**XVII ВСЕРОССИЙСКАЯ КОНФЕРЕНЦИЯ МОЛОДЫХ УЧЕНЫХ
ПО МАТЕМАТИЧЕСКОМУ МОДЕЛИРОВАНИЮ И
ИНФОРМАЦИОННЫМ ТЕХНОЛОГИЯМ**

Новосибирск, ИВТ СО РАН 30 октября — 3 ноября 2016 г.

ПИКОВЫЕ ХАРАКТЕРИСТИКИ ФУНКЦИИ ЭНТРОПИИ СЛОВ И КЛАСТЕРИЗАЦИЯ СЕМЕЙСТВ РАСТЕНИЙ

УЛЬЯНОВ МИХАИЛ ВАСИЛЬЕВИЧ

д.т.н., проф., в.н.с. ИПУ РАН им. В.А. Трапезникова,

профессор ВМК МГУ им. М.В. Ломоносова

2016

ВВЕДЕНИЕ И ОБЛАСТИ ПРИМЕНЕНИЯ

Объект исследования — слова конечной длины над конечным алфавитом.

Предмет исследования — информативные характеристики слов, отражающие разнообразие подслов переменной длины в окне сдвига 1.

Рассматривается подход к анализу информации, представленной с использованием символьного кодирования, при котором образы исследуемых объектов или процессов представляются словами над некоторым конечным алфавитом.

Изучением таких символьных представлений занимается раздел современной дискретной математики — *комбинаторика слов*. Предельный случай бесконечных последовательностей является предметом изучения в *символической динамике*.

Для слов конечной длины вводятся *пиковые характеристики функции энтропии слов*, и показывается их возможное применение в качестве осей кластерного пространства на примерах семейств растений.

ЭНТРОПИЯ В СИМВОЛИЧЕСКОЙ ДИНАМИКЕ

В символической динамике энтропия используется как характеристика разнообразия элементов пространств сдвигов в словах бесконечной длины.

Пусть S_F — пространство сдвигов — множество бесконечных последовательностей над конечным алфавитом Σ, не содержащих в качестве подслов конечных слов из заданного множества F, тогда энтропия такого пространства сдвигов определяется как

$$H(S_F) = \lim_{n\to\infty} \frac{1}{n} \log|B_n|,$$

где B_n — множество подслов длины n, встречающихся в последовательностях из S_F.

В частности, для периодических бесконечных последовательностей $H = 0$, поэтому энтропия длинных периодических слов также близка к нулю.

Задача исследования:

Вычисление энтропии конечных слов, как функции переменной длины окна и исследование ее особенностей.

ТЕРМИНОЛОГИЯ И ОБОЗНАЧЕНИЯ

Далее в докладе будет использоваться следующая терминология и обозначения:

Σ — алфавит, s — произвольный символ алфавита;

$L_k = L(\Sigma^k) = \{w \mid |w| = k\}$ — множество всех слов длины k над алфавитом Σ;

$SW(w,i,l)$ — оператор выделения подслова длины l в слове w, начиная с символа в позиции i. Пусть $|w| = n$, тогда оператор определен при $i + l - 1 \le n$:

$$SW(s_1 s_2 \ldots s_n, i, l) = u = s_i s_{i+1} \ldots s_{i+l-1};$$

$SH1(w,k)$ — оператор сдвига 1. Определенный при $|w| > k$ оператор порождает мультимножество подслов длины k мощности $|w| - k + 1$, выполняя сдвиг на единицу окна длины k по слову w, начиная с крайней левой позиции слова w:

$$SH1(w,k) = \{u_j \mid j = 1, |w| - k + 1; u_j = SW(w, j, k)\},$$

для результата оператора $SH1(w,k)$ мы, очевидно, допускаем мультимножества:

$$SH1(1101010,4) = \{1101, 1010, 0101, 1010\} = \{1101, 1010^{(2)}, 0101\}.$$

ЭНТРОПИЯ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ С КОНЕЧНЫМ НОСИТЕЛЕМ

Рассмотрим классическую вероятностную модель с конечным носителем $M = \langle \Omega, P(\cdot) \rangle$, в которой вероятностное пространство Ω конечно — $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$, а вероятностная мера $P(\cdot)$ определяет вероятности случайных событий — $P(\omega_i) = p_i$. Тогда, по определению, энтропия дискретного распределения определяется в виде

$$H_P \overset{def}{==} -\sum_{i=1}^{k} p_i \log p_i.$$

Выбрав основание логарифма равное мощности вероятностного пространства мы можем нормировать максимальное значение энтропии конечного дискретного распределения в единицу.

