АНАЛИЗ ПОСЛЕДОВАТЕЛЬНОСТЕЙ И ТАБЛИЦ ДАННЫХ
(Вычислительные системы)

1994 год Выпуск 150

УДК 519.766

ОБУЧЕНИЕ ПРИ РАСПОЗНАВАНИИ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

П.Г.Кузнецов

Введение

Методы распознавания образов широко применяются в научных исследованиях для введения эмпирических закономерностей в слабо формализованных системах [1]. Методы и алгоритмы распознавания существенно зависят от физической природы распознаваемых объектов или явлений. Большой класс объектов распознавания, особенно на верхних уровнях в иерархических распознающих системах, представляется в виде символьных последовательностей. Особенности символьных последовательностей, меры сходства и методы анализа подробно рассмстрены в работах [2-7].

В дальнейшем предполагается, что для каждого класса образов существует эталонная символьная последовательность $\varphi = x_1 x_2 \dots x_n$, $x_j \in E$, $j = \overline{1,n}$, где n - длина символьной последовательности; $E = \{e_i\}$, $i = \overline{1,N}$, - алфавит, состоящий из N символоз. Наблюдаемые реализации обучающей выборки образуются в результате воздействия на эталонную символьную последовательность допустимых преобразований в виде выпадений, вставок и замещений символов, т.е. последовательности обучающей выборки $\varphi_{ов}$ представлены в виде слов произвольной длины n_s: $\varphi_{ов} = \{\varphi_s\}$, $s = \overline{1,v}$, $\varphi = \{x_{sj}\}$, $j = \overline{1,n_s}$.

Возможны два подхода к обучению.

1. Можно использовать реализации обучающей выборки как эталоны и для распознавания применять известные решающие правила типа k - ближайших соседей. При изменении k от единицы до v происходит переход от локальной меры сходства распознаваемой реализации символьной последовательности с обучающей выборкой к глобальной. Однако использование правила k - ближайших соседей требует запоминания всей обучающей выборки и сравнения поступившей реализации со всеми членами обучающей выборки, что существенно увеличивает объем памяти для хранения эталонов и время распознавания.

2. При обучении в метрических пространствах широко применяется центроидная аппроксимация обучающей выборки в виде оценки вектора математического ожидания. Для символьных последовательностей аналогом такой центроидной реализации является эталонная символьная последовательность φ. Таким образом, возникает задача оценки φ по $\varphi_{ов}$. Известна [6,7] постановка такой задачи и некоторые алгоритмы решения. В частности, в [6] описывается последовательный алгоритм грубой начальной оценки и итерационный алгоритм улучшения этой оценки, не гарантирующий однако достижения глобального экстремума. Кроме того трудно прогнозировать число шагов итерационного процесса, т.е. оценить трудоемкость алгоритма.

1. Постановка задачи

В байесовской постановке задача обучения (оценки) может быть записана в виде

$$\varphi^* = \arg\max_{\varphi} P(\varphi/\varphi_{ов}), \tag{1}$$

где $P(\varphi/\varphi_{ов})$ - апостериорная вероятность эталонной символьной последовательности при заданной обучающей выборке $\varphi_{ов}$. Для независимых и равновероятных реализаций обучающей выборки байесовский критерий (1) переходит в критерий максимального правдоподобия

$$\varphi^* = \arg\max_{\varphi} \prod_{s=1}^{v} P(\varphi_s/\varphi), \qquad (2)$$

где $P(\varphi_s/\varphi)$ - условная вероятность реализации φ_s при заданной эталонной последовательности φ. Для решения уравнения (2) требуется оценка φ и вероятностей искажений в виде вектора вероятностей выпадений символов $P_D = \{P_D(x_j)\}$, $j = \overline{1,n}$; вектора вероятностей вставок символов $P_J = \{P_J(x_j)\}$, $j = \overline{1,n}$, и матрицы вероятностей замещения символов $P_s = \| P(e_k/x_j = e_i) \|_{n \times N}$, $j = \overline{1,n}$, $k = \overline{1,N}$, где $P(e_k/x_j = e_i)$ - вероятность замещения символа e_i в позиции j на символ e_k.

