ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 148 за 2019 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Гавриков М.Б., Пестрякова Н.В.

Исследование характеристик метода распознавания на модификации обучающего множества

Рекомендуемая форма библиографической ссылки: Гавриков М.Б., Пестрякова Н.В. Исследование характеристик метода распознавания на модификации обучающего множества // Препринты ИПМ им. М.В.Келдыша. 2019. № 148. 12 с. **http://doi.org/10.20948/prepr-2019-148**
URL: **http://library.keldysh.ru/preprint.asp?id=2019-148**

О р д е н а   Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й   а к а д е м и и   н а у к

М.Б. Гавриков, Н.В. Пестрякова

Исследование характеристик метода распознавания на модификации обучающего множества

Москва — 2019

М.Б. Гавриков, Н.В. Пестрякова

Исследование характеристик метода распознавания на модификации обучающего множества

Описываются результаты статистического исследования функции оценки и вектора первичных признаков, полученные при распознавании множества, являющегося заданной модификацией обучающей базы рукопечатных символов, с применением разработанного авторами метода распознавания.

Ключевые слова: классификация, полиномиальная регрессия, печатные и рукопечатные символы

Mikhail Borisovich Gavrikov, Nadejda Vladimirovna Pestryakova

A research of characteristics of the recognition method

The results of a statistical study of the evaluation function and the vector of primary features are described obtained by recognizing a set that is a given modification of the training base of hand-written characters using the recognition method developed by the authors.

Key words: classification, polynomial regression, printed and hand-printed symbols

Оглавление

1. Введение

Настоящая работа посвящена разработанному авторами статистическому методу распознавания символов [1] и исследованию его свойств [2–6]. Этот метод является точным, быстрым, удобным для распараллеливания, а также устойчивым к искажениям. Генерируемые им оценки обладают свойством монотонности, т.е. надежности. Данный подход имеет несомненные преимущества в сравнении с другими [2]. По точности и быстродействию он практически не уступает получившим в последнее время широкое распространение искусственным нейронным сетям, которые могут выставлять максимальные оценки при ошибочном распознавании.

Были разработаны новые эффективные приложения данного метода для объектов иной природы, не являющихся изображениями [7–11].

В основе рассматриваемого подхода лежит восстановленный с большой степенью достоверности некоторый неизвестный вероятностный закон, в соответствии с которым распределены элементы обучающей последовательности символов, моделирующей датчик случайных векторов. Высокий уровень точности распознавания на обучающем множестве служит подтверждением правомерности этого приближения.

Однако наше знание об оценке распознавания ограничено пониманием алгоритма построения механизма для ее вычисления, а также непосредственной его реализацией на компьютере.

Здесь исследуется «внутреннее устройство» указанного механизма. Для этого применяется численный подход. При обучении используется база рукопечатных символов, представляемых заданным способом в виде векторов признакового пространства. Это множество имеет большой объем, а набор его элементов является случайным. Распознаются искаженные изображения обучающих символов, полученные на основе разработанной авторами математической модели.

2. Метод распознавания

Решается следующая задача классификации. Нужно определить, какому из K символов соответствует поступающий на вход распознавателя растр изображения. Нормализованный растр состоит из N=$N1$×$N2$ серых пикселов, перенумерованных в диапазоне $1 \le i \le N$. Пусть $\mathbf{v} \in \mathbf{R}^N$ – вектор, i-я компонента которого равна яркости пиксела с таким же номером. Для рассматриваемых серых растров значения яркости принадлежат отрезку [0,1].

Символ, имеющий номер k ($1 \le k \le K$), отождествляем с базисным вектором $\mathbf{e}_k$=(0…1…0) из $\mathbf{R}^K$, где 1 стоит на k-м месте. Обозначаем Y={$\mathbf{e}_1$,…,$\mathbf{e}_K$}.

Пусть $\mathrm{p}_k(\mathbf{v})$ – вероятность того, что входящий растр изображает k-й символ для $1 \le k \le K$. Распознанным считается символ с номером k_o, причем

$$\mathrm{p}_{k_0}(\mathbf{v})=\max \mathrm{p}_k(\mathbf{v}),\ \ 1\le k\le K. \tag{1}$$

Значения компонент $(\mathrm{p}_1(\mathbf{v}),\ldots,\mathrm{p}_K(\mathbf{v}))$ можно приближенно представить в виде многочленов от координат $\mathbf{v}=(\mathrm{v}_1,\ldots,\mathrm{v}_\mathrm{N})$:

$$\mathrm{p}_k(\mathbf{v})\cong c_0^{(k)}+\sum_{i=1}^{N} c_i^{(k)} v_i\ +\sum_{i,j=1}^{N} c_{i,j}^{(k)} v_i v_j+\ldots,\ 1\le k\le K. \tag{2}$$

