Приближенное вычисление оценки вероятности сложного события в условиях объективной недостаточности статистических опытов

© Солодухин А.

Омский государственный технический университет
sandys13@yandex.ru

Аннотация

В работе рассматривается один из вариантов решения проблемы объективной недостаточности статистических опытов для вычисления оценки вероятности произведения большого количества элементарных событий путем дополнительных допущений.

1. Введение

В таких задачах как анализ покупательской корзины, экспертные системы, классификация текстов и других часто возникает задача вычисления оценки вероятности произведения элементарных событий.

Определим модель, часто используемую при решении данных задач. Пусть $A=\{a_1, a_2, \dots a_j, \dots a_N\}$ – неупорядоченное множество (алфавит), состоящее из элементарных событий a_j, между которыми имеется только отношение эквивалентности. Будем называть **сложным событием** или сочетанием событий произведение некоторых элементарных событий из A, то есть подмножество A. Пусть $T=\{t_1, t_2, \dots t_i, \dots t_n\}$ – неупорядоченное множество t_i подмножеств A, т.е. $t_i \subseteq A$. Назовем множество t_i **статистическим опытом**, также множество t_i является сложным событием. **Оценка вероятности** сложного события t_i есть отношение количества его

встреч n_i к размеру множества *T*. На практике [1] стоит задача оценки вероятностей различных сложных событий, как входящих во множество *T*, так и не входящих.

При решении практических задач [1,2] размер алфавита *N* может принимать значения порядка 10^3-10^6. Причем с течением времени возникают задачи, в которых актуальны все большие размеры алфавита. Известно, что при размере алфавита *N* количество возможных подмножеств событий равно 2^N (мощность булеана). Для того чтобы каждое подмножество *A* встретилось во множестве *T* в среднем один раз необходимо, чтобы размер множества *T* был порядка 2^{1000}-$2^{1000000}$, что на практике невозможно. Отсюда следует **проблема объективной недостаточности статистических опытов**, в данной модели задачи, для значимой оценки вероятностей сложных событий.

Известным приемом [3] в таких случаях является допущение независимости сочетаний событий. Например, имеется сочетания событий *a* и *b*, тогда $p^*(ab)=p^*(a)p^*(b)$ по теореме умножения вероятностей независимых событий, т.е. за $p^*(ab)$ принимается не отношение n_{ab}/n которое равно нулю из-за $n_{ab}=0$, а произведение $n_a/n \cdot n_b/n$. На практике, зачастую, неизвестна заранее взаимная независимость сочетаний событий. Тогда возникает вопрос: возможно ли оценить независимость сочетаний событий, основываясь на имеющихся статистических опытах? Из теории вероятностей известно, что если $p(ab) \neq p(a)p(b)$ то события *a* и *b* считаются зависимыми. Известна формула $p(ab)=p(a)p(b/a)=p(b)p(a/b)$ для зависимых *a* и *b*. Для оценок событий получается следующее равенство $p^*(ab)=p^*(a)p^*(b/a)=n_a/n \cdot n_{ab}/n_a=n_b/n \cdot n_{ab}/n_b=n_{ab}/n$, которое ни к чему не приводит, поскольку $n_{ab}=0$. Для определения независимости событий можно принять за правило условие-равенство $p^*(ab)=p^*(a)p^*(b)$, но в условиях недостаточности статистических опытов, когда $n_{ab}=0$, это правило неприменимо. Таким образом, можно сделать вывод, что определить опытным путем независимость сочетаний событий, входящих в большое сложное событие, в условиях недостаточности статистических опытов невозможно без дополнительных допущений.

2. Предлагаемый метод

Рассмотрим множество $T=\{t_1, t_2, \ldots t_i, \ldots t_n\}$ как последовательность сгенерированных **событий реализаций** некоторым **источником без памяти**. Допустим, что данный источник обладает свойствами **стационарности** и **эргодичности**. Тогда можно утверждать, что модель такого источника с алфавитом $A=\{a_1, a_2, \ldots a_j, \ldots a_N\}$ может быть представлена в виде распределения вероятностей множеств булеана 2^A с отрицаниями недос-

тающих событий до полного алфавита (как несовместных событий). Назовем данную модель как **распределение вероятностей множества совместных событий**. Например, *A={a,b,c}*, тогда распределение вероятностей будет множество {*p(abc), p(a~b~c), p(~ab~c), p(~a~bc), p(ab~c), p(~abc), p(a~bc), p(~a~b~c)*}, данное преобразование представлено на рис. 1.