Во многих реальных задачах распознавания символьных последовательностей объем обучающей выборки недостаточен для оценки вероятностей искажений, поэтому в этом случае естественным путем, например, в соответствии с принципом простоты [1], приходим к метрике Левенштейна [5] или взвешенного расстояния Левенштейна. При сравнении двух символьных последовательностей выпадение символа в одной из них эквивалентно вставке этого символа в другую последовательность, поэтому веса выпадений и вставок можно считать одинаковыми и равными единице. Замещение символа можно рассматривать как комбинацию выпадения и вставки, поэтому вес замещения можно взять равным двойному весу выпадения (вставки). Для этого случая между расстоянием Левенштейна d и длиной ρ максимально длинной подпоследовательности двух символьных подпоследовательностей существует связь [2,3]:

$$d = |\varphi_1| + |\varphi_2| - 2\rho. \qquad (3)$$

В рассматриваемом случае аналог критерия (2) будет иметь вид

$$\varphi^* = \arg\min_{\varphi} d_\Sigma = \arg\min_{\varphi} \sum_{s=1}^{v} d(\varphi_s,\varphi), \tag{4}$$

где d_Σ - суммарное внутриклассовое расстояние Левенштейна. Из (3) и (4) имеем

$$\varphi^* = \arg\min_{\varphi}(vn - 2\sum_{s=1}^{v} \rho_s + \sum_{s=1}^{v} |\varphi_s|), \tag{5}$$

где ρ_s - длина максимально длинной подпоследовательности между φ_s и φ. Учитывая, что для заданной обучающей выборки последний член в (5) не зависит от φ, получим

$$\varphi^* = \arg\min_{\varphi}(vn - 2\sum_{s=1}^{v} \rho_s). \tag{6}$$

Пусть имеется эталонная символьная последовательность k-го приближения $\varphi^{(k)}$. Введем дополнительный символ в $\varphi^{(k)}$, тогда приращение суммарного расстояния $\Delta d = v-2r$, где r - число максимально длинных общих подпоследовательностей у φ и φ_s, длина которых увеличилась на единицу. Чтобы обеспечить выполнение условия $\Delta d < 0$ в $\varphi^{(k)}$, можно добавлять очередной символ, если длина максимально длинных общих подпоследовательностей увеличивается на единицу более чем для половины реализаций обучающей выборки, т.е. при $r > 0{,}5v$.

Рассмотрим среднее внутриклассовое расстояние Левенштейна. Из (5) имеем:

$$\bar{d} = \frac{1}{v}\sum_{s=1}^{v} d(\varphi_s,\varphi) = n + \frac{1}{v}\sum_{s=1}^{v}|\varphi_s| - \frac{2}{v}\sum_{s=1}^{v}\rho_s = n + \bar{n} - 2\bar{\rho},$$

где $\bar{n}$ - средняя длина последовательностей из обучающей выборки; $\bar{\rho}$ - средняя длина максимально длинных подпоследовательностей для всех пар (φ, φ_s), где $\varphi_s \in \varphi_{ов}$.

Таким образом, эталонную символьную последовательность φ можно рассматривать как последовательность, для которой минимизируется суммарная стоимость операций выпадений, вставок и замещений, необходимых для превращения каждой реализации обучающей выборки в эту последовательность.

2. Теоретическое обоснование алгоритма

Любую конечную цепочку символов $\varphi = \{x_j\}$, $j = \overline{1,n}$, $x_j \in E = \{e_i\}$, $i = \overline{1,N}$, будем называть словом над алфавитом E.

Пусть даны слова $\varphi_s = \{x_{sj}\}$, $s = \overline{1,v}$, $j = \overline{1,n_s}$. Назовем массивом Ф таблицу

$$\Phi = \left\{ \begin{matrix} x_{11} & x_{12} & \dots & x_{1n_1} \\ x_{21} & x_{22} & \dots & x_{2n_2} \\ & . & . & . \\ x_{v1} & x_{v2} & \dots & x_{vn_v} \end{matrix} \right\}.$$

Нитью $\tilde{x}$ символа $x \in E$ в массиве Ф будем называть любое подмножество одинаковых символов $x_{sj} = x$, $s = \overline{1,v}$, $j = \overline{1,n_s}$, содержащее не более одного символа из каждой строки. Будем говорить, что нить $\tilde{x}$ пересекается со строкой φ_s, если элемент этой строки x_{sj} входит в нить, т.е. $\tilde{x} \cap \varphi_s = x_{sj} = x$. Если нить не содержит никакого символа из строки φ_s, будем говорить, что они не пересекаются.

ПРИМЕР.

$$\Phi = \left\{ \begin{array}{ccccccccc} б & д & а & с & е & а & д & & \\ р & е & д & а & б & с & р & м & \\ р & м & а & е & о & н & а & & \\ б & д & а & м & е & к & с & д & а \\ & & | & | & | & & | & & | \\ & & \tilde{а} & \tilde{м} & \tilde{е} & & \tilde{с} & & \tilde{а} \end{array} \right\}.$$

Весом нити $\tilde{x}$ назовем число $W(\tilde{x}) = P\tilde{x} - Q\tilde{x}$, где $P\tilde{x}$ - число строк, с которыми пересекается нить $\tilde{x}$, $Q\tilde{x}$ - число строк,с которыми нить не пересекается.

Будем говорить, что $\tilde{x}_1 < \tilde{x}_2$, если для каждой строки массива Φ из того, что $x_{sj} = \tilde{x}_1 \cap \varphi_s$ и $x_{sk} = \tilde{x}_2 \cap \varphi_s$, выполняется $j < k$.