Суммы в правых частях приближенных равенств (2) конечные и определяются выбором базисных мономов. Пусть

$$\mathbf{x}(\mathbf{v})=(1,v_1,\ \ldots\ ,v_N,\ \ldots\ )^\mathrm{T}$$

– вектор конечной размерности L из выбранных и упорядоченных некоторым образом базисных мономов. Тогда (2) можно записать в векторном виде:

$$\mathbf{p}(\mathbf{v})=(\mathrm{p}_1(\mathbf{v}),\ldots,\mathrm{p}_K(\mathbf{v}))^\mathrm{T}\cong A^\mathrm{T}\mathbf{x}(\mathbf{v}). \tag{3}$$

Здесь A – матрица размера $L\times K$, столбцами которой являются векторы $\mathbf{a}^{(1)},\ldots,\ \mathbf{a}^{(K)}$. Каждый такой вектор составлен из коэффициентов при мономах соответствующей строки (2) (с совпадающим верхним индексом), упорядоченных так же, как в векторе $\mathbf{x}(\mathbf{v})$. Приближенный поиск вектора вероятностей $\mathbf{p}(\mathbf{v})$ сводится к нахождению A.

Матрица A вычисляется приближенно в процессе обучения. Для этого используется обучающее множество, состоящее из пар векторов $[\mathbf{v}^{(1)},\mathbf{y}^{(1)}]$, …, $[\mathbf{v}^{(J)},\mathbf{y}^{(J)}]$ ($\mathbf{v}^{(j)}$ – образ символа с номером k ($1\le k\le K$), $\mathbf{y}^{(j)}=(0\ldots1\ldots0)$ – его базисный вектор, где 1 стоит на k-м месте, причем $1\le j\le J$):

$$A\cong\Big(\frac{1}{J}\sum_{j=1}^{J}\mathbf{x}^{(j)}(\mathbf{x}^{(j)})^\mathrm{T}\Big)^{-1}\Big(\frac{1}{J}\sum_{j=1}^{J}\mathbf{x}^{(j)}(\mathbf{y}^{(j)})^\mathrm{T}\Big). \tag{4}$$

Рекуррентная процедура вычисления данной матрицы и структура полиномиального вектора **x(v)** подробно описывались ранее [1–4].

3. Моделирование искажений изображений

Обучение проводилось на базе рукопечатных цифр из 174778 элементов, а распознавание – на множестве, полученном путем ее модификации.

Пусть на вход распознавателя поступают перенумерованные изображения ($1\le\ j\le\ J_{k_0}$) только одного определенного k_0-го символа (в данной работе это цифра «8»). Каждое из них будет распознано в качестве k-го символа с оценкой p_k^j для $1\le k\le K$, где K = 10. Согласно (2) и (3),

$$p_k^j = \sum_{l=1}^{L} a^{lk} x_l^j, \qquad 1 \le j \le J_{k_0},\ k = 1,\ldots,K, \tag{5}$$

где a^{lk} - весовой коэффициент, x_l^j - моном.

Исследуем, какие изменения происходят с совокупностью значений $\{a^{lk} x_l^j\}$ при распознавания множества, элементы которого получены в результате заданной модификации элементов обучающего множества, причем величина искажения постепенно увеличивается.

Моделирование процесса нарастания различия обучающего и распознаваемого множеств осуществляется следующим образом. При затемнении образа, поступающего для распознавания, значение яркости в каждом пикселе растра постепенно увеличивается: $v_i \to v_i + 0{,}1 \cdot n$, где $n = 0, 1, .., 9$. Если для каких-то пикселов начиная с некоторого n имеем: $v_i > 1$, то считаем $v_i = 1$.

В более ранних исследованиях [3, 4] показано, что картина распознавания для различных символов не имеет существенных различий. Поэтому рассмотрен только один символ. В табл. 1 показано, как изменяется количество неправильно распознанных его изображений с усилением затемнения.