		совместные события		
несовместные события	p_1	*a*	*b*	*c*
	p_2	*a*	*b*	*~c*
	p_3	*a*	*~b*	*c*
	p_4	*a*	*~b*	*~c*
	p_5	*~a*	*b*	*c*
	p_6	*~a*	*b*	*~c*
	p_7	*~a*	*~b*	*c*
	p_8	*~a*	*~b*	*~c*

Рис. 1. Преобразование алфавита совместных элементарных событий в алфавит несовместных сложных событий.

Очевидно, получение данного распределения является главной целью исходной рассматриваемой задачи. Учитывая условия задачи – объективная недостаточность статистических опытов, а также большой размер алфавита подсчет частот событий данного распределения приведет к недостаточно точным значениям оценок вероятностей. Делая допущения независимости подмножеств алфавита можно прогнозировать значения оценок вероятностей сложных событий, частоты которых малы. Вычисление распределения вероятностей множества совместных событий для независимых подмножеств *a* и *b* делается по следующим формулам:

$$p(a)\,p(b) = p(ab),$$
$$p(a)\,p(\bar{b}) = p(a\bar{b}),$$
$$p(\bar{a})\,p(b) = p(\bar{a}b),$$
$$p(\bar{a})\,p(\bar{b}) = p(\bar{a}\bar{b}).$$

Например, *A={a,b,c}* и разбиение на независимые подмножества *{{a,b},{c}}*, тогда используя теорему умножения независимых событий можно вывести распределение вероятностей множества совместных событий {*p(ab)p(c), p(a~b)p(~c), p(~ab)p(~c), p(~a~b)p(c), p(ab)p(~c), p(~ab)p(c), p(a~b)p(c), p(~a~b)p(~c)*}, данное распределение представлено на рис. 2.

		Независимые распределения вероятностей множеств совместных событий		
Распределение вероятностей множества совместных событий исходного алфавита	p_1	$p(ab)$	·	$p(c)$
	p_2	$p(ab)$	·	$p(\sim c)$
	p_3	$p(a\sim b)$	·	$p(c)$
	p_4	$p(a\sim b)$	·	$p(\sim c)$
	p_5	$p(\sim ab)$	·	$p(c)$
	p_6	$p(\sim ab)$	·	$p(\sim c)$
	p_7	$p(\sim a\sim b)$	·	$p(c)$
	p_8	$p(\sim a\sim b)$	·	$p(\sim c)$

Рис. 2. Распределение вероятностей множества совместных событий с независимыми подмножествами.

Выбрав в качестве целевой функции критерий правильности разбиения алфавита на независимые подмножества, можно использовать алгоритм локального поиска, тогда шагом может являться перенос набора событий из одного независимого подмножества в другое. Для поиска оптимального разбиения алфавита на независимые подмножества возможно применение таких алгоритмов [4] как поиск с восхождением к вершине, поиск с эмуляцией отжига, локальный лучевой поиск, генетический алгоритм. Можно ожидать, что процесс поиска оптимального разбиения алфавита на независимые подмножества будет достаточно продолжительным, поэтому скорость вычисления целевой функции представляется актуальной. Известно [4], что еще важным свойством целевой функции является наличие локальных оптимумов-ловушек и пологих областей.

В качестве целевой функции можно использовать эффективность решения задачи – если модель позволяет решать задачу лучше значит она более правильная. Например, эффективность решения задачи, может быть, характеризоваться количеством правильно спрогнозированных некоторых контрольных значений из тестовой выборки. Контрольные значения необходимо выбирать так чтобы они могли охарактеризовать максимально полно разбиение алфавита, иначе будут образовываться локальные оптимумы-ловушки и пологие области по причине того, что изменения, какой то части модели никак не будут оцениваться. Выбор теоретических обоснований для использования такой целевой функции в данной работе рассматриваться не будет.

В качестве целевой функции можно использовать энтропию Шэннона [5] – минимум энтропии будет означать наиболее правильное разбиение. Другими словами, целевая функция будет являться длинной кода для выделения последовательности *T*. Свойство энтропии – энтропия независимых источников равна сумме энтропий этих источников, можно использовать для оптимизации скорости вычисления целевой функции.

Попробуем сформулировать теоретические обоснования с точки зрения теории вероятностей. Пусть источник имеет алфавит, состоящий из только несовместных событий, т.е. когда разбиение алфавита на независимые подмножества не имеет смысла, причем источник без памяти и также обладает свойствами стационарности и эргодичности – рис. 3. Тогда нашей целью будет являться определение распределения вероятностей исходного алфавита $A=\{a_1, a_2, \dots a_j, \dots a_N\}$.