Набор нитей $\tilde{x}_1 \ldots \tilde{x}_n$ назовем кортежом,если $\tilde{x}_j < \tilde{x}_k$ тогда и только тогда, когда $j < k$. Другими словами, кортеж - это упорядоченное множество нитей. Всякому кортежу $\tilde{x}_1 \ldots \tilde{x}_n$ соответствует слово $x_1 \ldots x_n$.

Весом кортежа $\tilde{x}_1\tilde{x}_2 \ldots \tilde{x}_n = \tilde{\varphi}$ назовем число

$$W(\tilde{\varphi}) = \sum_{j=1}^{n} W(\tilde{x}_j).$$

Расстоянием слова φ до слов массива Φ назовем величину

$$d(\varphi,\Phi) = \sum_{s=1}^{v} d(\varphi,\varphi_s).$$

Слово φ назовем ближайшим к словам $\Phi = \{\varphi_s\}, s = \overline{1,v}$, если расстояние $d(\varphi,\Phi)$ минимально.

Оказывается, что нахождение ближайшего слова к набору слов $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$, тесно связано с нахождением кортежа максимального веса. Справедлива следующая

ТЕОРЕМА 1. *Если вес кортежа* $\tilde{\varphi} = \{x_j\}$, $j = \overline{1,n}$, *максимален в массиве* $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$, *то соответствующее слово* $\varphi = \{x_j\}$ *- ближайшее к словам массива* $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$. *Обратно, если слово* $\varphi = \{x_j\}$, $j = \overline{1,n}$, *- ближайшее к словам* $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$, *то в массиве* Φ *существует кортеж* $\tilde{\varphi} = \{\tilde{x}_j\}$, $j = \overline{1,n}$, *максимального веса.*

Эта теорема сводит задачу нахождения ближайшего слова к задаче нахождения кортежа максимального веса. Доказательству теоремы предпошлем несколько лемм.

Пусть $\varphi = \{x_j\}$, $j = \overline{1,n}$, и $\psi = \{y_k\}$, $k = \overline{1,m}$, - два слова. Для каждого символа $x_j \in \varphi$, $y_k \in \psi$ вводим вес

$$P_{МДП}(x_j) = \begin{cases} 1, & \text{если } x_j \in МДП^{*)}, \\ -1, & \text{если } x_j \notin МДП; \end{cases}$$

$$P_{МДП}(y_k) = \begin{cases} 1, & \text{если } y_k \in МДП, \\ -1, & \text{если } y_k \notin МДП. \end{cases}$$

Легко доказывается

ЛЕММА 1.

$$d(\varphi,\psi) = |\varphi| - \sum_{k=1}^{m} P_{МДП}(y_k) = |\psi| - \sum_{j=1}^{n} P_{МДП}(x_j).$$

ЛЕММА 2. *Пусть* $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$, *- массив,* $\psi = \{y_k\}$, $k = \overline{1,m}$, *- произвольное слово,* $МДП_s$ *- максимально длинная подпоследовательность слов* ψ *и* φ_s. *Тогда*

$$d(\psi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - \sum_{s=1}^{v} \left(\sum_{k=1}^{m} P_{МДП_s}(y_k)\right).$$

Доказательство следует из леммы 1.

ЛЕММА 3. *Пусть* $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$, *- массив,* $\psi = \{y_k\}$, $k = \overline{1,m}$, *- произвольное слово. Тогда существует кортеж* $\tilde{\psi}' = \{\tilde{y}_{k_1} \tilde{y}_{k_2} \ldots \tilde{y}_{k_l}\}$ *из символов слова* ψ *такой, что*

$$d(\psi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}') + v(m-1).$$

---

*) МДП - максимально длинная подпоследовательность.

ДОКАЗАТЕЛЬСТВО. Найдем для каждого $\varphi_s \subset \Phi$ и ψ МДП$_s \subseteq \psi$, в которой номер членов МДП$_s$ наследуются из ψ. Составим кортеж $\tilde{y}_{k_1}\tilde{y}_{k_2}\ldots\tilde{y}_{k_l}$ из членов соответствующих МДП$_s$, $s = \overline{1,v}$. Тогда

$$d(\psi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - \sum_{s=1}^{m}\left[\sum_{k=1}^{m} P_{\text{МДП}_s}(y_k)\right] =$$

$$= \sum_{s=1}^{v} |\varphi_s| - \sum_{k=1}^{m}\left[\sum_{s=1}^{v} P_{\text{МДП}_s}(y_k)\right] =$$

$$= \sum_{s=1}^{v} |\varphi_s| - \sum_{q=1}^{l} W(\tilde{y}_{k_q}) - \sum_{y_k \neq y_{k_q}}\left[\sum_{s=1}^{v} P_{\text{МДП}_s}(y_k)\right].$$