При $P(\omega_i) = 1/|\Omega| = 1/k$ имеем

$$H_p = -\sum_{i=1}^{k} p_i \log_{|\Omega|} p_i = -\sum_{i=1}^{k} \frac{1}{k} \log_k \frac{1}{k} = 1.$$

ФУНКЦИЯ ЭНТРОПИИ СЛОВ (I)

Построение этой функции для слова w над конечным алфавитом Σ выполняется в два этапа:

— на первом этапе вычисляется энтропия слова w для подслов длины k;

— на втором этапе происходит обобщение на подслова произвольной длины.

На первом этапе мы фиксируем длину подслова k. Множество всех слов длины k над алфавитом Σ есть $L_k = L\left(\Sigma^k\right)$. Пусть $U = L\left(\Sigma^k\right)$. Введем в рассмотрение формальное мультимножество

$$\widetilde{U} = \left\{ u_i^{(0)} \mid i = \overline{1, |\Sigma|^k} \right\},$$

элементы которого (все слова длины k над алфавитом Σ) имеют нулевую кратность.

Далее применим к исследуемому слову w оператор сдвига 1 с окном ширины k и получим порожденное оператором $SH1(w,k)$ мультимножество $\widetilde{V}$

$$\widetilde{V} = SH1(w,k) = \left\{ v_i^{(c_i)} \mid i = \overline{1, l} \right\}.$$

ФУНКЦИЯ ЭНТРОПИИ СЛОВ (II)

Пусть m есть число позиций окна ширины k на слове w, отметим, что m равно сумме кратностей элементов мультимножества $\widetilde{V}$

$$m = \sum_{i=1}^{l} c_i = |w| - k + 1 = n - k + 1.$$

Построим объединенное мультимножество $\widetilde{V} \cup \widetilde{U}$, но поскольку $\widetilde{U}$ содержит все возможные слова длины k, то $\widetilde{V} \cup \widetilde{U}$ не будет содержать новых по отношению к $\widetilde{U}$ элементов. Тем самым объединение мультимножеств приведет только к изменению кратностей некоторых, или быть может всех элементов из $\widetilde{U}$. На этой основе введем вероятностную модель

$$M_k = \left\langle \Omega_k = \widetilde{V} \cup \widetilde{U},\ P_k(\cdot) \right\rangle,$$

где мощность $|\Omega_k| = |\Sigma|^k$, а вероятностная мера определяется через частоту появления различных слов длины k в слове w на основе кратности элементов мультимножества $\widetilde{V}$

ФУНКЦИЯ ЭНТРОПИИ СЛОВ (III)

$$P_k(\cdot):\ \begin{cases} p(\omega_i) = \dfrac{c_i}{m}, & \omega_i \in \widetilde{V} \\ p(\omega_i) = 0, & \omega_i \in \widetilde{U} \setminus \widetilde{V} \end{cases}$$

Для полученной вероятностной модели с вероятностной мерой $P_k(\cdot)$ мы определяем нормированную энтропию, используя $|\Omega_k| = |\Sigma|^k$ в качестве основания логарифма

$$H_{P_k} = -\sum_{i=1}^{|\Sigma|^k} p_i \log_{|\Sigma|^k} p_i = -\sum_{i=1}^{m} \left(\frac{c_i}{m}\right) \log_{|\Sigma|^k} \left(\frac{c_i}{m}\right).$$

Просто показать, что значение $H_{P_k} = 1$ может быть получено только при равночастотности *всех возможных* подслов длины k в исследуемом слове w.

Значение $H_{P_k} = 0$ означает, что все подслова, порожденные оператором $SH1(w,k)$ одинаковы и, следовательно, состоят из одного и того же символа.

ФУНКЦИЯ ЭНТРОПИИ СЛОВ (IV)

На втором этапе мы используем естественное расширение энтропии H_{P_k} путем введения функции $H(k) = H_{P_k}$, аргументом которой является длина подслова k, с областью определения: $1 \le k \le n$.

Значения функции $H(k)$ при фиксированном значении аргумента k вычисляются по формуле для H_{P_k} на основе вероятностной модели $M_k = \langle \Omega_k = \tilde{V} \cup \tilde{U},\ P_k(\cdot) \rangle$, полученной в результате применения оператора $SH1(w,k)$ к исходному слову w.