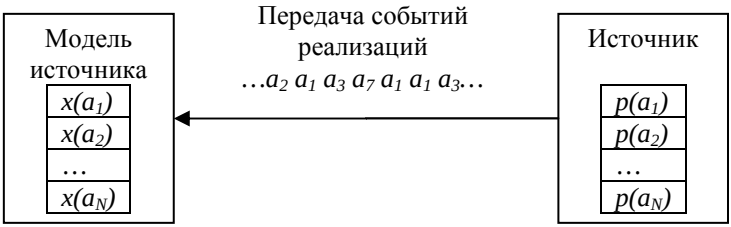

Рис. 3. Построение модели источника с алфавитом, состоящим из несовместных событий.

В данном случае очевидное решение задачи в ситуации, когда есть возможность наблюдать только последовательность ограниченной длины n это использование формулы: $x(a_i) = \frac{n_i}{n}$, где n_i есть количество повторений a_i в наблюдаемой последовательности. Это приведет к приемлемым на практике оценкам $x(a_i)$ сходящимся к $p(a_i)$ с увеличением n.

Случай, когда источник имеет алфавит, состоящий из совместных событий (и возможно из несовместных как частный случай) в условиях ограниченной выборки представлен на рис. 4 (для простоты обозначений $A=\{a,b,c\}$).

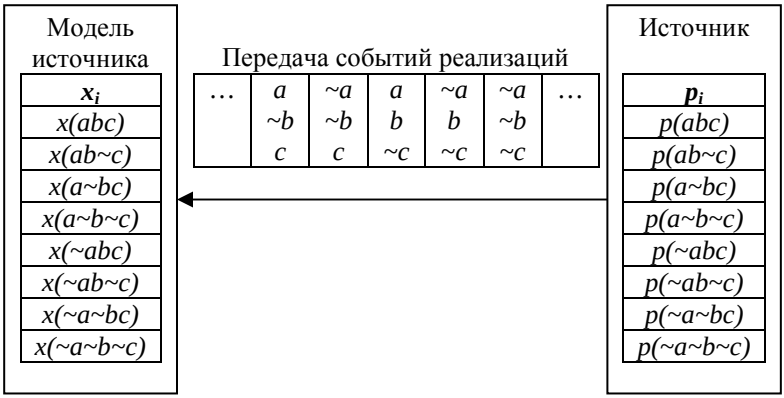

Рис. 4. Построение модели источника с алфавитом, состоящим из совместных событий.

В данном случае по-прежнему стоит цель определения x_i как наилучшего приближения p_i. Но применение формулы $x_i = \frac{n_i}{n}$ не представляется эффективным ввиду количества возможных событий реализаций источника равного 2^N (где N размер исходного алфавита) в условиях ограниченной выборки. Можно предположить, что источник имеет алфавит, состоящий из набора взаимно независимых подмножеств событий и, как уже было сказано, стоит задача поиска оптимального разбиения исходного алфавита на независимые подмножества. Некоторое выбранное разбиение исходного алфавита позволяет вычислять распределение x_i с использованием теоремы умножения независимых событий, как это показано на рис. 2 (причем в произведение подставляются частоты событий посчитанных по классической формуле вероятности), таким образом, меняя разбиение, мы меняем значения распределения x_i. Известно, что вероятность последовательности событий реализаций $T=\{t_1, t_2, \dots t_i, \dots t_n\}$, для данной модели, равна произведению их вероятностей $p(T)=p(t_1)p(t_2)\dots p(t_i)\dots p(t_n)$. Величина $p'(T)= x(t_1)x(t_2)\dots x(t_i)\dots x(t_n)$ зависит от разбиения исходного алфавита на независимые подмножества. Можно утверждать, что величина $p'(T)$ максимальна только в том случае, когда значения распределения x_i равны соответствующим значениям распределения p_i это доказано в теореме 1. Очевидно, что данное свойство следует из того, что в данной модели источника, при генерации события реализации, ожидаемость случайного события **прямо пропорциональна** его доли в распределении вероятностей, т.е. источник больше всего «ожидает»

последовательность, сгенерированную по его собственному распределению.