Так как для каждого y_k: $y_k \neq y_{k_q}$, $P_{\text{МДП}_s}(y_k) = -1$, то

$$\sum_{y_k \neq y_{k_q}}\left[\sum_{s=1}^{v} P_{\text{МДП}_s}(y_k)\right] = -v(m-1).$$

Сумма $W(\tilde{y}_{k_1}\tilde{y}_{k_2}\ldots\tilde{y}_{k_l}) = \sum_{q=1}^{l} W(\tilde{y}_{k_q})$ - вес кортежа $\tilde{\psi}' = \tilde{y}_{k_1}\tilde{y}_{k_2}\ldots\tilde{y}_{k_l}$. Окончательно получим

$$d(\psi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}') + v(m-1).$$

ЛЕММА 4. *Если* $\tilde{\psi} = \{\tilde{y}_k\}$, $k = \overline{1,m}$ *- кортеж массива* $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$, *то для слова* $\psi = \{y_k\}$, $k = \overline{1,m}$, *справедливо*

$$d(\psi,\Phi) \leq \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}).$$

ДОКАЗАТЕЛЬСТВО

$$\sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}) = \sum_{s=1}^{v} |\varphi_s| - \sum_{k=1}^{m} W(\tilde{y}_k) =$$

$$= \sum_{s=1}^{v} |\varphi_s| - \sum_{k=1}^{m} \sum_{s=1}^{v} P(\tilde{y}_k \cap \varphi_s) =$$

$$= \sum_{s=1}^{v} \{|\varphi_s| - \sum_{k=1}^{m} P(\tilde{y}_k \cap \varphi_s)\},$$

где

$$P(\tilde{y}_k \cap \varphi_s) = \begin{cases} 1, \text{если } \tilde{y}_k \cap \varphi_s \neq \emptyset; \\ -1, \text{если } \tilde{y}_k \cap \varphi_s = \emptyset. \end{cases}$$

Но для каждого $s = \overline{1,v}$

$$|\varphi_s| - \sum_{k=1}^{m} P(\tilde{y}_k \cap \varphi_s) \geq d(\varphi_s,\psi).$$

Отсюда получим

$$\sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}) \geq \sum_{s=1}^{v} d(\varphi_s,\psi) = d(\psi,\Phi).$$

ЛЕММА 5. *Если слово* $\psi = \{y_k\}$, $k = \overline{1,m}$,- *ближайшее к массиву* $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$, *то существует кортеж* $\tilde{\psi} = \{\tilde{y}_k\}$, $k = \overline{1,m}$, *такой, что*

$$d(\psi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}).$$

ДОКАЗАТЕЛЬСТВО. Построим кортеж $\tilde{\psi}' = \{\tilde{y}_{k_1} \tilde{y}_{k_2} \ldots \tilde{y}_{k_1}\}$ для слова ψ как в лемме 3. Тогда

$$d(\psi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}') + v(m-1).$$

Отсюда следует, что $m = 1$. Действительно, иначе для слова $y_{k_1} y_{k_2} \ldots y_{k_l}$, $1 < m$, выполнялось бы

$$d[(y_{k_1} y_{k_2} \ldots y_{k_l}),\Phi] \le \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}') \le d(\psi,\Phi),$$

что противоречит тому, что ψ - ближайшее слово. Итак, наш кортеж $\tilde{y}_{k_1} \tilde{y}_{k_2} \ldots \tilde{y}_{k_l}$ содержит все символы слова $\psi = \{y_k\}$, $k = \overline{1,m}$, т.е. $\tilde{\psi} = \{\tilde{y}_k\}$, $k = \overline{1,m}$, и

$$d(\psi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}).$$

ДОКАЗАТЕЛЬСТВО теоремы 1. Пусть кортеж $\tilde{\varphi} = \{\tilde{x}_j\}$, $j = \overline{1,n}$, имеет минимальный вес в массиве $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$. Покажем, что слово $\varphi = \{x_j\}$, $j = \overline{1,n}$, - ближайшее к массиву Φ. В силу леммы 4,

$$d(\varphi,\Phi) \le \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\varphi}).$$

Пусть слово ψ - ближайшее к массиву Φ и $d(\psi,\Phi) < d(\varphi,\Phi)$. По лемме 5, для ψ существует кортеж $\tilde{\psi}$ такой, что

$$d(\psi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}).$$

Если $d(\psi,\Phi) < d(\varphi,\Phi)$, то имеем

$$\sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}) < \sum_{s=1}^{v} |\varphi_s| - W(\varphi)$$

или $W(\tilde{\psi}) > W(\tilde{\varphi})$, что противоречит максимальности веса кортежа $\tilde{\varphi}$ в условиях теоремы.

