Выбор оптимального алфавитного классификатора при минимизации общего числа операций

В.А. Тищенко

Аннотация. Определяется функционал $S_{оп}$ общего числа операций в классификаторе. Находится оптимальный классификатор в смысле максимального количества вершин в классе n_{max} и числа вершин в группе n при минимальном значении функционала $S_{оп}$. Вид функционала $S_{оп}$ дается как для случая одноуровневого, так и многоуровневого алфавитного классификатора. Приводится пример нахождения оптимальных значений средней длины ключа классификатора k* и максимального количества вершин в классе n_{max}* для поля ФИО.

Ключевые слова: *оптимальный алфавитный классификатор, максимальное число вершин в классе, средняя длина ключа алфавитного классификатора, число вершин в группе.*

1. Описание алфавитного классификатора на основе префиксного дерева сочетаний

Рассмотрим ключевой массив или индекс, состоящий из алфавитного списка вершин. Пусть этот индекс разбивается на классы по лексикографическому признаку посредством алфавитного классификатора [1,2]. Алфавитные ключи классификатора составляются из ключей префиксного дерева сочетаний (ПДС) для всевозможных путей в дереве. Каждый ключ – это соединение букв или буквенных сочетаний от корня до некоторого для каждого сочетания своего уровня. При этом длина алфавитного ключа класса выбирается минимальной настолько, чтобы соответствующий класс содержал не более, чем заданное количество вершин n_{max}. В каждый класс входят все вершины индекса, начинающиеся с алфавитного ключа класса. Каждый класс разбивается на группы по n вершин. Таким образом, в классе может содержаться несколько групп по n вершин, кроме последней группы, в которой может содержаться менее, чем n вершин (также и в случае одной группы).

2. Выбор оптимального алфавитного классификатора

Введем следующие обозначения. Пусть N – число вершин в индексе; k – число букв в алфавитном ключе, k=1,…,k_m, где k_m обозначает максимальную длину ключа; n(k) – число ключей длины k. Предполагается, что ключи классификатора нумеруются по длинам ключей в порядке левого обхода ПДС [3]. Пусть n(k,i) – число вершин под i-м ключом классификатора длиной k; r(k,i) – число вершин, входящих в последнюю группу для ключа длиной k с номером i: $r(k,i) = n(k,i) \bmod n$; m(k,i) – число групп по n вершин для ключа длиной k с номером i: $m(k,i) = \lceil n(k,i)/n \rceil + l(\{n(k,i)/n\})$, где квадратные и фигурные скобки – это целая и дробная часть числа соответственно, l(0)=0, l(x)=1 при x>0. Тогда общее число операций по поиску всех вершин на ключевом уровне массива определяется суммой:

$$S_{оп}(n) = \sum_{k=1}^{k_m} \sum_{i=1}^{n(k)} \left(S_g(k,i,n) + S_{gk}(k,i,n) + S_{gr}(k,i) \right) + S_k(n) . \quad (1)$$

Здесь первая сумма означает суммирование по всем длинам ключей от 1 до k_m, вторая сумма означает суммирование по всем ключам длины k от 1 до n(k), где n(k) – общее число ключей длины k. Первое слагаемое S_g(k,i,n) – суммарное число операций прохода по группам вершин для класса с ключом длины k и номером i:

$$S_g(k,i,n) = \sum_{j=0}^{m(k,i)-1} jn = \frac{(m(k,i)-1)m(k,i)}{2} n. \quad (2)$$

Второе S_{gk}(k,i,n) слагаемое – суммарное число операций прохода по вершинам групп для класса с ключом длины k и номером i, кроме последней группы, в которой может быть число вершин, меньшее n:

$$S_{gk}(k,i,n) = \sum_{w=1}^{n} w(m(k,i)-1) = \frac{n(n+1)}{2}(m(k,i)-1). \quad (3)$$