Заметим, что $H(n) = 0$, поскольку при $k = n$ мы наблюдаем всего одно слово с частотой 1. Значение $H(1)$ будет близко к 1, в случае, если в исследуемом слове частотная встречаемость символов алфавита будет приблизительно одинакова.

Тем самым в целом функция $H(k)$ будет невозрастающей функцией на области ее определения — мы можем характеризовать ее как «убывающую по совокупности».

КОНЕЧНАЯ РАЗНОСТЬ ФУНКЦИИ ОЦЕНКИ ЭНТРОПИИ СДВИГОВ

Интерес представляет изучение характера убывания значений $H(k)$ с ростом аргумента. Поскольку функция $H(k)$ — «убывающая по совокупности», рассмотрим конечную разность функции $H(k)$, взятую с обратным знаком (инверсная конечная разность)

$$\Delta H(k) = -(H(k+1) - H(k)) = H(k) - H(k+1),\ k = \overline{1, n-1}.$$

По определению функции $H(k)$ значения $\Delta H(k)$ ограничены, и $0 \le \Delta H(k) \le 1$, но поведение $\Delta H(k)$ может быть достаточно сложным. Однако функция $H(k)$ не может долго «держать единицу». Всего в слове длины n мы имеем $n-k+1$ позиций окна сдвига 1. Тогда максимально возможная длина подслова при котором еще можно наблюдать полное разнообразие подслов, определяется из уравнения $|\Sigma|^{\hat{k}} = n - \hat{k} + 1$, что с учетом целочисленности $\hat{k}$ приводит к уравнению $\hat{k} = \lfloor \log_{|\Sigma|}(n - \hat{k} + 1) \rfloor$. В предположении, что $\hat{k} << n$, пороговое значение $\hat{k} \approx \lfloor \log_{|\Sigma|} n \rfloor$.

МОДЕЛЬНЫЙ ПРИМЕР (I)

Рассмотрим модельный пример для введенных функций. Пусть $w = (abbaab)_4$, $|w| = 24$ — периодическое слово в алфавите $\Sigma = \{a, b\}$, тогда $H(k)$:

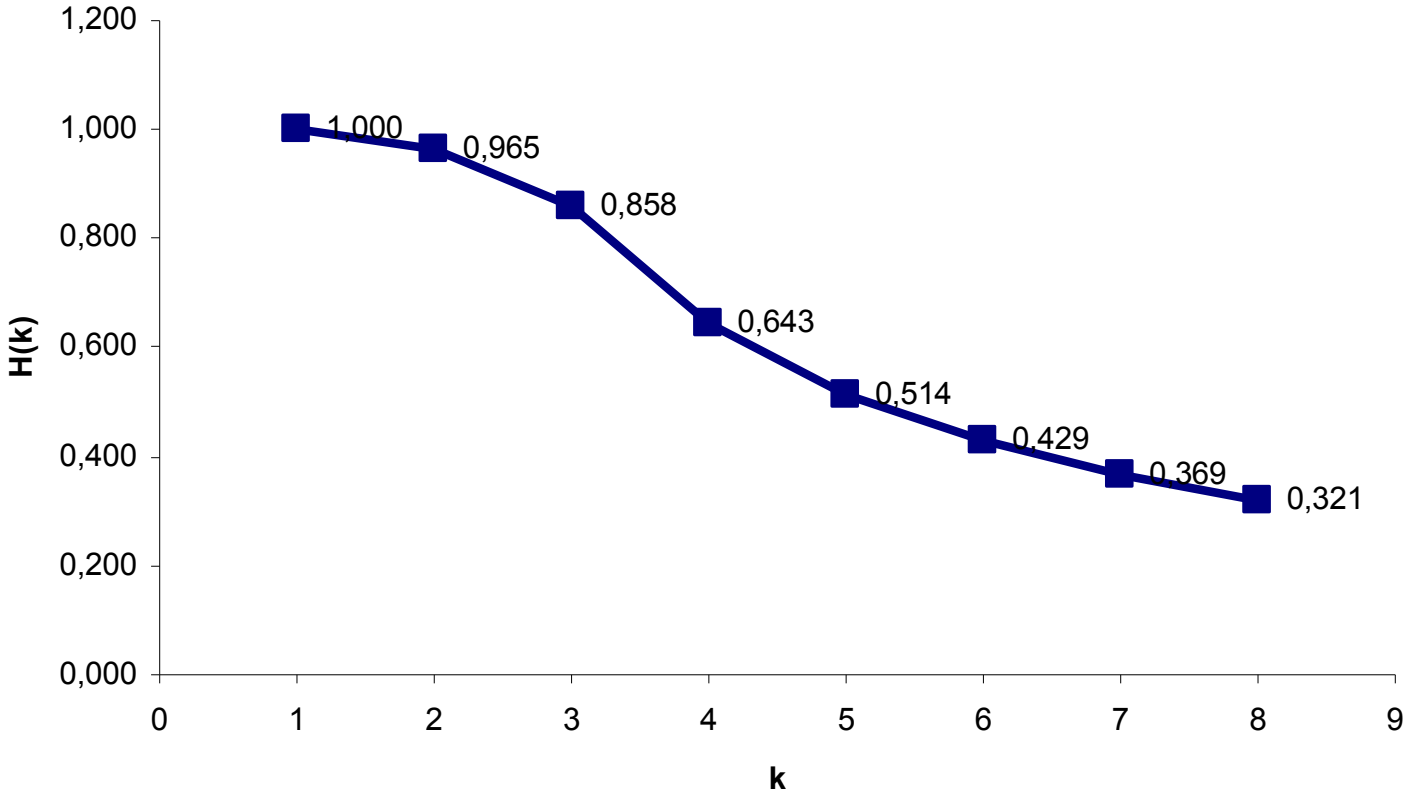

Рис.1. График функции энтропии слов $H(k)$ для модельного слова.

МОДЕЛЬНЫЙ ПРИМЕР (II)