Теорема 1. Пусть имеется алфавит источника $A=\{a_1, a_2, \dots a_j, \dots a_N\}$ совместных событий, распределение истинных вероятностей множества совместных событий $\{p_1, p_2,\dots, p_K\}$, варьируемое распределение оценок вероятностей множества совместных событий $\{x_1, x_2,\dots,x_K\}$ и последовательность событий реализаций источника $T=\{t_1, t_2, \dots t_i, \dots t_n\}$, сгенерированных на основе распределения $\{p_1, p_2,\dots, p_K\}$. Тогда вероятность произведения $x(t_1)x(t_2)\dots x(t_i)\dots x(t_n)$ будет максимальна только в том случае, когда $p_i=x_i$, где $n\to\infty$ и $K=2^N$.

Доказательство теоремы 1. Легко доказать, проведя дифференцирование по x_i, что

$$E = -\frac{\log(p(T))}{n} = -\sum_{i=1}^{2^N} p_i \log(x_i),$$

где $\sum_{i=1}^{2^N} p_i = 1, \sum_{i=1}^{2^N} x_i = 1$ минимально только тогда, когда $p_i=x_i$.

Варьируемое распределение оценок вероятностей множества совместных событий $\{x_1, x_2,\dots,x_K\}$ представляет собой вычисляемое распределение на основе разных разбиений алфавита на независимые подмножества. Из теоремы 1 следует, что энтропия будет иметь экстремум в том случае, когда искомое распределение будет найдено.

Для вычисления энтропии источника с заданным алфавитом совместных событий необходимо вычислить частоты вхождений в T каждого из исходов распределения вероятностей множества совместных событий. В случае если источники независимы, то их энтропии вычисляются раздельно и затем складываются. Формула вычисления энтропии источника без памяти для последовательности ограниченной длины имеет вид:

$$E = -\frac{\log(p(T))}{n} = -\sum_{i=1}^{2^N} p_i \log(p_i) = -\sum_{i=1}^{2^N} \frac{n_i + d}{n + d2^N} \log(\frac{n_i + d}{n + d2^N}),$$

где d величина смещения вероятностей всех возможных исходов, которая характеризует априорное распределение вероятностей исходов, когда отсутствует статистика. Величина d определяет порог зависимости/независимости множеств событий и от ее значения во многом зависит разбиение алфавита с минимальным значением энтропии.

3. Эксперименты

Поскольку предлагаемый метод очень требователен к вычислительным ресурсам, рассмотрим процесс построения модели источника без памяти на простейшем примере исходных данных «Golf» представленных в таблице 1. Автору неизвестны способы оптимизации предлагаемого метода для решения задач с размером алфавита порядка $N>1000$ средствами одного персонального компьютера.

Таблица 1. «Golf»

	Outlook	**Temperature**	**Humidity**	**Wind**	**Play**
1	sunny	85	85	false	no
2	sunny	80	90	true	no
3	overcast	85	80	false	yes
4	rain	70	95	false	yes
5	rain	70	80	false	yes
6	rain	65	70	true	no
7	overcast	65	65	true	yes
8	sunny	70	95	false	no
9	sunny	70	70	false	yes
10	rain	75	80	false	yes
11	sunny	75	70	true	yes
12	overcast	70	90	true	yes
13	overcast	80	75	false	yes
14	rain	70	80	true	no

Для значений колонок «Outlook»,«Wind»,«Play» поставим в соответствие идентификаторы событий для каждого варианта значения каждой колонки. Для значений колонок «Temperature»,«Humidity» для каждого значения x поставим в соответствие набор идентификаторов событий: «$x<f$» и «$x>f$» где f принимает все дискретные значения из данной колонки. В таблице 2 представлены возможные идентификаторы событий для выбранных исходных данных, зарезервированные для отражения значений во множестве T – алфавит источника (видно, что размер алфавита N=27).

Таблица 2. Возможные идентификаторы событий для таблицы «Golf»

	Outlook	**Temperature**	**Humidity**	**Wind**	**Play**
Зарезервированные идентификаторы	{0, 1, 2}	{3, 4, 5, 6, 7, 8, 9, 10}	{11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}	{23, 24}	{25, 26}

В итоге множество T будет выглядеть следующим образом (колонки и строки оставлены для наглядности) – таблица 3.

