	25,95	15	0,05	16	1	519
6	26,487	15	0,624	15	0	42,44712
7	26,717	15	5,793	15	0	4,611945
8	26,04	15	0,05	15	0	520,8
9	24,56	15	0,03	16	1	818,6667
10	23,096	14	0,587	14	0	39,34583

Алгоритм был также опробован на графах с известным типом структур (на двумерных решётках, на графах Гретша, Хватала, Хивуда и др.). Результаты экспериментов позволяют говорить о том, что вычисленные значения оценок являются хорошими и часто достижимыми. При этом время нахождения оценки существенно меньше, чем при использовании метода полного перебора. Однако было замечено, что алгоритм даёт бóльшую ошибку на графах с регулярной структурой (например, на двумерной решётке).

ЛИТЕРАТУРА

1. *Быкова В. В.* О мерах целостности графа // Прикладная дискретная математика. 2014. № 4(26). С. 96–111.
2. *Barefoot C. A., Entringer R., and Swart H. C.* Vulnerability in graphs — a comparative survey // J. Combin. Math. Combin. Comput. 1987. No. 1. P. 13–22.
3. *Clark L. H., Entringer R. C., and Fellows M. R.* Computational complexity of integrity // J. Combin. Math. Combin. Comput. 1987. No. 2. P. 179–191.
4. *Berry A. and Bordat J.-P.* Structuring the minimal separators of an undirected graphs. Technical Report 152, LIM, Marseille, 1996.

УДК 681.5.015 DOI 10.17223/2226308X/8/56

ПОСТРОЕНИЕ ФУНКЦИИ ОШИБКИ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ИДЕНТИФИКАЦИИ АЛГОРИТМА РАНЖИРОВАНИЯ

О. А. Кожушко

Предлагается функция ошибки для постановки задачи идентификации алгоритма ранжирования результатов текстового поиска. Приводится обоснование некорректности применения функций ошибки, используемых в задачах, где выходные значения принимают действительные значения. В качестве альтернативы предлагается функция ошибки, которая учитывает не абсолютное, а относительное изменение результатов ранжирования. Приводится частный случай постановки задачи идентификации, в которой результат ранжирования рассматривается как класс релевантности.

Ключевые слова: *алгоритм ранжирования, идентификация системы, функция ошибки.*

Задача идентификации системы подразумевает построение модели, устанавливающей взаимосвязь между входными и выходными значениями данной системы, и в общем виде ставится следующим образом [1].

Пусть исходная система реализует неизвестное отображение $F : DX \to DY$. Необходимо построить отображение $M_F : DX \to DY$ таким образом, что $M_F(x) = F(x)$

для всех $x \in DX$. В этом случае исходная система рассматривается как «чёрный ящик», для которого определены входы и выходы, однако неизвестны принципы его функционирования.

Задача идентификации системы в терминах машинного обучения ставится следующим образом. Пусть исходная модель F реализует функцию вида $F : DX \to DY \subset \mathbb{R}^m$. Необходимо построить модель $M_F : DX \to DY \subset \mathbb{R}^m$, такую, что на заданном множестве примеров $X = \{x_i : x_i \in DX, i = 1, \ldots, N\}$

$$E(F, M_F, X) < \varepsilon,$$

где E — функция ошибки; ε — заданная константа.

В данной работе в качестве исследуемой системы рассматривается алгоритм ранжирования. В общем виде алгоритм ранжирования осуществляет отображение вида

$$F_D(q, d) = \mathrm{rank}_D(f(q, d)),$$

где D — рассматриваемая коллекция документов, а функция rank сопоставляет документу порядковый номер в списке документов коллекции, отсортированном по убыванию значения функции релевантности. Функция релевантности f получает на вход пару векторов $\langle q, d\rangle$, описывающих текстовый запрос и документ соответственно, и вычисляет числовую оценку релевантности $r = f(q, d) \in \mathbb{R}$.

Ключевым отличием данной задачи идентификации алгоритма ранжирования от задачи построения ранжирующей функции является то, что выходные значения системы принимают ранговые значения. В настоящее время известно несколько работ, посвящённых решению задачи идентификации алгоритма ранжирования [2], однако в них нет чёткой постановки задачи.