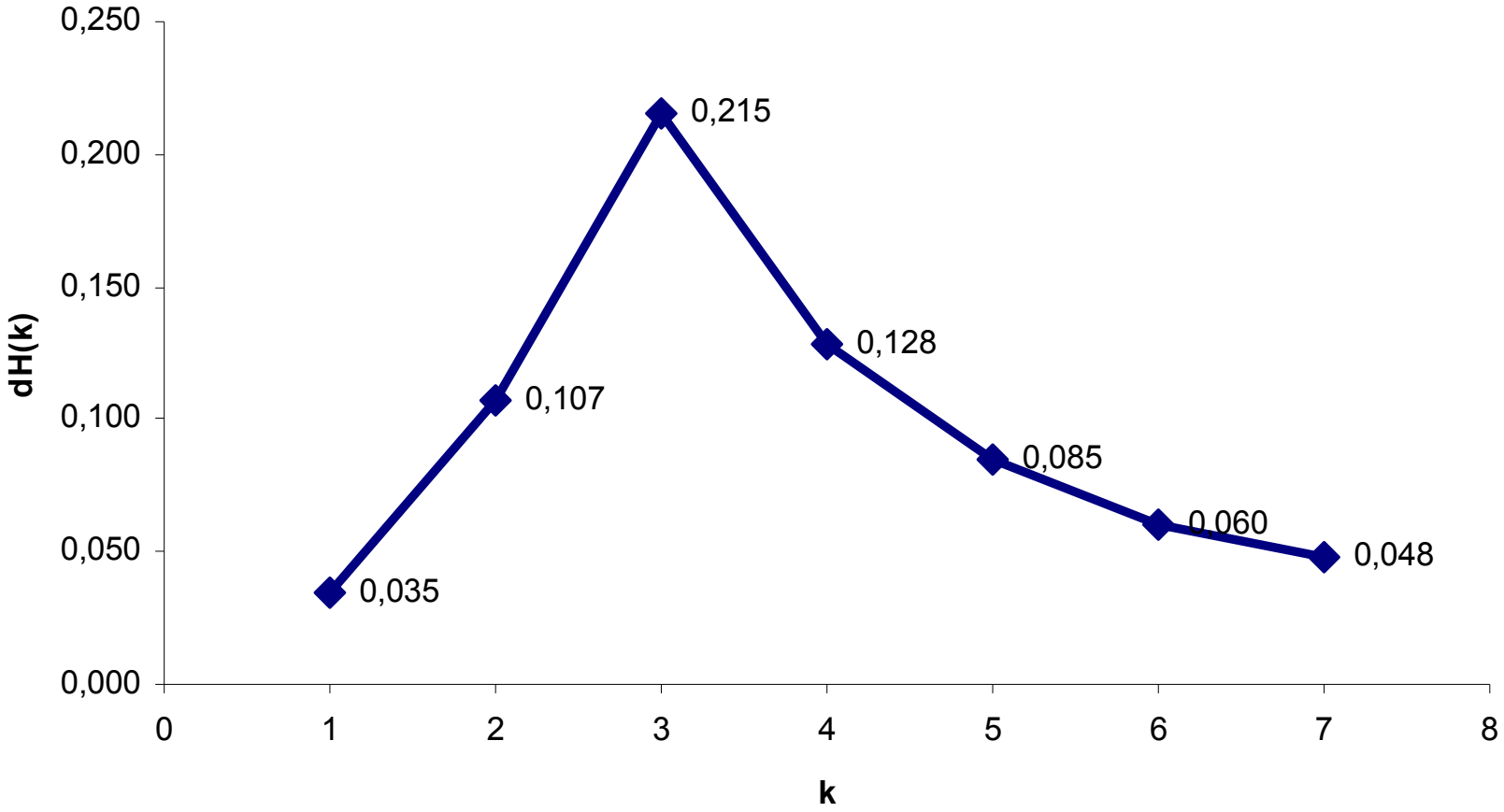

Рис. 2. График функции инверсной конечной разности $\Delta H(k)$ для модельного слова.

ПИКОВЫЕ ХАРАКТЕРИСТИКИ ФУНКЦИИ ЭНТРОПИИ СЛОВ

На основе исследования функции $\Delta H(k)$ мы, совместно с д.ф.-м.н. Ю. Г. Сметанин-ным, вводим следующие *пиковые характеристики функции энтропии слов* (компонен-ты меры символьного разнообразия):

Рассмотрим функционал, определенный на множестве слов W

$$\mu_s(w): W \to \aleph \times (0,1).$$

Формально, пусть

$$k^* = \arg\max_{1 \le k \le n} \Delta H(k),$$

тогда

$$\mu_s(w) = \left(k^*, \Delta H\left(k^*\right)\right).$$

Значение k^* будем называть *пиковой шириной окна* (оператора сдвига 1), а значе-ние $\Delta H\left(k^*\right)$ — *пиковым значением функции энтропии слов*.

КЛАСТЕРНОЕ ПРОСТРАНСТВО НА ОСНОВЕ ПИКОВЫХ ХАРАКТЕРИСТИК ФУНКЦИИ ЭНТРОПИИ СЛОВ

Начальным этапом решения задачи кластеризации объектов является построение координатного кластерного пространства. Как правило, признаки, которые определяют оси кластерного координатного пространства, задаются экспертами в данной проблемной области. Более независимый от предметных областей подход состоит в создании набора признаков, имеющих достаточно универсальный характер. Этот подход, очевидно, предполагает, что используемые признаки основаны на исследовании некоторых универсальных кодов объектов, например, — конечных слов над конечным алфавитом, полученных на основе символьного кодирования.

В рамках такой универсализации предлагается использовать кластерное координатное пространство, осями которого являются пиковые характеристики функции энтропии слов. В этом случае координатами точки исследуемого объекта в двумерном пространстве являются значения k^* и $\Delta H\left(k^*\right)$.

ПРИМЕНЕНИЕ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ СЕМЕЙСТВ РАСТЕНИЙ

В качестве примера построения кластерного пространства на основе введенных пиковых характеристик были выбраны геномы растений. Были исследованы представители семейства Пасленовых и семейства Капустных (Крестоцветных).

Поскольку $\hat{k} \approx \lfloor \log_{|\Sigma|} n \rfloor$, то для определения $\mu_s(w)$ достаточно вычислить функцию энтропии слов для значений ширины окна несколько превышающих $\hat{k}$. В реальном эксперименте вычисления проводились для аргумента $H(k)$ от 1 до $2\hat{k}$.

В ходе работы использовались данные из следующих генбанков:

— NCBI www.ncbi.nlm.nih.gov/genbank/;

— ENA http://www.ebi.ac.uk/ena;

— DDBJ http://www.ddbj.nig.ac.jp/.

Экспериментальное исследование выполнено студенткой ФКН НИУ ВШЭ А. С. Пестовой.