Пусть теперь слово $\varphi = \{x_j\}, j = \overline{1,n}$, - ближайшее к массиву $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$. По лемме 5 существует кортеж $\tilde{\varphi}=\{\tilde{x}_j\}$, $j = \overline{1,n}$, такой, что

$$d(\varphi,\Phi) = \sum_{s=1}^{v} |\varphi_s| - W(\varphi).$$

Покажем, что вес $W(\tilde{\varphi})$ максимален среди всех кортежей массива Φ. Пусть $\tilde{\psi}$ - кортеж максимального веса и $W(\tilde{\psi}) > W(\tilde{\varphi})$. По сказанному выше слово ψ - ближайшее к массиву Φ. По лемме 4

$$d(\psi,\Phi) \leq \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\psi}) < \sum_{s=1}^{v} |\varphi_s| - W(\tilde{\varphi}) = d(\varphi,\Phi).$$

Тогда $d(\psi,\Phi) < d(\varphi,\Phi)$, что противоречит тому, что φ - ближайшее слово. Теорема доказана.

ЗАМЕЧАНИЕ 1. Если $\tilde{\varphi} = \{\tilde{x}_j\}$, $j=\overline{1,n}$, - кортеж максимального веса в массиве $\Phi = \{\varphi_s\}$, $s = \overline{1,v}$, то мы можем без ограничения общности считать, что для пересечения кортежа $\tilde{\varphi}$ со строкой φ_s выполняются следующие условия: символ $x_1 = \tilde{x}_1 \cap \varphi_s$ имеет наименьший номер для этого символа в строке φ_s, т.е. это самый левый символ вида $x_1 \in E$; символ $x_j = \tilde{x}_j \cap \varphi_s$ имеет наименьший номер после символа $x_{j-1} = \tilde{x}_{j-1} \cap \varphi_s$ из всех символов x_j, расположенных правее символа $x_{j-1} \in E$.

ЗАМЕЧАНИЕ 2. Если $\tilde{\varphi} = \{\tilde{x}_j\}$, $j = \overline{1,n}$, - кортеж максимального веса, то для каждой нити $\tilde{x}_j$ ее вес $W(\tilde{x}_j) \geq 0$, т.е. нить пересекается не менее, чем с половиной строк массива Φ. Отбросив все нити нулевого веса, получим кортеж максимального веса, в котором вес каждой нити строго больше нуля.

3. Алгоритмы восстановления эталонной символьной последовательности

Итак, в силу сказанного выше,нахождение эталонной символьной последовательности или ближайшего к массиву слова сводится к нахождению кортежа максимального веса. При этом будем учитывать замечания 1 и 2.

Введем следующие обозначения. Пусть $\tilde{\varphi}_s = \{\tilde{x}_k\}$, $k = \overline{1,l}$ - произвольный кортеж массива Φ, тогда массив $\Phi(\tilde{\varphi}_l) =$ $= \Phi \setminus \{x_{sj}: \exists \tilde{x}_k \subset \tilde{\varphi}_l$ такое,что $\tilde{x}_k \cap \varphi_s \neq \emptyset$ и $j \leq k\}$ назовем рабочим массивом.

ПРИМЕР.

$$\Phi = \begin{array}{lcllllll} \varphi_1 & = & с & а & б & е & & \\ \varphi_2 & = & д & е & а & б & с & \\ \varphi_3 & = & а & д & с & е & & \Phi_{раб} \\ \varphi_4 & = & б & с & а & д & е & б \\ \varphi_5 & = & д & а & б & е & с & \\ \tilde{\varphi}_1 & = & & \tilde{а} & \tilde{б} & & & \end{array}$$

Строки массива $\Phi_{раб}$ будем обозначать через $\varphi_{sраб}$.

$$\Phi_{раб} = \begin{array}{lcl} \varphi_{1раб} & = & е \\ \varphi_{2раб} & = & д\ е\ а\ б\ с \\ \varphi_{3раб} & = & д\ с\ а \\ \varphi_{4раб} & = & с\ а\ д\ е\ б \\ \varphi_{5раб} & = & е\ с \end{array}$$

АЛГОРИТМ 1.

1. Положим $\Phi_{раб} = \Phi$.

2. Нить $\tilde{x}$ массива $\Phi_{раб}$ будем называть левой, если для каждой строки $\varphi_{sраб} \in \Phi_{раб}$ $\tilde{x} \cap \varphi_{sраб}$ имеет наименьший номер

среди всех вхождений данного символа в строку $\varphi_{\dot{s}раб}$. Пусть $LEV(\Phi_{раб})$ - множество всех левых нитей массива $\Phi_{раб}$ с весом $W(\tilde{x}) > 0$.

