ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2018 Управление, вычислительная техника и информатика № 42

УДК 004.852
DOI: 10.17223/19988605/42/9

В.Е. Уваров

РАСПОЗНАВАНИЕ НЕПОЛНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ОПИСЫВАЕМЫХ СКРЫТЫМИ МАРКОВСКИМИ МОДЕЛЯМИ, В ПРОСТРАНСТВЕ ПЕРВЫХ ПРОИЗВОДНЫХ ОТ ЛОГАРИФМА ФУНКЦИИ ПРАВДОПОДОБИЯ

Предлагается метод распознавания неполных последовательностей, который заключается в классификации последовательностей в пространстве признаков, образованном первыми производными от логарифма функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал распознаваемую неполную последовательность. В качестве классификатора в предлагаемом методе применяется метод опорных векторов.

Ключевые слова: скрытые марковские модели; машинное обучение; последовательности; пропущенные наблюдения; неполные данные.

Теория скрытых марковских моделей (СММ) была представлена еще в 1970-х гг. Л. Баумом и его коллегами [1]. Изначально СММ применяли для распознавания речи. В конце 1980-х гг. СММ начали использовать в биоинформатике, например для обработки цепочек ДНК. Тем не менее наибольшую популярность СММ обрели после 1990-х гг., и данная тенденция продолжается и в настоящее время, что можно подтвердить частотой упоминания термина «hidden Markov model» в публикациях [2].

Однако в теории СММ остается малоизученная область, касающаяся вопросов применения СММ для анализа неполных данных. Данные вопросы являются актуальными, поскольку в сложных системах, например при приеме данных с космических и авиационных аппаратов, а также других источников, приходится иметь дело с потоками данных от различных датчиков в сложной помеховой обстановке, когда возможно пропадание информации или ее искажение. В настоящей работе рассматривается такой случай неполных данных, как наличие пропусков в распознаваемых последовательностях. Такие последовательности с пропусками будем называть неполными. В рассматриваемой ситуации пропуски не генерируются самим случайным процессом, описываемым СММ, а возникают в произвольных местах последовательностей за счет внешних условий.

Данная статья является продолжением исследований по распознаванию последовательностей, описываемых СММ, проводимых на кафедре теоретической и прикладной информатики Новосибирского государственного технического университета [3]. Отличие проводимого исследования заключается в том, что распознаваемые последовательности могут содержать пропуски.

1. Описание скрытой марковской модели

1.1. Структура скрытой марковской модели

Скрытой марковской моделью называют модель, описывающую случайный процесс, находящийся в каждый момент времени $t \in \{1, ..., T\}$ в одном из N скрытых состояний $s \in \{s_1, ..., s_N\}$ и в новый момент времени переходящий в другое или в прежнее состояние согласно некоторым вероятностям переходов. Состояния считаются скрытыми, однако они проявляются в тех или иных особенностях наблюдаемых последовательностей. В данной работе рассматриваются СММ с непрерывной плотностью распределения наблюдений, когда в общем случае многомерные наблюдения – это векторы действительных чисел. Значения наблюдаемых величин при условии того, что СММ находится в конкретном скрытом состоянии, подчиняются некоторым вероятностным законам. В случае СММ с непрерывной

плотностью распределения наблюдений эти вероятностные законы описываются функциями условной плотности распределений наблюдений.

Рассмотрим параметры, которыми можно полностью задать конкретную СММ. Обозначим скрытое состояние, в котором находится описываемый СММ процесс в момент *t*, символом q_t, многомерное наблюдение, которое он сгенерировал в момент времени *t*, – символом $\boldsymbol{o}_t$, а многомерное наблюдение, не привязанное к конкретному времени, – символом $\boldsymbol{o}$. СММ с непрерывной плотностью распределения характеризуется вектором вероятностного распределения начального скрытого состояния $\Pi = \left\{\pi_i = p\left(q_1 = s_i\right),\ i = \overline{1,N}\right\}$, матрицей вероятностей переходов из одного скрытого состояния в другое $A = \left\{a_{ij} = p\left(q_{t+1} = s_j \mid q_t = s_i\right),\ \ i, j = \overline{1,N}\right\}$, а также функциями условной плотности распределений многомерных наблюдений $B = \left\{b_i\left(\boldsymbol{o}\right) = f\left(\boldsymbol{o} \mid q = s_i\right), i = \overline{1,N}, \boldsymbol{o} \in R^Z\right\}$ [4]. В данной работе в качестве функций условной плотности распределения наблюдений рассматривается смесь многомерных нормальных распределений: $b_i(\boldsymbol{o}) = \sum_{m=1}^{M} \tau_{im} g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}),\ \ i = \overline{1,N}, \boldsymbol{o} \in R^Z$, где *M* – число компонент в смеси для каждого скрытого состояния, $\tau_{im} \geq 0$ – вес *m*-й компоненты смеси в *i*-м скрытом состоянии ($\sum_{m=1}^{M} \tau_{im} = 1,\ i = \overline{1,N}$), μ_{im} – математическое ожидание нормального распределения, соответствующего *m*-й компоненте смеси в *i*-м скрытом состоянии, Σ_{im} – ковариационная матрица нормального распределения, соответствующая *m*-й компоненте смеси в *i*-м скрытом состоянии, а $g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}), \boldsymbol{o} \in R^Z$ – функция плотности многомерного нормального распределения, т.е. $g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}) = \frac{1}{\sqrt{(2\pi)^Z \left|\Sigma_{im}\right|}} e^{-0{,}5(\boldsymbol{o}-\mu_{im})^T \Sigma_{im}^{-1} (\boldsymbol{o}-\mu_{im})}, \boldsymbol{o} \in R^Z$. Таким образом, некоторую конкретную СММ будем задавать в виде набора определяющих ее параметров $\lambda = \{\Pi, A, B\}$.