Третье слагаемые S_{gr}(k,i) – суммарное число операций прохода по вершинам последней группы числом r(k,i)≤n для класса с ключом длины k и номером i:

$$S_{gr}(k,i) = \frac{r(k,i)(r(k,i)+1)}{2} . \quad (4)$$

Эти три слагаемых стоят под двойной суммой. Последнее слагаемое S_k(n) – общее число

операций для алфавитных ключей классификатора состоит из трех слагаемых (см. формулу 5), подобных (2), (3) и (4). Однако весь уровень алфавитных ключей представляет собой один класс, разбитый на группы по *n* алфавитных ключей. Таким образом, у величин m_a числа групп по *n* алфавитных ключей и r_a числа алфавитных ключей, входящих в последнюю группу, соответствующих величинам m(k,i) и r(k,i), будут отсутствовать индексы *k* и *i*:

$$S_k(n) = \frac{(m_a - 1)m_a}{2} n + \frac{n(n+1)}{2}(m_a - 1) + \frac{r_a(r_a + 1)}{2}. \quad (5)$$

Величины m(k,i) и r(k,i) являются случайными величинами и случайным образом зависят от максимального числа вершин в классе n_{max} в силу неравномерности распределения текстовых вершин по буквенным сочетаниям.

В качестве оптимальных значений параметров алфавитного классификатора рассматриваются n_{max}^* (максимальное число вершин в классе, при котором достигается минимум $S_{оп}^*$) и связанное с ним регрессионной зависимостью [1,2] значение k^*. При n_{max}=N длина алфавитного ключа k=0 и сумма (1) сводится к трем слагаемым $S_{on}(N) = S_g(0,0,n) + S_{gk}(0,0,n) + S_{gr}(0,0)$ и n(0,0)=N.

На примере поля ФИО[34 657] (данные взяты из базы данных «За Христа пострадавшие»*) рассмотрим процесс нахождения оптимальных значений k^* и n_{max}^* в указанном выше смысле. Процесс нахождения минимального значения функционала (1) $S_{оп}(n_{max},n)$ сводится к полному перебору значений максимального числа вершин в классе n_{max}, например, от 1 до 1000. Предполагается, что число вершин в группе n также меняется в некотором диапазоне, например, 10≤n≤100.

На рис.1 показаны графики зависимости максимального числа вершин в классе при минимальном значении $S_{оп}$ от числа вершин в группе $n_{max}^*(n)$ и суммарного числа операций, минимального по величине относительно n_{max}^*, от числа вершин в группе $S_{оп}(n_{max}^*,n)$. Как видно из графика $S_{оп}(n_{max}^*,n)$ минимум (при целых значениях n на диапазоне от 10 до 100 с шагом 10) достигается при n^*=20 и соответствующего значения n_{max}^*=176 (или n_{max}^*=177) $S_{оп}^*=S_{оп}(n_{max}^*,n^*)$=505080. Фактически минимум $S_{оп}^*$ достигается при n=15. При этом величина $n_{max}^*(n)$ имеет сильные колебания. На рисунке показана сглаженная зависимость.

Таким образом, оптимальный классификатор в данном случае будет содержать максимальное число вершин в группе, равное 176, и средняя длина ключа класса в этом случае равна 3,106.