Таблица 1(1)

n	0	1	2	3	4
число ошибок	138	71	106	339	2957

Таблица 1(2)

n	5	6	7	8	9
число ошибок	6457	9761	10116	10121	10121

4. Статистический анализ функции оценки

Для изображений символа «8», поступающих на вход распознавателя, определим, в каком диапазоне лежат значения составляющих, входящих в выражение для оценки в виде слагаемых $a^{lk_0} x_l^j$, а также как они распределены внутри этого диапазона для случая, когда в качестве альтернативы распознавания выбирается только сам символ. Нижняя и верхняя границы соответствующего диапазона t^{k_0} и T^{k_0} вычисляются по формулам:

$$t^{k_0} = \min_{l,j} (a^{lk_0} x_l^j),$$

$$1 \le l \le L,\ 1 \le j \le J_{k_0}. \qquad (6)$$

$$T^{k_0} = \max_{l,j} (a^{lk_0} x_l^j),$$

В табл. 2 приведены указанные диапазоны для различных значений n:

Таблица 2

	n = 0	n ≥ 1
$\min_{l,j} (a^{lk_0} x_l^j)$	-0,037	-0,037
$\max_{l,j} (a^{lk_0} x_l^j)$	0,051	0,061

Для единообразия будем рассматривать только наибольший диапазон (при n≥1). Делим диапазон [-0,037; 0,061] на 10 равных частей и для каждого из n (n≥0) определяем, какое количество значений указанных слагаемых $a^{lk_0} x_l^j$ оказывается на каждом таком маленьком отрезке (табл. 3).

Из табл. 1 видно, что на начальном этапе затемнения точность распознавания улучшается, при этом для n = 0,1 число неправильных распознаваний минимально (выделено серым цветом). Такое поведение точности распознавания коррелирует с особенностью изменения картины распределения числа слагаемых $a^{lk_0} x_l^j$. А именно, на первых шагах увеличения n значение максимума количества указанных слагаемых, наблюдающегося в четвертом интервале (пик распределения в окрестности нуля [5], в табл. 3 выделен серым цветом), увеличивается и достигает наибольшей величины, когда n = 0,3 (в табл. 3 обозначено жирным курсивом). Нарастание n приводит к тому, что в близлежащих интервалах (третьем, пятом и шестом) число мономов падает и достигает минимума при n =0,1 в третьем интервале, при n = 0,3 – в пятом и n = 0,2 – в шестом (в табл. 3 выделено жирным курсивом). При дальнейшем увеличении n картина распределения постепенно «расплывается», число слагаемых $a^{lk_0} x_l^j$ в четвертом (пиковом) интервале распределения уменьшается, а во всех других, включая близлежащие, – увеличивается.

Определим, в каком диапазоне лежат составляющие $a^{lk} x_l^j$, а также как они распределены внутри этого диапазона для случая, когда в качестве альтернативы распознавания выбирается любой из возможных символов «0», «1», …, «9». Нижняя и верхняя границы соответствующего диапазона t и T вычисляются по формулам:

Таблица 3(1)

n	1 интервал	2 интервал	3 интервал	4 интервал	5 интервал
0	124	659	126309	221647273	2667904
0,1	215	1094	***107223***	221802217	2531081
0,2	343	1564	111413	221931741	2397075
0,3	562	2281	124882	***222059446***	***2253006***
0,4	799	3395	176639	221622472	2623139
0,5	1105	5028	227901	221269448	2899881
0,6	1523	7829	311359	220687930	3357950
0,7	2122	12580	457786	219748030	4067998
0,8	3419	24895	758029	218359711	5030651
0,9	6285	74769	1313762	216762545	5769642

Таблица 3(2)

n	6 интервал	7 интервал	8 интервал	9 интервал	10 интервал
0	49653	1926	40	13	0
0,1	49079	2928	46	14	3
0,2	***47650***	4020	61	22	7
0,3	48436	5121	103	33	22
0,4	59349	7210	620	217	34
0,5	80067	9264	816	291	57
0,6	112792	12255	1644	425	113
0,7	184364	17054	2233	1352	237
0,8	278477	30515	4761	2672	478
0,9	479803	75670	2117	5640	3115

$$t = \min_{l,j,k} (a^{lk} x_l^j),$$

$$1 \le l \le L,\ 1 \le j \le J_{k0},\ k = 1,\ldots,K. \qquad (7)$$

$$T = \max_{l,j,k} (a^{lk} x_l^j),$$

Оказалось, что диапазон изменения значений слагаемых $a^{lk} x_l^j$ не изменяется при увеличении числа n. Он имеет следующие границы:

$$t = -0{,}070; \quad T = 0{,}082.$$

Делим этот диапазон на 10 равных частей и исследуем, как изменяется число составляющих $a^{lk} x_l^j$ на каждом таком отрезке с нарастанием n (табл. 4).