1.2. Распознавание целых последовательностей, описываемых СММ

Пусть определено несколько классов, соответствующих некоторым различным случайным процессам с номерами $\overline{1,D}$, которые описываются соответствующими СММ $\lambda_1, ..., \lambda_D$, а также имеется последовательность многомерных наблюдений $O = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_T\}$. Для классификации последовательности, т.е. определения того, каким именно процессом, описываемым соответствующей СММ, она была порождена, как правило, применяют критерий максимума функции правдоподобия (МФП). В этом случае последовательность *O* относят к тому классу r^*, для которого значение логарифма функции правдоподобия является максимальным: $r^* = \arg \max_{r \in 1,...D} \left(\ln p\left(O \mid \lambda_r\right)\right)$.

Для расчета логарифма функции правдоподобия того, что последовательность *O* была сгенерирована процессом, описываемым СММ λ, т.е. $p(O \mid \lambda) = \ln \sum_{q_1, q_2, ..., q_T} p\left(\{\boldsymbol{o}_1, ..., \boldsymbol{o}_T\}, \{q_1, q_2, ..., q_T\} \mid \lambda\right)$, обычно применяют алгоритм forward-backward [5]. Для вычисления самого значения $\ln p\left(O \mid \lambda\right)$ необходима лишь первая часть forward-backward алгоритма, поэтому приведем только ее.

Во оригинальном алгоритме forward-backward вероятности умножаются друг на друга, т.е. числа меньше единицы, имеющие, как правило, значения, обратные количеству скрытых состояний, умножаются в количестве, пропорциональном длине последовательности. Для длинных последовательностей (длиной более 100) данные произведения достаточно быстро становятся меньше минимальных аппаратно реализуемых чисел современных машин. Для исправления этой проблемы необходимо либо использовать длинную арифметику, что значительно замедлит вычисления, либо масштабировать все

промежуточные произведения, чтобы они не стремились к нулю. Эффективные методы масштабирования, которые практически не замедляют обучения, известны и приведены в [4].

Первая часть forward-backward алгоритма (ее достаточно для вычисления логарифма функции правдоподобия) производит вычисление отмасштабированных прямых вероятностей $p(\{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t\}, q_t = s_i \mid \lambda)$, $t = \overline{1,T}$, $i = \overline{1,N}$, т.е. вероятностей того, что последовательность многомерных наблюдений $\{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t\}$ была порождена процессом, описываемым моделью λ, и что данный процесс находился в скрытом состоянии s_i в момент времени t. Алгоритм вычисления отмасштабированных прямых вероятностей и логарифма функции правдоподобия:

1) инициализация:

$$\tilde{\alpha}_1(i) = \pi_i b_i(\boldsymbol{o}_1),\ i = \overline{1,N}; \tag{29}$$

2) индукция:

$$\tilde{\alpha}_{t+1}(i) = b_i(\boldsymbol{o}_{t+1}) \left[\sum_{j=1}^{N} \alpha'_t(j) a_{ji} \right],\ i = \overline{1,N},\ t = \overline{1,T-1}, \tag{30}$$

где

$$\alpha'_t(j) = \frac{\tilde{\alpha}_t(j)}{\sum_{n=1}^{N} \tilde{\alpha}_t(n)},\quad j = \overline{1,N},\ t = \overline{1,T-1}. \tag{31}$$

Определим параметр масштаба:

$$c_t = \left(\sum_{i=1}^{N} \tilde{\alpha}_t(i) \right)^{-1},\ t = \overline{1,T}, \tag{32}$$

тогда

$$\alpha'_t(i) = c_t \tilde{\alpha}_t(i), \qquad i = \overline{1,N},\quad t = \overline{1,T-1},$$
$$\alpha'_t(i) = \left(\prod_{\tau=1}^{t} c_\tau \right) \alpha_t(i), \qquad i = \overline{1,N},\ t = \overline{1,T-1}.$$