Таблица 3. Возможное множество T для таблицы «Golf»

	Outlook	**Temperature**	**Humidity**	**Wind**	**Play**
1	0	3,4,5,6	11,12,13,14,21,22	23	25
2	0	3,4,5,10	11,12,13,14,15,22	24	25
3	1	3,4,5,6	11,12,13,20,21,22	23	26
4	2	3,8,9,10	11,12,13,14,15,16	23	26
5	2	3,8,9,10	11,12,13,20,21,22	23	26
6	2	7,8,9,10	11,18,19,20,21,22	24	25
7	1	7,8,9,10	17,18,19,20,21,22	24	26
8	0	3,8,9,10	11,12,13,14,15,16	23	25
9	0	3,8,9,10	11,18,19,20,21,22	23	26
10	2	3,4,9,10	11,12,13,20,21,22	23	26
11	0	3,4,9,10	11,18,19,20,21,22	24	26
12	1	3,8,9,10	11,12,13,14,15,22	24	26
13	1	3,4,5,10	11,12,19,20,21,22	23	26
14	2	3,8,9,10	11,12,13,20,21,22	24	25

Используя оговоренные выше алгоритмы локального поиска и энтропию источника в качестве целевой функции, вычисляемой по указанной ранее формуле, построена зависимость оптимального разбиения алфавита A на независимые подмножества от параметра d, которая представлена в таблице 4.

Таблица 4. Зависимость оптимального разбиения алфавита *A* на независимые подмножества от параметра *d*

d	**Оптимальное разбиение алфавита *A* на независимые подмножества**	**Максимальный коэффициент сжатия последовательности *T* (N/E)**
10^{-9}	{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14, 15,16,17,18,19,20,21,22,23,24,25,26}	6.577307
10^{-7}	{3,4,7,8,11,17} {0,1,2,5,6,9,10,12,13,14,15,16, 18,19,20,21,22,23,24,25,26}	4.801769
10^{-5}	{0,1,2,4,5,6,8,9,10,23,24,25,26} {3,7,11,12,13,14,15,16,17,18,19,20,21,22}	4.144976
10^{-3}	{3,7,12,13,18,19} {4,5,6,8,9,10} {14,15,16,20,21,22,23,24} {0,1,2,11,17,25,26}	3.280769
10^{-1}	{0,1,2} {6,10} {11,17} {25,26} {12,13,18,19} {16,22} {23,24} {14,15,20,21} {4,5,8,9} {3,7}	2.199000
10^{0}	{0,1,2} {23,24} {25,26} {20,21} {11} {3,7} {12,18} {14,15} {17} {5,9} {16,22} {13,19} {6,10} {4,8}	1.407803
10^{1}	{1} {10} {7} {23,24} {25,26} {5,9} {20,21} {17} {16} {3} {4,8} {6} {13,19} {22} {2} {14,15} {0} {11} {12,18}	1.039986
10^{3}	{1} {10} {11} {14,20} {3} {12} {7} {9} {17} {16} {2} {21} {5} {4,8} {25,26} {13,19} {22} {18} {23,24} {15} {6} {0}	1.000010

Из таблицы видно, как влияет параметр *d* на оптимальное разбиение – уменьшение *d* уменьшает энтропию. В случае отсутствия разбиения на всем алфавите распределение выглядит, как представлено на рис. 5.

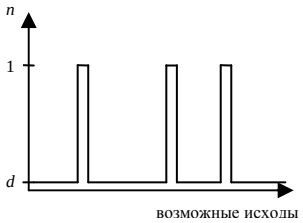

Рис. 5. Возможный вид распределения в случае отсутствия разбиения алфавита на независимые подмножества.

Чем меньше значение *d*, тем меньше вклад в распределение исходов с нулевой частотой, что говорит о том, что те исходы, которые не встречались, маловероятно встретятся в будущем и наоборот. Параметр *d* необходимо выбирать либо постоянным для каждой конкретной задачи, либо выполнять оптимизацию по данному параметру, а в качестве целевой функции использовать критерий эффективности решения задачи.

Литература

[1] Bart Goethals. Survey on Frequent Pattern Mining. HIIT Basic Research Unit. Department of Computer Science. University of Helsinki, 2003.

[2] Солодухин А.С. Классификация текстов на основе приближенных оценок вероятностей классов // Системы управления и информационные технологии, 2007, N3.3(29). - С. 379-384.

[3] Peter Jackson. Introduction to Expert Systems. Harlow, England: Addison Wesley, Longman,1999.

[4] Рассел С., Норвиг П. Искусственный интеллект: современный подход, 2-изд. – М.: Издательский дом «Вильямс», 2006. – 1408 с.

[5] Claude Shannon A Mathematical Theory of Communication. Bell System Technical Journal, 1948, vol. 27, July, pp. 379–423, October, pp. 623–656.

Approximate complex event probability estimation in conditions objective insufficiency statistical tests

In the given work is presented one of few variants decision the problem objective insufficiency statistical tests for calculating probability estimation of product a big number elementary events by additional assumption.