При постановке задачи идентификации необходимо решить две следующих подзадачи. Первая состоит в определении множества входных факторов. Задача подбора факторов, задающих значения компонент входных векторов q и d, обычно решается с помощью экспертной оценки, исходя из эмпирических соображений и априори известной информации. Итоговый набор значимых факторов может быть получен следующими методами:

1) Подбором всех возможных факторов и дальнейшим исключение малозначимых. Значимость фактора может быть определена с помощью корреляционного анализа зависимости между значениями фактора и результатами ранжирования.
2) Методом AdDel [3], суть которого состоит в последовательном добавлении факторов, улучшающих качество идентификации, и удалении факторов, негативно влияющих на качество идентификации.

Вторая подзадача, решению которой посвящена данная работа, вытекает из свойств функции F_D, задающей частичный порядок на множестве пар векторов $\langle q, d\rangle$. Опишем эти свойства в виде леммы.

Лемма 1. Если $F_D(q, d_1) - F_D(q, d_2) = F_D(q, d_3) - F_D(q, d_4)$, то в общем случае $f_D(q, d_1) - f_D(q, d_2) \neq f_D(q, d_3) - f_D(q, d_4)$. То есть если разность рангов документов одинакова для двух пар документов, из этого в общем случае не следует равенство разности значений функции релевантности. Верно и обратное утверждение: если $f_D(q, d_1) - f_D(q, d_2) = f_D(q, d_3) - f_D(q, d_4)$, то в общем случае $F_D(q, d_1) - F_D(q, d_2) \neq$ $\neq F_D(q, d_3) - F_D(q, d_4)$.

Следует отметить, что значение ранга по конкретному запросу напрямую зависит от коллекции документов. Добавление в коллекцию одного документа, релевантного определённому запросу, изменит на 1 ранги множества документов по этому запросу.

Эти свойства функции F_D делают невозможным использование стандартных функций ошибки (таких, как среднеквадратичная функция ошибки MSE), применяемых для неранговых величин. В данной задаче применима следующая функция ошибки:

$$E(F, M_F, Q, D) = \frac{1}{N_D^2 N_Q} \sum_{q \in Q} \sum_{d_i, d_j \in D} |(F_D(q, d_j) - F_D(q, d_j)) - (M_F(q, d_j) - M_F(q, d_j))|,$$

где N_Q — количество запросов в множестве Q; N_D — количество документов в множестве D. С помощью данной функции оценивается разница в последовательностях ранжирования. Попарно сравниваются ранги документов, используемых в ранжировании по тестовым запросам. В случае, когда последовательные в исходном ранжировании документы следуют в модели M_F через k рангов друг от друга, значение функции ошибки увеличивается на $\dfrac{|k-1|}{N_D^2 N_Q}$. Следует отметить, что резкое изменение ранга одного документа в последовательности слабо влияет на значение функции ошибки. Значимые значения функция ошибки принимает в том случае, когда пары документов получают существенно различные ранги.

Задача идентификации может быть сведена к задаче классификации документов по степени релевантности. В этом случае определяется несколько степеней релевантности, например высоко-релевантные и низко-релевантные документы. Выход алгоритма ранжирования определяется не как ранг, а как степень релевантности. Такой подход обоснован тем, что при конструировании исходной системы ранжирования также используются оценки релевантности. Различие состоит в том, что при решении задачи идентификации степень релевантности задается тем, в какой промежуток значений рангов по данному запросу попадает данный документ, а в случае задачи конструирования алгоритма ранжирования — асессорами [4].

Задача классификации имеет следующий вид. Пусть определено M классов релевантности, а функция $\mathrm{class}(\mathrm{rank}(q, d))$ задаёт номер класса релевантности по рангу, присвоенному алгоритмом ранжирования. Необходимо построить идентифицирующую модель M_F, такую, что на заданном множестве примеров $X = \{\langle q, d\rangle_i \in \mathbb{R}^{n+m}, i = 1, \ldots, N\}$

$$E(F, M_F, X) = \frac{1}{N} \sum_{\langle q,d\rangle} I(\mathrm{class}(F_D(q, d), \mathrm{class}(M_F(q, d)) < \varepsilon,$$

где E — функция ошибки; ε — заданная константа;

$$I(x_1, x_2) = \begin{cases} 1, & \text{если } x_1 = x_2, \\ 0 & \text{иначе.} \end{cases}$$

В этом случае используется классическая функция ошибки для задачи классификации. Такой подход был успешно использован в работе [5]. Данная функция позволяет также избежать резких скачков при смене ранга одного документа за счёт интерпретации ранга как степени релевантности. Однако её использование влечёт за собой потерю части информации о разнице в последовательностях документов внутри классов релевантности.