Логарифм функции правдоподобия для последовательности наблюдений может быть вычислен с помощью параметров масштаба:

$$\ln\left[p(O|\lambda) \right] = -\sum_{t=1}^{T} \ln c_t. \tag{33}$$

2. Распознавание неполных последовательностей, описываемых СММ

Прежде чем производить распознавание неполных последовательностей, описываемых СММ, необходимо оценить параметры соответствующих СММ, т.е. обучить их. Вполне вероятно, что в реальной ситуации обучение придется также проводить на неполных последовательностях, соответственно, необходимо иметь алгоритмы обучения СММ по неполным последовательностям. Тем не менее в данной статье будет рассмотрен только вопрос распознавания неполных последовательностей. Вопрос обучения СММ на неполных последовательностях был рассмотрен автором в предыдущих работах [6, 7, 8], где был предложен алгоритм обучения СММ, основанный на маргинализации пропущенных наблюдений.

2.1. Распознавание неполных последовательностей с помощью маргинализации пропущенных наблюдений

Как и ранее, будем называть неполной, или «дефектной», последовательностью такую последовательность O, в которой значение некоторых наблюдений не определено. Обозначим пропуск символом $\varnothing$. Тогда $O = \left\{ \boldsymbol{o}_t \in R^*, t = \overline{1,T} \right\}$, $R^* = R^Z \cup \{\varnothing\}$.

Для получения алгоритма распознавания неполных последовательностей с помощью СММ необходимо прежде всего обратиться к формулам (29)–(33), по которым производится расчет прямых и

обратных вероятностей. Видно, что вычисление значений $b_i(\boldsymbol{o}_t)$, $i=\overline{1,N}$, $t=\overline{1,T}$, в формулах (29)–(33), которые используются как в алгоритме обучения СММ, так и в алгоритме распознавания последовательностей, невозможно, если $\boldsymbol{o}_t=\varnothing$, где символ $\varnothing$ означает пропущенное наблюдение, так как не определено конкретное наблюдаемое значение, а значит, нельзя рассчитать значение $b_i(\boldsymbol{o}_t)$, которое соответствует данному наблюдению. Чтобы можно было использовать эти формулы в случае неполных последовательностей, нужно каким-то образом доопределить значение сомножителя $b_i(\varnothing)$, $i=\overline{1,N}$, для тех прямых вероятностей, которые рассчитываются по отсутствующим в последовательности наблюдениям.

Предлагаемый в данной работе подход состоит в том, чтобы считать, что на месте пропуска могло стоять любое наблюдение из R^Z [9]. Руководствуясь этой идеей, представим значение $b_i(\varnothing)$, $i=\overline{1,N}$, как интеграл по всем возможным значениям пропущенного наблюдения:

$$b_i(\varnothing)=\int b_i(\boldsymbol{x})d\boldsymbol{x}=1,\quad i=\overline{1,N}.$$

Справедливость данного равенства обусловлена тем, что в один момент времени имеется только одно наблюдение $\boldsymbol{x}$, а также тем, что $b_i(\boldsymbol{x})$ – условная плотность распределения наблюдения $\boldsymbol{x}$ в скрытом состоянии s_i, $i=\overline{1,N}$. Руководствуясь теми же соображениями, определим значение плотности нормального распределения, входящего в смесь, для наблюдения-пропуска [9]:

$$g(\varnothing,\mu_{im},\Sigma_{im})=\int g(\boldsymbol{x},\mu_{im},\Sigma_{im})d\boldsymbol{x}=1,\quad i=\overline{1,N},\ m=\overline{1,M}.$$

Теперь выражение $b_i(\boldsymbol{o}_t)$, $i=\overline{1,N}$, $t=\overline{1,T}$, определено для всех $\boldsymbol{o}_t\in R^*$, и формулы (29)–(33) расчета прямых и обратных вероятностей можно расширить на случай неполных последовательностей.

Модифицированный алгоритм вычисления прямых вероятностей (отмасштабированный), используемый при распознавании неполных последовательностей:

1) инициализация:

$$\tilde{\alpha}_1(i)=\begin{cases}\pi_i, & \boldsymbol{o}_1=\varnothing,\\ \pi_i b_i(\boldsymbol{o}_1), & \text{иначе},\end{cases}\qquad i=\overline{1,N};$$

2) индукция:

$$\tilde{\alpha}_{t+1}(i)=\begin{cases}\sum\limits_{j=1}^{N}\alpha'_t(j)a_{ji}, & \boldsymbol{o}_{t+1}=\varnothing,\\ b_i(\boldsymbol{o}_{t+1})\left[\sum\limits_{j=1}^{N}\alpha'_t(j)a_{ji}\right], & \text{иначе},\end{cases}\qquad i=\overline{1,N},\ t=\overline{1,T-1},$$

где $\alpha'_t(j)=\dfrac{\tilde{\alpha}_t(j)}{\sum\limits_{n=1}^{N}\tilde{\alpha}_t(n)}$, $j=\overline{1,N}$, $t=\overline{1,T-1}$.