3. Проверяем, является ли нить $\tilde{x} \in LEV(\Phi_{раб})$ минимальной во множестве $LEV(\Phi_{раб})$, т.е. существует ли нить $\tilde{y} \in LEV(\Phi_{раб})$ такая, что $\tilde{y} < \tilde{x}$. Если такой нити не существует, то $\tilde{x}$ - мини - мальная левая нить. Пусть $MIN\ LEV(\Phi_{раб})$ - множество всех минимальных левых нитей.

4. Каждую нить из $MIN\ LEV(\Phi_{раб})$ назовем кортежем единичной длины и вычислим его вес.

5. Пусть построены все кортежи длины $j-1$ и вычислены их веса. Для каждого кортежа $\tilde{\varphi}_{j-1} = \{\tilde{x}_1\tilde{x}_2\ldots\tilde{x}_{j-1}\}$ полагаем $\Phi_{раб} = \Phi(\tilde{\varphi}_{j-1})$ и переходим к п.2.

6. Каждую нить из $MIN\ LEV(\Phi_{раб})$ присоединяем к кортежу $\tilde{\varphi}_{j-1}$ и вычисляем его вес, добавляя к весу $W(\tilde{\varphi}_{j-1})$ вес нити $W(\tilde{x}_j)$ из $MIN\ LEV(\Phi_{раб})$.

После окончания работы алгоритма получим семейство максимально длинных или неудлиняемых кортежей. Среди них будут кортежи максимального веса. Таким образом, получим все слова, ближайшие к массиву Φ, или эталонные символьные последовательности.

АЛГОРИТМ 2.

Предыдущий алгоритм гарантирует нахождение оптимальной эталонной символьной последовательности, что достигается за счет большого перебора возможных вариантов. Оценим максимальную мощность массива $LEV(\Phi_{раб})$ всех левых нитей с положительным весом. Нетрудно видеть, что может быть одна левая нить вида $\tilde{x}$ с максимальным весом $W(\tilde{x}) = v$, v нитей с весом $v-1$, C_v^2 нитей с весом $v-2$ и т.д. Отсюда

$$M = |LEV(\Phi_{раб})| = N \sum_{\alpha=1}^{\beta} C_v^{\alpha},$$

где $\beta = (v/2)-1$ для четных v и $\beta = (v-1)/2$ для нечетных v.

Пусть число минимальных левых нитей $L = |MIN\ LEV(\Phi_{раб})| = \lambda M$, где $\emptyset < \lambda \leq 1$. Тогда число вариантов на шаге j алгоритма равно L^j. Для уменьшения вычислительной сложности предлагается следующий эвристический алгоритм. Введем некоторые определения.

1. Текущий массив символов $\Phi_{тек}$ - это таблица символов из v строк, каждая строка $\varphi_{sтек}$ в которой начинается с символа, следующего за последним символом, вошедшим в восстанавливаемую эталонную символьную последовательность в соответствующей строке φ_s массива Φ. С правой стороны текущий массив ограничен некоторым очередным столбцом символов массива Φ. Некоторые строки $\varphi_{sтек}$ могут быть пустыми.

2. Расширение текущего массива - добавление к нему справа очередного столбца символов из массива Φ. В качестве начального приближения текущего массива для первого шага служит первый столбец массива Φ. Частота встречаемости Px символа x в текущем массиве равна количеству строк этого массива,в которых данный символ встречается хотя бы один раз (число строк, с которыми пересекается нить $\tilde{x}$).

Алгоритм 2 выглядит следующим образом.

1. Определяем частоты встречаемости Px для всех символов $x \in E$ в текущем массиве символов и находим $\max\limits_{x} Px$. Если $\max Px > 0,5v$, включаем соответствующий символ в искомую эталонную символьную последовательность. Если в некоторой строке $\varphi_{sтек}$ символ x встречается несколько раз,выбираем для включения в φ самый левый символ.

2. После включения символа $x \in \varphi$ корректируем текущий массив и переходим к п.1

3. Если $\max Px \le 0{,}5v$, производим расширение текущего массива и переходим к п.1.

4. Выполняем пп. 1-3 до окончания массива.

Если условие $\max Px > 0{,}5v$ выполняется для нескольких символов, то для включения в φ выбираем символ с минимальной суммой Sx номеров его позиций в строках текущего массива $\Phi_{\text{тек}}$. (Примечание: если в строке несколько символов данного типа, то суммируется номер позиции самого левого вхождения символа).

На рис.1 приведен пример обучающей выборки $\varphi_{\text{ов}} = \Phi$, состоящей из 9 реализаций. Каждая реализация получена путем равновероятного выбора одного из пяти символов в первом столбце и одного из четырех оставшихся в последующих столбцах. На массиве Φ показаны нити $\tilde{x}$, образующие кортеж $\tilde{\varphi} = \{y_k\}$, $k = \overline{1,20}$, максимального веса. На рис.2 показано несколько первых шагов алгоритма 2 восстановления эталонной символьной последовательности φ.