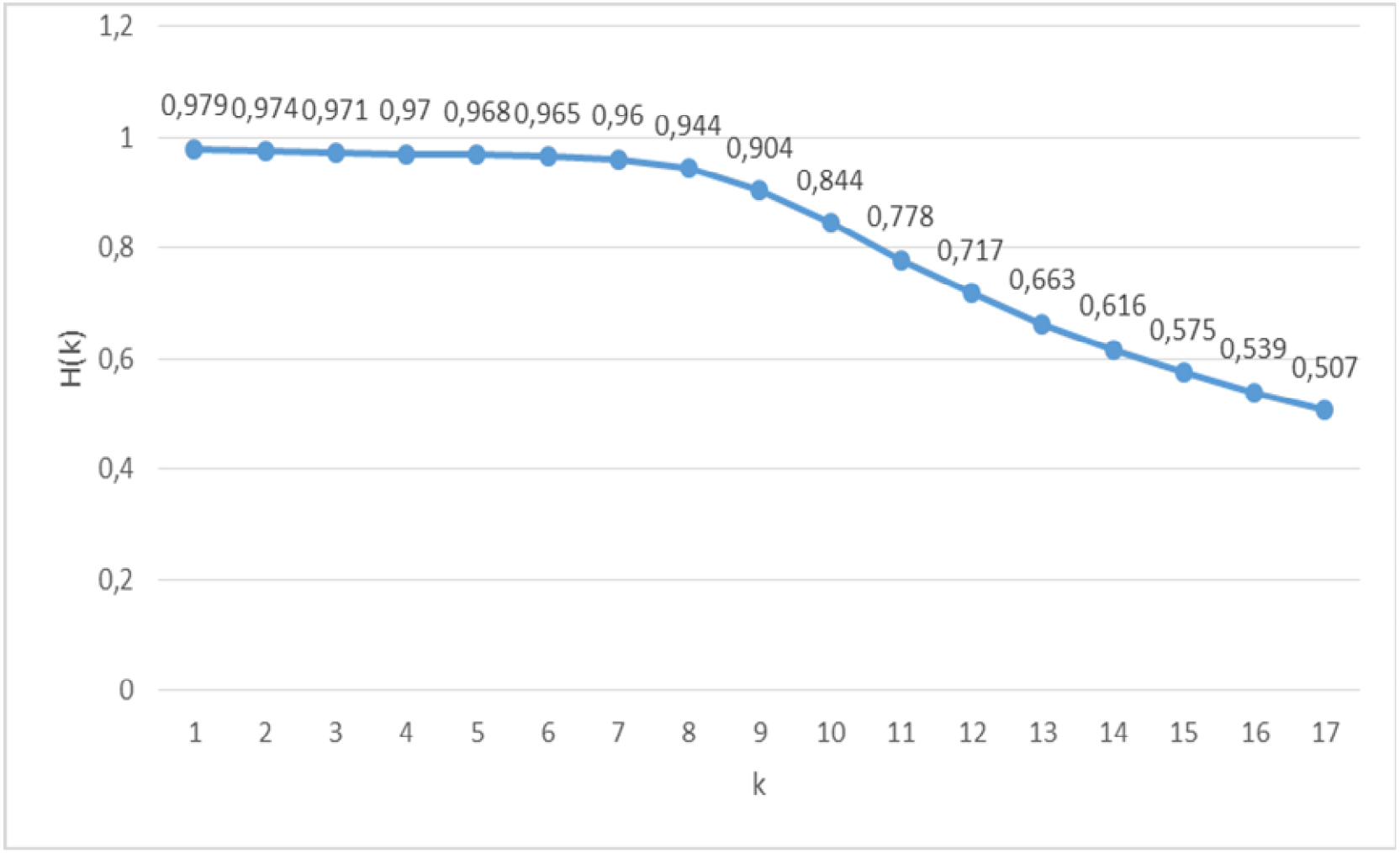

Рис.3. График функции $H(k)$ для генома паслена черного.

Далее мы интерпретировали значения $\mu_s(w) = \left(k^*, \Delta H\left(k^*\right)\right)$ как координаты точки в пространстве кластеризации для всех исследованных геномов.