Параметр масштаба вычисляется по формуле: $c_t=\left(\sum\limits_{i=1}^{N}\tilde{\alpha}_t(i)\right)^{-1}$, $t=\overline{1,T}$. Логарифм функции правдоподобия вычисляется по формуле $\ln\left[p(O|\lambda)\right]=-\sum\limits_{t=1}^{T}\ln c_t$.

Назовем описанный выше прием доопределения неизвестных величин «маргинализацией пропущенных наблюдений», так как здесь вычисляется маргинальное распределение $b_i(\varnothing)$, $i=\overline{1,N}$, для случайной величины $\varnothing$, которая может принимать любое значение из множества R^Z. Легко видеть, что с помощью процедуры маргинализации можно проводить распознавание неполных последовательностей по критерию МФП, поскольку необходимые формулы для вычисления логарифма функции правдоподобия доопределены на случай пропущенных наблюдений.

2.2. Распознавание неполных последовательностей в пространстве первых производных от логарифма функции правдоподобия

Распознавание последовательностей, описываемых СММ, можно проводить не только с помощью критерия максимума функции правдоподобия. Ранее был разработан и успешно применен метод распознавания последовательностей в пространстве первых производных от логарифма функции правдоподобия того, что случайный процесс, описываемый СММ, сгенерировал распознаваемую последовательность, по различным параметрам СММ. Данный метод распознавания показал преимущество над критерием МФП в случаях близости СММ, описывающих классы, по параметрам, а также в условиях, когда СММ обучались на последовательностях, подверженных разного рода помехам [3]. Тем не менее случая полностью пропущенных наблюдений в последовательностях в данном исследовании не рассматривалось. Поскольку пропуски в наблюдениях также можно интерпретировать как своего рода помехи, то целесообразно исследовать применимость данного метода к анализу неполных последовательностей, описываемых моделями, обученными на неполных последовательностях.

Далее приведено описание упомянутого выше метода. Для наглядности рассмотрим случай двухклассовой классификации. Для каждой последовательности наблюдений *O* принадлежность к одной из двух моделей будем определять с помощью значений производных по различным параметрам модели. Пусть имеются обучающая выборка $\left\{O^1, O^2, ..., O^K\right\}$ и две СММ λ_1 и λ_2. Для каждой обучающей последовательности *O* из $\left\{O^1, O^2, ..., O^K\right\}$ будет построен вектор вида $\begin{pmatrix} \frac{\partial \ln p(O|\lambda_1)}{\partial \eta}\Big|_{\lambda_1} \\ \frac{\partial \ln p(O|\lambda_2)}{\partial \eta}\Big|_{\lambda_2} \end{pmatrix}$. Транспонированные версии этих векторов объединяются вместе в обучающую матрицу *X*, в которой столбцы соответствуют признакам (производным по параметрам моделей), а строки – последовательностям наблюдений. Также составляется вектор правильных ответов $Y = \left\{y_1, ..., y_K\right\}$, где $y_K \in \{1,2\}$ – это номер СММ, которая соответствует случайному процессу, породившему O^k, $k = \overline{1,K}$. Затем производится обучение классификатора по методу опорных векторов с помощью обучающей матрицы *X* и вектора правильных ответов *Y*. Для распознавания строится аналогичный вектор для рассматриваемой последовательности *O* и определяется, к какой группе этот вектор ближе по методу опорных векторов [10]. Описанный двухклассовый случай легко обобщить на многоклассовый случай, используя стратегии «каждый против каждого» или «один против всех», часто применяемые для бинарных классификаторов.

Далее приводится способ вычисления производных от логарифма функции правдоподобия по параметрам СММ [3].

Исходя из формулы (33),

$$\frac{\partial \ln p(O|\lambda)}{\partial \eta} = -\sum_{k=1}^{K}\left(\sum_{t=1}^{T}\frac{1}{c_t^k}\frac{\partial c_t^k}{\partial \eta}\right). \tag{34}$$

Вычисление производной от параметра масштаба по некоторому параметру модели η производится следующим образом:

$$\frac{\partial c_t}{\partial \eta} = -c_t^2\sum_{i=1}^{N}\frac{\partial \tilde{\alpha}_t(i)}{\partial \eta}, t = \overline{1,T}. \tag{35}$$

Для вычисления $\frac{\partial \tilde{\alpha}_t(i)}{\partial \eta}, i = \overline{1,N},$ продифференцируем по шагам алгоритм вычисления прямых переменных с масштабом:

Шаг 1:

$$\frac{\partial \tilde{\alpha}_1(i)}{\partial \eta} = \frac{\partial \alpha_1(i)}{\partial \eta}, \quad i = \overline{1,N}. \tag{36}$$

Шаг 2:

$$\frac{\partial\tilde{\alpha}_t(i)}{\partial\eta}=\left[\sum_{j=1}^{N}\left(\frac{\partial\alpha'_{t-1}(j)}{\partial\eta}a_{ji}+\alpha'_{t-1}(j)\frac{\partial a_{ji}}{\partial\eta}\right)\right]b_i(t)+\sum_{j=1}^{N}\left(\alpha'_{t-1}(j)a_{ji}\right)\frac{\partial b_i(t)}{\partial\eta}, \tag{37}$$

где $\frac{\partial\alpha'_{t-1}(j)}{\partial\eta}=\frac{\partial c_{t-1}}{\partial\eta}\tilde{\alpha}_{t-1}(j)+\frac{\partial\tilde{\alpha}_{t-1}(j)}{\partial\eta}c_{t-1},\ i=\overline{1,N},\ \ t=\overline{2,T}$.