После восстановления φ можно произвести оценку вероятностей искажений эталонной символьной последовательности φ. Рассматривая реализации обучающей выборки как результат действия заданных допустимых преобразований на полученную оценку эталонной символьной последовательности, получим

$$\tilde{P}_D = \frac{n_D}{v|\varphi|}, \quad \tilde{P}_J = \frac{n_J}{v|\varphi|}, \quad \tilde{P}_s = \frac{n_s}{v|\varphi|},$$

где n_D, n_J, n_s - количество выпадений, вставок и замещений в обучающей выборке, $\tilde{P}_D$, $\tilde{P}_J$, $\tilde{P}_s$ - оценки вероятностей выпадений, вставок и замещений.

В частности, в нашем примере для модели выпадений и вставок $\tilde{P}_D = 0{,}36$, $\tilde{P}_J = 0{,}47$; для модели выпадений, вставок и замещений $\tilde{P}_D = 0{,}18$, $\tilde{P}_J = 0{,}29$, $\tilde{P}_s = 0{,}18$.

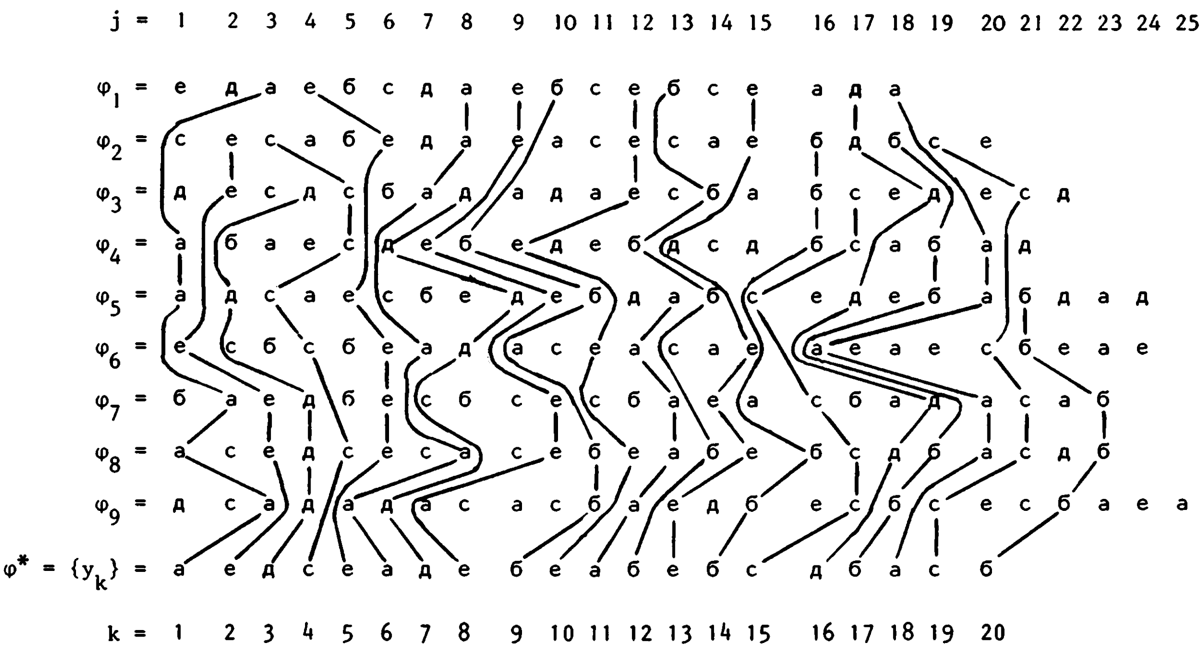

Рис.1 . Обучающая выборка для символьных последовательностей (массив Φ)

	Первое приб-лиже-ние	Расши-рение	Расши-рение	Коррек-ция	Кор-рек-ция	Расши-рение	Коррек-ция	Расши-рение	Кор-рек-ция	Расши-рение	Кор-рек-ция	Расши-рение
	е с д а а е б а д	е д с е д е а б а д е с б а а с д с	е д а с е с д е с а б а а д с е с б б а е а с е д с а	 с е с д е с б а д с е.с б е с е 	 с с б а д с с б 	е с а с д б а е д с а е б с д д д	е с а б а е с а е б с 	е б с а б с б а е с с а е е б с б б с а	е б а б а е б б а	е б с а б е б д а е с б е б е е а д	б с б д с а д	б с д д б а д е с б а с с а д а
x												
Px а	3	4	6 Sa=11	1	1	3	3	4	3	3	1	3
б	1	2	3	2	2	2	2	5	4	5	2	3
с	1	4	6 Sc=13	5 Sc=9	4	4	3	6	0	2	2	4
Sx д	2	4	4	2	1	5	0	0	0	2	2	4
е	2	4	6 Se=12	5 Se=8	0	3	3	4	2	6	0	2
max Px	3	4	6	5	4	5	3	6	4	6	2	4
φ*			а	е	д		с		е			

Рис.2. Первые шаги алгоритма 2

Алгоритм 2 использовался для решения модельных задач распознавания символьных последовательностей и реальных задач распознавания слов произвольного диктора [8-10]. При распознавании слов использовалось амплитудно-формантное описание речевого сигнала, состоящее из восьми параметров: огибающие речевого сигнала в низкочастотной и высокочастотной областях спектра, в областях первой и второй формант и общая огибающая; частота нулевых переходов речевого сигнала в областях первой и второй формант и в высокочастотной области [8].