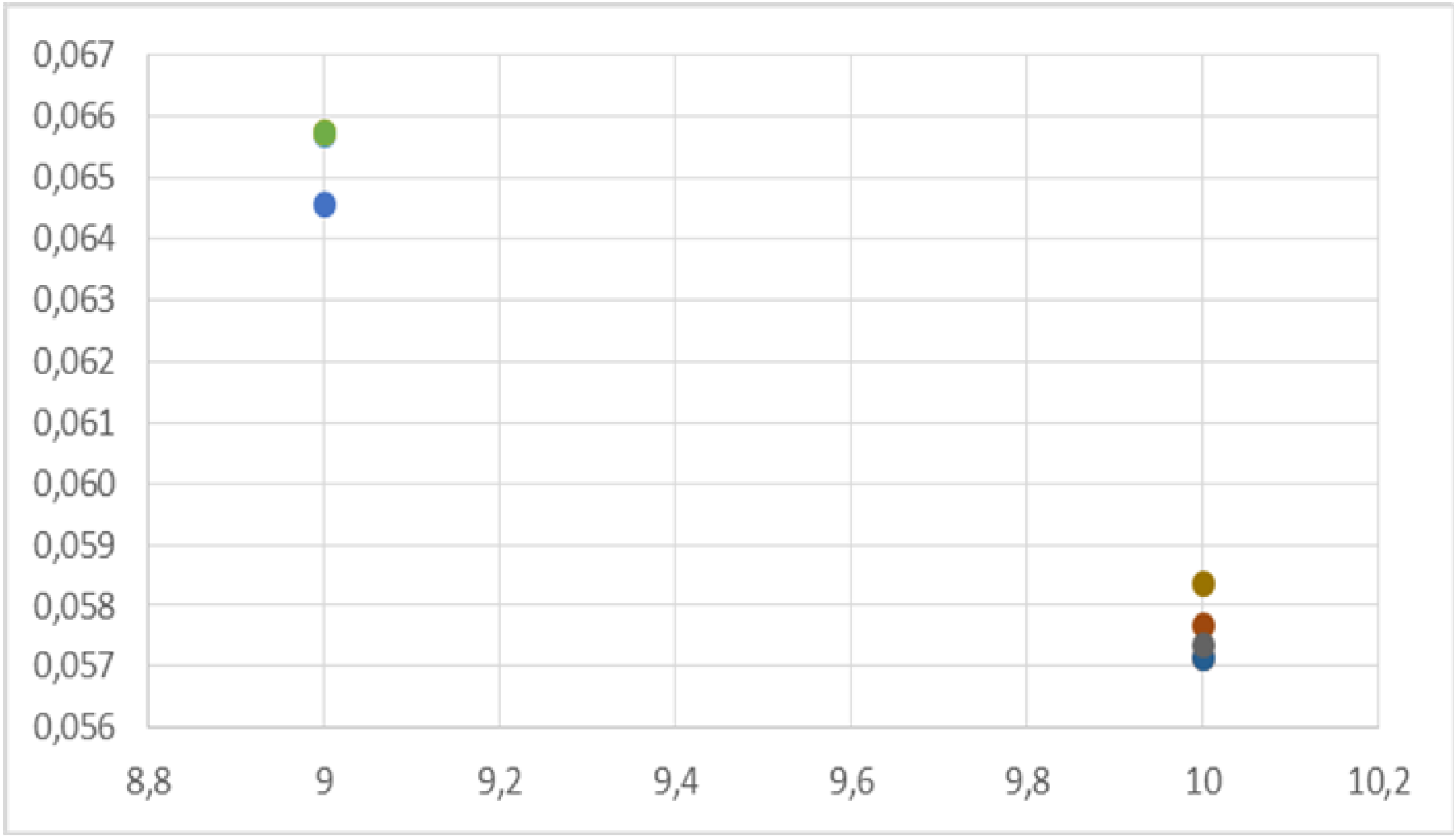

Рис. 4. Кластерное пространство геномов растений из 2-х семейств на основе значений пиковых характеристик функции энтропии слов $\mu_s(w)$.

Библиографический список

1. **Lind D., Marcus B.** An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge, UK. 1995. — 495 pp.

2. **Lothaire M.** Algebraic Combinatorics on Words. 2005. — 610 с.

3. **Сметанин Ю.Г., Ульянов М.В., Пестова А.С.** Энтропийный подход к построению меры символьного разнообразия слов и его применение к кластеризации геномов растений // Математическая биология и биоинформатика. 2016. Т. 11. № 1. С. 114–126.

4. **Сметанин Ю.Г., Ульянов М.В.** Мера символьного разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов // Бизнес-информатика. 2014. № 3(29). С. 40–46.

СПАСИБО ЗА ВНИМАНИЕ!