Таким образом, для вычисления значений $\frac{\partial\tilde{\alpha}_t(i)}{\partial\eta},\ i=\overline{1,N},$ нам потребуется вычислить производные $\frac{\partial\tilde{\alpha}_1(i)}{\partial\eta},\frac{\partial b_i(t)}{\partial\eta},\frac{\partial a_{ij}}{\partial\eta},\quad i,j=\overline{1,N},t=\overline{1,T}.$

В случае недиагональной матрицы при вычислении производной по элементу ковариационной матрицы придется дифференцировать элементы обратной матрицы, поэтому будем рассматривать случай, когда матрицы $\Sigma_{im},\ i=\overline{1,N},\ \ m=\overline{1,M}$ являются диагональными.

Далее приведен способ вычисления производных $\frac{\partial\tilde{\alpha}_1(i)}{\partial\eta},\frac{\partial b_i(t)}{\partial\eta},\frac{\partial a_{ij}}{\partial\eta}$ для указанных значений параметра η (параметр η может принимать значения $\pi_i,a_{ij},\tau_{im},\mu^z_{im},\Sigma^{zz}_{im},\quad i,j=\overline{1,N},m=\overline{1,M},z=\overline{1,Z}$):

$$\frac{\partial\alpha_1(i)}{\partial\pi_j}=\begin{cases}b_i(1), & i=j,\\ 0, & i\neq j,\end{cases}\quad i,j=\overline{1,N}, \tag{38}$$

$$\frac{\partial b_i(t)}{\partial\pi_j}=0,\quad i,j=\overline{1,N},\ \ t=\overline{1,T}, \tag{39}$$

$$\frac{\partial\alpha_1(i)}{\partial a_{i_1j_1}}=0,\ i,i_1,j_1=\overline{1,N}, \tag{40}$$

$$\frac{\partial b_i(t)}{\partial a_{i_1j_1}}=0,\ \ i,i_1,j_1=\overline{1,N},\ t=\overline{1,T}, \tag{41}$$

$$\frac{\partial a_{ij}}{\partial x}=\begin{cases}1, & x=a_{ij},\\ 0, & x\neq a_{ij},\end{cases}\quad i,j=\overline{1,N}, \tag{42}$$

$$\frac{\partial b_i(t)}{\partial\tau_{i_1m}}=\begin{cases}g(o_t;\mu_{im},\Sigma_{im}), & i=i_1,\\ 0, & i\neq i_1,\end{cases}\quad i,i_1=\overline{1,N},\ t=\overline{1,T},\ m=\overline{1,M}, \tag{43}$$

$$\frac{\partial\alpha_1(i)}{\partial\tau_{i_1m}}=\begin{cases}\pi_i\dfrac{\partial b_i(1)}{\partial\tau_{i_1m}}, & i=i_1,\\ 0, & i\neq i_1,\end{cases}\quad i,i_1=\overline{1,N},\ m=\overline{1,M}, \tag{44}$$

$$\frac{\partial b_i(t)}{\partial\mu^z_{i_1m}}=\begin{cases}0{,}5\tau_{im}g(o_t;\mu_{im},\Sigma_{im})\dfrac{o^z_t-\mu^z_{im}}{\Sigma^{zz}_{im}}, & i=i_1,\\ 0, & i\neq i_1,\end{cases}\quad i,i_1=\overline{1,N},\ t=\overline{1,T},\ m=\overline{1,M},\ \ z=\overline{1,Z}, \tag{45}$$

$$\frac{\partial\alpha_1(i)}{\partial\mu^z_{i_1m}}=\begin{cases}\pi_i\dfrac{\partial b_i(1)}{\partial\mu^z_{i_1m}}, & i=i_1,\\ 0, & i\neq i_1,\end{cases}\quad i,i_1=\overline{1,N},\ m=\overline{1,M},\ \ z=\overline{1,Z}, \tag{46}$$

$$\frac{\partial b_i(t)}{\partial\Sigma^{zz}_{i_1m}}=\begin{cases}0{,}5\tau_{im}g(o_t;\mu_{im},\Sigma_{im})\left(\left(\dfrac{o^z_t-\mu^z_{im}}{\Sigma^{zz}_{im}}\right)^2-\dfrac{1}{|\Sigma_{im}|}\right), & i=i_1,\\ 0, & i\neq i_1,\end{cases}\quad i,i_1=\overline{1,N},\ t=\overline{1,T},\ m=\overline{1,M},\ \ z=\overline{1,Z}, \tag{47}$$

$$\frac{\partial \alpha_1(i)}{\partial \Sigma_{i_1 m}^{zz}} = \begin{cases} \pi_i \dfrac{\partial b_i(1)}{\partial \Sigma_{im}^{zz}}, i = i_1, \\ 0, \qquad i \neq i_1, \end{cases} \quad i, i_1 = \overline{1, N},\ m = \overline{1, M},\quad z = \overline{1, Z}. \tag{48}$$