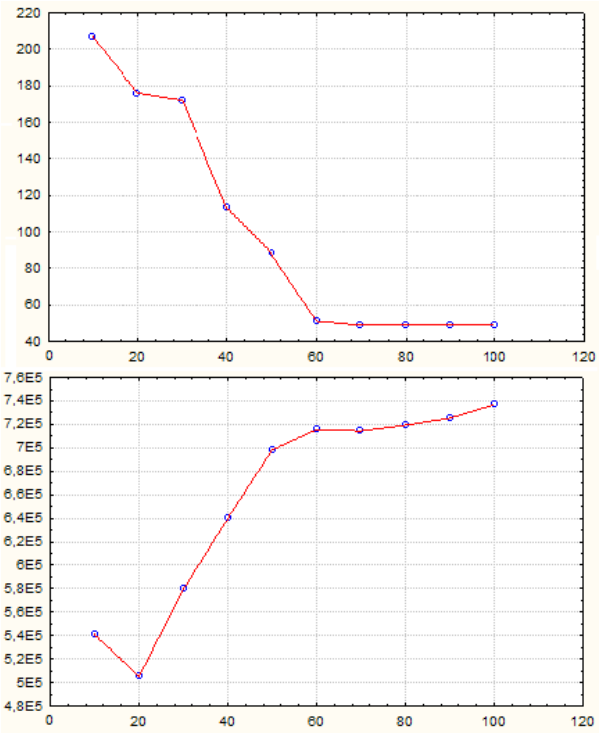

Рис.1. Графики $n_{max}^*(n)$ и $S_{оп}(n_{max}^*,n)$

На рис. 2 качественно показана зависимость $S_{оп}(n_{max},20)$ при $n=n^*=20$, т.е. для значения n, когда достигается оптимальное число операций $S_{оп}^*$. Минимум при n_{max}^*=176 условно показан при n_{max}=200. Шаг делений по оси абсцисс также неодинаковый. Кроме того, необходимо отметить, что действительный минимум $S_{оп}^*$ достигается при n^*=15 и n_{max}^*=178 (или n_{max}^*=179) и равен $S_{оп}^*$=487 117.

3. Вид функционала общего числа операций в общем случае

Необходимо отметить, что формула (1) предполагает наличие одного уровня классификатора, который состоит из алфавитных ключей, ссылающихся на соответствующие классы вершин ключевого массива. При более общем рассмотрении можно обобщить результат (1) для многоуровневого классификатора. При этом каждый более высокий уровень классификатора будет являться одноуровневым классификатором для подчиненного уровня. Формула (1) в этом случае перепишется в виде трех слагаемых, соответствующих (2), (3) и (4), которые стоят под тремя суммами. Внешняя сумма будет соответствовать сумме по всем уровням многоуровневого классификатора, а две внутренние суммы будут соответствовать суммам в формуле (1). Также величины m(k,i), r(k,i) и n(k) будут уже зависеть еще от номера уровня в клас-

* БД «За Христа пострадавшие», http://martyrs.pstbi.ru

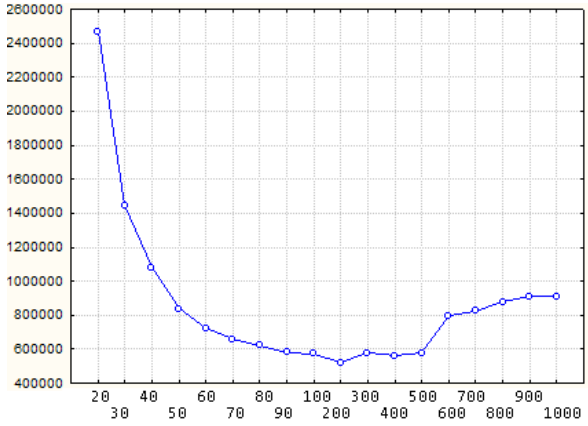

Рис. 2. Схематичная форма графика $S_{on}(n_{max},20)$

сификаторе h. Верхний уровень алфавитного классификатора будет состоять из одного класса. Число уровней классификатора $h_m = [\bar{k}/\Delta k]+1$. Здесь $\bar{k}$ обозначает среднюю длину ключа на последнем уровне классификатора, а Δk — среднее число букв, которое добавляется к ключу на каждом уровне классификатора. В величину h_m включается также ключевой уровень массива, поэтому добавляется 1:

$$S_{on}(n) = \sum_{h=1}^{h_m} \sum_{k=1}^{k_m} \sum_{i=1}^{n(h,k)} \left(S_g(h,k,i,n) + S_{gk}(h,k,i,n) + S_{gr}(h,k,i) \right) \quad (6)$$

В рассмотренном примере применялся одноуровневый классификатор, так как ключевой массив состоит из относительно небольшого количества вершин – 34 657 и число ключей в оптимальном алфавитном классификаторе (n^*=20) получается относительно небольшим – порядка 1,5 тысяч. При необходимости этот классификатор можно надстроить двумя-тремя уровнями алфавитных ключей и получить многоуровневый классификатор, например, первый уровень – однобуквенный классификатор, а второй – двух-трех буквенный классификатор.