Обучающая выборка состояла из 100 произнесений для 50 дикторов каждого из трех словарей управления техническими системами. Из них путем объединения был сформирован четвертый словарь. Таким образом, распознавались слова четырех словарей объемом от 13 до 34 слов.

В соответствии с заданными порогами слова сегментировались на 8 видов сегментов, отвечающих групповым фонемам. Каждая символьная последовательность характеризовалась вектором длительности символов. При обучении длительности символов не учитывались. Обучающая выборка из 100 реализаций для каждого слова словаря разбивалась на 5 кластеров алгоритмом, использующим максимальное расстояние [11]. Число кластеров выбиралось эмпирическим путем. Для каждого кластера вычислялась эталонная символьная последовательность с помощью алгоритма 2 и оценивались вероятности искажений символов. Кроме того, вычислялись средние длительности символов эталонных последовательностей каждого кластера.

Сравнение текущей реализации с эталонными символьными последовательностями осуществлялось методами динамического программирования на взвешенном графе, где веса дуг пропорциональны логарифмам оценок вероятностей искажений. При каждом сравнении веса дуг корректировались в пределах 20% от начальной величины в зависимости от разницы в средней длительности символа эталонной последовательности и соответствующего символа текущей реализации. Решение принималось по правилу двух ближайших соседей. При распознавании слов, произнесенных 10 дикторами, не участвовавшими в обучении, оценки надежности распознавания оказались равны 97-99% при средней вероятности отказов от распознавания, равной 1,2%. Это соответствует результатам, полученным классическими методами динамического программирования в работах других исследователей, но время распознавания существенно (в 20-30 раз) меньше.

Литература

1. ЗАГОРУЙКО Н.Г. Информатика и МОЗ //Проблемы обработки информации. - Новосибирск, 1983. - Вып. 100: Вычислительные системы. - С. 34-45.

2. ГУСЕВ В.Д. Характеристики символьных последовательностей //Машинные методы обнаружения закономерностей. - Новосибирск, 1981. - Вып. 88: Вычислительные системы. - С. 112-123.

3. ГУСЕВ В.Д. Механизмы обнаружения структурных закономерностей в символьных последовательностях //Проблемы обработки информации. - Новосибирск, 1983. - Вып. 100: Вычислительные системы. - С. 47-66.

4. REUHKALA E. Recognition of String of Discrete Symbols with Special Application to Isolated Word Recognition //Acta Polytechn. Scandinavica Math. and Comput. Sci. - 1983. - N.38. - 92 p.

5. ЛЕВЕНШТЕЙН В.И. Бинарные коды, способные к исправлению пропусков, вставок и замен //Докл. АН СССР. - 1965. - Т. 163. - С. 845-848.

6. МОТТЛЬ В.В., МУЧНИК И.Б. Лингвистический анализ экспериментальных кривых //ТИИЭР. - 1979. - Т. 67, Т 5. - С.12-39.

7. WATERMAN M.S. Consensus Pattern in Sequences //Mathematical methods for DNA sequences /Ed. M.S. Waterman. CRS Press, Jnc. USA, 1989. - P. 93-116.

8. КУЗНЕЦОВ П.Г., ПОЗДЕЕВ В.С. Выбор системы признаков для малогабаритного устройства распознавания ограниченного набора слов //Автоматическое распознавание слуховых образов (АРСО-13). - Новосибирск: Институт математики СО АН СССР, 1984. - Ч.2. - С. 101-102.

9. КУЗНЕЦОВ П.Г., ХАТБУЛЛИН Р.А. Обучение при распознавании символьных последовательностей //Тез. докл. 16-го Всесоюзного семинара (АРСО-16) (1; 1991; Москва). - Москва: МГУ, 1991. - С. 46-47.

10. КУЗНЕЦОВ П.Г., ПОЗДЕЕВ В.С. Распознавание слов для произвольного диктора //Речевая информатика: Сб.науч.трудов /Институт кибернетики АН УССР. - Киев, 1989. - С. 101-104.

11. ТУ Дж., ГОНСАЛЕС Р. Принципы распознавания образов.-М.: Мир, 1978. - 416 с.

Поступила в ред.-изд.отд.

1 ноября 1992 года