Таблица 4(1)

n	1 интервал	2 интервал	3 интервал	4 интервал	5 интервал
0	38	244	29291	173573188	2071292360
0,1	29	206	30351	169888827	2074974533
0,2	21	208	33411	***169721542***	***2075132823***
0,3	33	247	38795	172440255	2072401159
0,4	456	1872	69322	179933587	2064831059
0,5	488	2035	87164	193699713	2051019893
0,6	542	2403	119566	207798255	2036834981
0,7	647	3428	197098	217912916	2026513816
0,8	994	11161	412953	224575455	2019340942
0,9	3721	71570	977204	230136962	2012470329

Таблица 4(2)

n	6 интервал	7 интервал	8 интервал	9 интервал	10 интервал
0	43400	481	6	2	0
0,1	44626	421	14	3	0
0,2	50549	420	31	5	0
0,3	57934	527	43	16	1
0,4	98841	3749	105	15	4
0,5	125171	4317	193	30	6
0,6	176978	5839	375	57	14
0,7	299912	10163	845	158	27
0,8	567677	25522	3846	389	71
0,9	1189203	67255	18934	3467	365

Аналогично табл. 3, отмеченное ранее поведение точности распознавания коррелирует с особенностью изменения картины распределения числа слагаемых $a^{lk} x_l^j$ для всех альтернатив. А именно, на начальных этапах увеличения n значение максимума количества таких слагаемых, наблюдающегося в пятом интервале (пик распределения в окрестности нуля [5], выделен серым цветом), увеличивается и достигает наибольшего значения при n = 0,2 (обозначен жирным курсивом). При этом в одном из близлежащих интервалов, а именно, в четвертом, число указанных слагаемых падает и достигает минимума также при n = 0,2 (выделен жирным курсивом). Далее увеличение n приводит к тому, что картина распределения постепенно «расплывается», при этом число таких слагаемых в пятом (пиковом) интервале распределения уменьшается, а во всех других, включая четвертый, увеличивается.

Для случая, когда в качестве альтернативы распознавания выбирается только сам символ «8», определим, как слагаемые $a^{lk_0} x_l^j$ распределены внутри данного диапазона [-0,070; 0,082] (табл. 5).

Таблица 5(1)

n	1 интервал	2 интервал	3 интервал	4 интервал	5 интервал
0	0	1	670	17764935	206717860
0,1	0	1	1133	***17593244***	***206886879***
0,2	0	1	1721	17786765	206690641
0,3	0	1	2607	18256263	206217862
0,4	0	1	3868	19861280	204607010
0,5	0	1	5635	21855712	202604729
0,6	0	1	8349	23934499	200512824
0,7	0	1	13444	25133748	199282436
0,8	0	1	24988	25724663	198629015
0,9	0	1	71419	26196133	198010758

Таблица 5(2)

n	6 интервал	7 интервал	8 интервал	9 интервал	10 интервал
0	10417	18	0	0	0
0,1	12608	32	4	0	0
0,2	14713	49	11	0	0
0,3	17058	81	29	0	0
0,4	21225	460	57	0	0
0,5	27082	647	95	0	0
0,6	36684	1355	189	0	0
0,7	61380	2529	363	0	0
0,8	108476	6047	711	0	0
0,9	205195	6922	3473	0	0

Результаты аналогичны табл. 3, 4. А именно, наблюдается корреляция с зависимостью точности распознавания от n, проявляющаяся в том, что в «пиковом» пятом интервале (выделен серым цветом) достигается максимум в распределении числа указанных слагаемых при n = 0,1 (обозначен жирным курсивом), а в четвертом интервале имеется минимум при n = 0,1 (выделен жирным курсивом). Последующее увеличение n делает распределение все более расплывшимся, так что количество таких слагаемых в пятом (пиковом) интервале падает, а во всех других, включая четвертый, – нарастает.

5. Статистический анализ вектора первичных признаков

Определим, в каком диапазоне лежат компоненты вектора первичных признаков и как они распределены внутри этого диапазона. Величина x_l^j для $1 \le l \le L$, $1 \le j \le J_{k0}$ находится в диапазоне [s, S]:

$$s = \min_{j,l}(x_l^j),\ S = \max_{j,l}(x_l^j),\ \text{где } 1 \le l \le L,\ 1 \le j \le J_{k0}. \tag{8}$$

Получены следующие значения величин: $s = -1$, $S = 2$. Заметим, что они одинаковы для различных входящих символов и не изменяются при нарастании степени затемнения *n*.

Поделим отрезок [*s*, *S*] на десять равных частей и выясним, какое количество значений мономов x_l^j относится к каждому такому маленькому отрезку (табл. 6) для различных *n*.