Полученный качественный вид зависимости $S_{оп}(n_{max},n)$ с одним характерным глобальным минимумом можно считать достаточно общим. Например, в случае равномерного распределения текстовых ключей по сочетаниям при общем числе ключей 810 000 минимум $S_{оп}^*$=24 732 450 достигается при n^*=35 или n^*=36 и n_{max}=900. С другой стороны, случайные распределения длины ключа классификатора и числа вершин в классе [1,2] могут изменяться в зависимости от типа рассматриваемых полей, т.е., например, ключевой массив фамилий или ключевой массив адресов.

Литература

1. *Емельянов Н.Е., Тищенко В.А.* Методология построения многоуровневого индекса ключевого массива по лексикографическому признаку на основе метода регрессионного анализа на примере СУБД НИКА // Обработка информационных и графических ресурсов / Сб. трудов ИСА РАН. Т.58. Под ред. Арлазарова В.Л. – М. 2010. С. 6-17.
2. *Соловьев А.В, Тищенко В.А.* Проблемы построения многоуровневого алфавитного классификатора (на примере ключевого уровня массива СУБД НИКА).
3. *Богачева А.Н., Емельянов Н.Е.* Семантическая модель документа // Системные исследования. Ежегодник 2001 / М.: Едиториал УРСС. 2003. С. 360-375.

Тищенко Владимир Александрович. Научный сотрудник ИСА ФИЦ ИУ РАН. Закончил МИФИ в 1993г. Количество печатных работ: 17. Область научных интересов: средства создания и поддержки электронных библиотек и электронных изданий. E-mail: vtischenko@isa.ru

The selection of optimal alphabetical classifier while minimizing of the total number of operations

Tishchenko V.A.

Abstract. The functional S_{op} of the total number of operations in the classifier is defined. There is an optimal classifier in the sense of the maximum number of vertices in the class n_{max} and the number of vertices in the group n with the minimum value of the functional S_{op}. The form of the functional S_{op} is given both for the case of a single-level and multilevel alphabetic classifier. An example of finding the optimal values of the average key length of the classifier k^* and the maximum number of vertices in the class n_{max}^* for the field Name is given.
Keywords: *optimal alphabetic classifier, the maximum number of vertices in a class, the average length of an alphabetical classifier key, the number of vertices in a group*

References

1. *Emelyanov N.E., Tischenko V.A.* 2010. Metodologiya postroeniya mnogourovnevogo indeksa klyuchevogo massiva po leksikograficheskomu priznaku na osnove metoda regressionnogo analiza na primere SUBD NIKA [Methodology for constructing a multilevel index of a key array based on the lexicographic characteristic based on the regression analysis method on the example of the NIKA database]. Trudy ISA RAN "Obrabotka informatsionnih i graficheskih resursov" [ISA RAS "Processing of information and graphics resources" Proceedings]. 58:6–17.
2. *Solovyov A.V., Tishchenko V.A.* 2018. Problemi postroeniya mnogourovnevogo alfavitnogo klassifikatora (na primere klyuchevogo urovnya massiva SUBD NIKA) [Problems of constructing a multilevel alphabetic classifier (for example, the key level of an array of NIKA DBMS)].
3. *Bogacheva A.N., Emelyanov N.E.* 2001. Semanticheskaya model' dokumenta [Semantic model of the document]. Sistemnie issledovaniya. Ezhegodnik [System Research. Yearbook]. M: Editorial URSS. 360-375.

Tishchenko Vladimir Alexandrovich. Researcher, ISA FRC CSC RAS.Graduated from the MEPhI in 1993. Number of publications: 17. Research interests: means of creation and support of electronic libraries and electronic publications. E-mail: vtischenko@isa.ru