Формулы (34)–(48) можно доопределить на случай неполных последовательностей, воспользовавшись приемом маргинализации пропущенных наблюдений, описанным в предыдущем подразделе. Таким образом, в формулах (34)–(48) будем считать $b_i(\varnothing) = 1,\quad i = \overline{1, N}$, а $g(\varnothing, \mu_{im}, \Sigma_{im}) = 1,\quad i = \overline{1, N},\ m = \overline{1, M}$, где символ $\varnothing$ означает пропущенное наблюдение. К тому же будут внесены дополнительные изменения в формулу (45):

$$\frac{\partial b_i(t)}{\partial \mu_{i_1 m}^{z}} = \begin{cases} 0{,}5\tau_{im} g(o_t; \mu_{im}, \Sigma_{im}) \dfrac{o_t^z - \mu_{im}^z}{\Sigma_{im}^{zz}}, & i = i_1 \text{ и } o_t \neq \varnothing, \\ 0, & \text{иначе}, \end{cases}$$

$$i, i_1 = \overline{1, N},\ t = \overline{1, T},\ m = \overline{1, M},\quad z = \overline{1, Z},$$

и формулу (47):

$$\frac{\partial b_i(t)}{\partial \Sigma_{i_1 m}^{zz}} = \begin{cases} 0{,}5\tau_{im} g(o_t; \mu_{im}, \Sigma_{im}) \left(\left(\dfrac{o_t^z - \mu_{im}^z}{\Sigma_{im}^{zz}} \right)^2 - \dfrac{1}{|\Sigma_{im}|} \right), & i = i_1 \text{ и } o_t \neq \varnothing, \\ 0, & \text{иначе}, \end{cases}$$

$$i, i_1 = \overline{1, N},\ t = \overline{1, T},\ m = \overline{1, M},\quad z = \overline{1, Z}.$$

3. Результаты вычислительного эксперимента

В данном разделе разработанный метод распознавания неполных последовательностей в пространстве первых производных от логарифма функции правдоподобия сравнивается с методом распознавания неполных последовательностей с помощью маргинализации пропущенных наблюдений.

В качестве истинных СММ были взяты модели λ_1 и λ_2 со следующими характеристиками. Число скрытых состояний $N = 3$, количество компонент в смесях $M = 3$. Размерность векторов наблюдений $Z = 2$. Вектор распределения начального состояния: $\Pi = [1, 0, 0]$, матрица вероятностей переходов:

$$A = \begin{bmatrix} 0{,}1 + \Delta A & 0{,}7 - \Delta A & 0{,}2 \\ 0{,}2 & 0{,}2 + \Delta A & 0{,}6 - \Delta A \\ 0{,}8 - \Delta A & 0{,}1 & 0{,}1 + \Delta A \end{bmatrix},$$

веса компонентов смесей:

$$\left\{ \tau_{im},\ i = \overline{1, N},\ m = \overline{1, M} \right\} = \begin{pmatrix} 0{,}3 + \Delta\tau & 0{,}4 - \Delta\tau & 0{,}3 \\ 0{,}3 & 0{,}4 + \Delta\tau & 0{,}3 - \Delta\tau \\ 0{,}3 - \Delta\tau & 0{,}4 & 0{,}3 + \Delta\tau \end{pmatrix}$$

(номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), векторы математических ожиданий компонент смесей:

$$\left\{ \mu_{im},\ i = \overline{1, N},\ m = \overline{1, M} \right\} = \begin{pmatrix} (0 - \Delta\mu \quad 0 + \Delta\mu)^T & (1 - \Delta\mu \quad 1 + \Delta\mu)^T & (2 - \Delta\mu \quad 2 + \Delta\mu)^T \\ (3 - \Delta\mu \quad 3 + \Delta\mu)^T & (4 - \Delta\mu \quad 4 + \Delta\mu)^T & (5 - \Delta\mu \quad 5 + \Delta\mu)^T \\ (6 - \Delta\mu \quad 6 + \Delta\mu)^T & (7 - \Delta\mu \quad 7 + \Delta\mu)^T & (8 - \Delta\mu \quad 8 + \Delta\mu)^T \end{pmatrix}$$

(номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), все ковариационные матрицы компонент смесей $\left\{ \Sigma_{im},\ i = \overline{1, N},\ m = \overline{1, M} \right\}$ были выбраны диагональными, значения всех элементов на диагонали были равны $0{,}1 + \Delta\sigma$. При этом у первой модели

$\Delta A = 0$, $\Delta\tau = 0$, $\Delta\mu = 0$, $\Delta\sigma = 0$, а у второй модели $\Delta A = 0{,}05$, $\Delta\tau = 0{,}05$, $\Delta\mu = 0{,}01$, $\Delta\sigma = 0{,}01$. Такой выбор параметров максимально усложняет задачу распознавания, поскольку случайные процессы, описываемые такими моделями, очень близки по свойствам и порождаемые ими последовательности трудно различить. С помощью каждой из моделей λ_1 и λ_2 было сгенерировано $K = 100$ обучающих и тестовых последовательностей длиной $T = 100$, причем каждая из последовательностей содержала G пропусков (число G изменялось от 0 до 90 в ходе эксперимента) в случайных местах. С помощью обучающих неполных последовательностей были получены оценки моделей $\hat{\lambda}_1$ и $\hat{\lambda}_2$ по алгоритму обучения СММ по неполным обучающим последовательностям, основанному на маргинализации пропущенных наблюдений [6, 7, 8]. Также с помощью производных от обучающих последовательностей и оценок СММ был обучен классификатор метода опорных векторов, гиперпараметры которого подобраны с помощью кросс-валидации по четырем блокам. Затем с помощью $\hat{\lambda}_1$ и $\hat{\lambda}_2$ проводилось распознавание неполных тестовых последовательностей с помощью метода маргинализации пропущенных наблюдений по критерию максимума функции правдоподобия (сплошная линия) и с помощью первых производных от логарифма функции правдоподобия, используя метод опорных векторов в качестве классификатора (рис. 1, штриховая линия), причем использовались производные по всем параметрам моделей. Фиксировался процент верно распознанных последовательностей. На рис. 1 приведены усредненные результаты после 50 запусков описанного выше эксперимента с различными начальными значениями генератора случайных чисел.

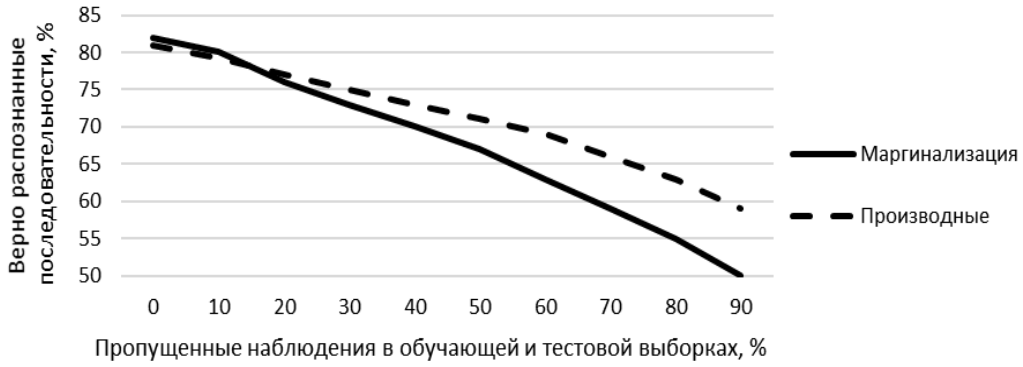

Рис. 1. Зависимость процента верно распознанных тестовых последовательностей от доли пропусков в обучающих и тестовых последовательностях

Как видно, метод распознавания, основанный на производных, начинает превосходить метод, основанный на маргинализации пропущенных наблюдений, начиная примерно с 20% пропусков в обучающих и тестовых последовательностях. При этом преимущество метода на основе производных увеличивается с увеличением процента пропусков, достигая 10% при 90% пропусков в последовательностях.

Заключение

В данной статье был предложен метод распознавания неполных последовательностей, который заключается в классификации последовательностей в пространстве признаков, образованном первыми производными от логарифма функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал распознаваемую неполную последовательность. Сравнительный анализ предложенного метода и разработанного автором ранее метода распознавания неполных последовательностей, основанного на маргинализации пропущенных наблюдений, показал, что предложенный метод позволяет достичь большего процента верно распознанных последовательностей, чем метод маргинализации, начиная с некоторого (в проведенном эксперименте – более 20%) процента пропусков в обучающих и тестовых последовательностях. Таким образом, предложенный метод может быть рекомендован к применению в условиях сильных помех, когда имеется много пропущенных данных, однако распознавание неполных последовательностей все же необходимо проводить.