Таблица 6(1)

n	1 интервал	2 интервал	3 интервал	4 интервал	5 интервал
0	282450	1818259	4396597	158383616	17502304
0,1	251659	1755056	4235217	158099618	18251197
0,2	198096	1398375	4369833	156455946	23192333
0,3	100725	1449984	4340849	**156369685**	25494290
0,4	169057	1350596	4331584	***163592860***	21349697
0,5	169057	1275793	4217447	157412159	23341780
0,6	169057	1045923	4111449	154438650	24260830
0,7	169057	1058009	3917827	150972632	24541884
0,8	169057	1072025	3062731	149467887	14351597
0,9	169057	1208650	2787987	148946109	9746844

Таблица 6(2)

n	6 интервал	7 интервал	8 интервал	9 интервал	10 интервал
0	10459153	29675642	1100160	615148	260572
0,1	14804423	25201879	1106710	592514	195628
0,2	12275326	24753447	1145045	581070	124430
0,3	10476565	24375736	1271563	507472	107032
0,4	8760659	23128718	1237288	462632	110810
0,5	11932934	24427298	1251577	355046	110810
0,6	12768527	26078003	1155606	355046	110810
0,7	14600060	27723117	1045459	355046	110810
0,8	20409316	34593549	901883	355046	110810
0,9	14740885	45818832	609681	355046	110810

Картина распределения перестраивается, причем в имеющемся пике распределения в окрестности нуля [5] (на четвертом интервале) вначале значение максимума постепенно уменьшается (при n=0,3 достигается минимум, выделенный прямым жирным шрифтом), затем нарастает (наибольшая величина при n=0,4, обозначено жирным курсивом) и далее уменьшается. Значения в граничных (первом и десятом) интервалах и приграничном (девятом) постепенно стабилизируются.

6. Заключение

Для разработанного авторами метода распознавания символов изучен механизм генерации оценок. На обучающем множестве рукопечатных цифр большого объема проведено статистическое исследование функции оценки. Найдены особенности распределения составляющих $a^{lk} x_l^j$, входящих в виде слагаемых в формулу для ее вычисления, а также компонент вектора первичных признаков x_l^j [5]. В данной работе аналогичные исследования проведены на множествах, полученных в результате искажения исходного, причем величина различия между ними постепенно нарастает. Показано, что поведение точности распознавания коррелирует с характером изменения картины распределения числа слагаемых $a^{lk0} x_l^j$ и $a^{lk} x_l^j$.

7. Библиографический список

1. Гавриков М. Б., Пестрякова Н. В. Метод полиномиальной регрессии в задачах распознавания печатных и рукопечатных символов // Препринты ИПМ им. М.В.Келдыша. 2004. №22. 12 с.
URL: http://keldysh.ru/papers/2004/prep22/prep2004_22.html

2. Об одном методе распознавания символов, основанном на полиномиальной регрессии / М. Б. Гавриков [и др.] // Автоматика и Телемеханика. 2006. №2. С. 119-134.

3. Пестрякова Н.В. Структуры в распознавании // Информационные технологии и вычислительные системы. 2009. №1. С. 58-71.

4. Пестрякова Н.В. Динамика качества распознавания при нарастании степени различия баз обучения и распознавания // Информационные технологии и вычислительные системы. 2010. №2. С. 75-82.

5. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Статистический анализ характеристик метода распознавания на обучающем множестве // Информационные технологии и вычислительные системы. 2015, №1. С. 82-88.

6. Гавриков Б.М., Пестрякова Н.В. Многовариантное численное моделирование в задаче исследования устойчивости методов статистического распознавания к искажениям образов // Вестник РФФИ. 2016. Вып. 92. № 4. С.124-134.

7. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Статистический метод распознавания и классификации на основе полиномиальной регрессии // Материалы XXI Международной конференции по вычислительной механике и современным прикладным программным системам (ВМСППС'2019), 24–31 мая 2019 г., Алушта. — М.: Изд-во МАИ, 2019. С. 143-145.

8. Гавриков Б.М., Пестрякова Н.В. О построении признакового пространства в задаче обучения // Информационные технологии и вычислительные системы. 2018. №1. С. 22-29. DOI: 10.14357/20718632180104

9. Гавриков Б.М., Пестрякова Н.В., Ставицкий Р.В. О свойствах обучающих множеств // Информационные технологии и вычислительные системы. 2018. №4. С. 97-107. DOI: 10.14357/207186321804010

10. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. О структуре базы обучения классификатора для оценивания состояния здоровья человека // Препринты ИПМ им. М.В.Келдыша. 2018. №126. 18 с. DOI:10.20948/prepr-2018-126

11. Гавриков Б.М., Гавриков М.Б., Пестрякова Н.В. Структура базы обучения статистического классификатора состояний систем организма человека // Препринты ИПМ им. М.В.Келдыша. 2018. №255. 40 с. DOI:10.20948/prepr-2018-255