В дальнейшем приведённые подходы могут быть использованы исследователями при построении идентифицирующих моделей как для алгоритмов ранжирования, так и для других функций, задающих частичный порядок на множестве векторов.

ЛИТЕРАТУРА

1. *Семенов А. Д., Артамонов Д. В., Брюхачев А. В.* Идентификация систем управления. Учеб. пособие. Пенза: Изд-во Пенз. ун-та, 2003. 211 с.
2. Материалы компании AlterTrader Research [Электронный ресурс]. `http://www.altertrader.com/` — 21.04.2015.
3. *Загоруйко Н. Г., Кутненко О. А.* Методы распознавания, основанные на алгоритме AdDel // Сиб. журн. индустр. матем. 2004. Т. 7. № 1. С. 39–47.
4. *Воронцов К. В.* Методы обучения ранжированию (Learning to rank). Курс лекций. [Электронный ресурс]. 2013. `http://www.machinelearning.ru/wiki/images/8/89/Voron-ML-Ranking-slides.pdf`
5. *Кожушко О. А., Тарков М. С.* Использование иерархической временной памяти для идентификации системы ранжирования документов // Проблемы информатики. 2015. №1(26). С. 47–54.

УДК 519.688 DOI 10.17223/2226308X/8/57

ПОЛИНОМЫ ХОЛЛА БЕРНСАЙДОВЫХ ГРУПП ПЕРИОДА 3[1]

А. А. Кузнецов, К. В. Сафонов

Пусть $B_k = (k, 3)$ — бернсайдова k-порождённая группа периода 3. В работе вычислены полиномы Холла для B_k при $k \leqslant 4$.

Ключевые слова: *периодическая группа, собирательный процесс, полиномы Холла.*

Пусть $B_k = (k, 3)$ — бернсайдова k-порождённая группа периода 3. Ф. Леви и ван дер Варден доказали [1], что $|B_k| = 3^{k+\binom{k}{2}+\binom{k}{3}}$ и ступень нильпотентности B_k не превышает 3.

Для каждой B_k несложно получить pc-представление (*power commutator presentation*), используя систему компьютерной алгебры GAP или MAGMA.

Пусть $a_1^{x_1} \dots a_n^{x_n}$ и $a_1^{y_1} \dots a_n^{y_n}$ — два произвольных элемента в группе B_k, записанные в коммутаторном виде. Тогда их произведение равно

$$a_1^{x_1} \dots a_n^{x_n} \cdot a_1^{y_1} \dots a_n^{y_n} = a_1^{z_1} \dots a_n^{z_n}.$$

Основой для нахождения степеней z_i является собирательный процесс [2, 3], который реализован в указанных системах компьютерной алгебры. Кроме того, существует альтернативный способ для вычисления произведений элементов группы, предложенный Ф. Холлом [4]. Холл показал, что z_i представляют собой полиномиальные функции (в нашем случае над полем $\mathbb{Z}_3$), зависящие от переменных $x_1, \dots, x_i, y_1, \dots, y_i$, которые принято сейчас называть *полиномами Холла*. Согласно [4],

$$z_i = x_i + y_i + p_i(x_1, \dots, x_{i-1}, y_1, \dots, y_{i-1}).$$

Необходимость применения полиномов Холла возникает при решении задач, требующих многократного умножения элементов группы. Исследование структуры графа Кэли некоторой группы является одной из таких задач. Вычислительные эксперименты на ЭВМ в группах периода пять и семь [5, 6] выявили, что метод полиномов Холла

[1]Работа выполнена при поддержке Министерства образования и науки РФ (проект Б 112/14) и гранта Президента РФ (проект МД-3952.2015.9).