ЛИТЕРАТУРА

1. Baum L.E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains // The Annals of Mathematical Statistics. 1966. V. 37. P. 1554–1563.
2. Упоминания ключевого слова «hidden Markov models» между 1800 и 2008 годами : данные из Google Ngram Viewer. URL: http://tinyurl.com/gmq5snv
3. Gultyaeva T.A., Popov A.A., Kokoreva V.V., Uvarov V.E. Classification of observation sequences described by Hidden Markov Models // Proc. of the Int. Workshop Applied Methods of Statistical Analysis Nonparametric approach AMSA-2015. Novosibirsk, Belokuriha, 14–19 Sep. 2015. P. 136–143.
4. Rabiner L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // Proc. of the IEEE. 1989. V. 77. P. 257–285.
5. Baum L.E., Egon J.A. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology // Bulletin of the American Meteorological Society. 1967. V. 73. P. 360–363.
6. Попов А.А., Гультяева Т.А., Уваров В.Е. Исследование подходов к обучению скрытых марковских моделей при наличии пропусков в последовательностях // Обработка информации и математическое моделирование : материалы Рос. науч.-техн. конф. Новосибирск, 21–22 апр. 2016. С. 125–139.
7. Popov A., Gultyaeva T., Uvarov V. A comparison of some methods for training hidden Markov models on sequences with missing observations // Proc. of 11th Int. Forum on Strategic Technology IFOST-2016. 2016. V. 1. P. 431–435.
8. Попов А.А., Гультяева Т.А., Уваров В.Е. Исследование методов обучения скрытых марковских моделей при наличии пропусков в последовательностях // Актуальные проблемы электронного приборостроения (АПЭП-2016) : труды XIII международной конференции : в 12 т. Новосибирск, 2016. Т. 8: Моделирование и вычислительная техника. Информационные системы и технологии. С. 149–152.
9. Cooke M., Green P., Josifovski L., Vizinh A. Robust automatic speech recognition with missing and unreliable acoustic data // Speech Communication. 2001. V. 34, No. 3. P. 267–285.
10. Boser B.E.; Guyon I.M., Vapnik V.N. A training algorithm for optimal margin classifiers // Proc. of the fifth annual workshop on Computational learning theory – COLT '92. 1992. P. 144.

Уваров Вадим Евгеньевич. E-mail: uvarov.vadim42@gmail.com
Новосибирский государственный технический университет

Поступила в редакцию 7 апреля 2017 г.

Uvarov Vadim E. (Novosibirsk State Technical University, Russian Federation).

Recognition of incomplete sequences described by hidden Markov models using first derivatives of likelihood function logarithm.

Keywords: hidden Markov models; machine learning; sequences; missing observations; incomplete data.

DOI: 10.17223/19988605/42/9

Hidden Markov model (HMM) conception was presented yet in 1970-s, however problems which concern using HMMs in case of incomplete data remain poorly investigated. These problems are quite relevant since in complex systems, e.g. when receiving signals from spacecrafts or aircrafts, one has to deal with datastreams of various sources in noisy environments when there is a high possibility of data loss or corruption. In this paper, we deal with the problem of missing observations in sequences. From now on we will refer to such sequences as incomplete. We consider a case when such missing observations are not generated by random process itself but rather occur randomly in sequences because of some external interference.

We propose a method for recognition of incomplete sequences which is based on classification of incomplete sequences using first derivatives of likelihood function logarithm with respect to various HMM parameters. We use a support vector machine classifier for that purpose. The likelihood in that case is the probability of incomplete sequence being generated by a HMM.

The proposed method was compared to a previously developed method for recognition based on marginalization of missing observations. The proposed method proved to be more effective than the other method in situation when the number of missing observations in training and testing sequences is high (more than 20% in our particular experiment). Thus, we propose to prefer the usage of the proposed method in situations when there is big loss of data but the recognition is still had to be done.

REFERENCES

1. Baum, L.E. & Petrie, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics.* 37. pp. 1554–1563. DOI: 10.1214/aoms/1177699147. https://projecteuclid.org/euclid.aoms/1177699147
2. Google Ngram Viewer. (n.d.) *Frequencies of "hidden Markov models" keyword in literature published between 1800 and 2008 year provided by Google Ngram Viewer*. [Online] Available from: http://tinyurl.com/gmq5snv
3. Gultyaeva, T.A., Popov, A.A., Kokoreva, V.V. & Uvarov, V.E. (2015) Classification of observation sequences described by Hidden Markov Models. *Proc. of the Int. Workshop Applied Methods of Statistical Analysis Nonparametric approach AMSA-2015*. Novosibirsk, Belokuriha. September 14–19, 2015. pp. 136-143.

4. Rabiner, L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE.* 77. pp. 257–285. DOI: 10.1109/5.18626
5. Baum, L.E. & Egon, J.A. (1967) An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bulletin of the American Meteorological Society*. 73. pp. 360-363. DOI: 10.1090/S0002-9904-1967-11751-8
6. Popov, A., Gultyaeva, T. & Uvarov, V. (2016) [Training hidden Markov models on incomplete sequences]. *Obrabotka informatsii i matematicheskoe modelirovanie* [Information processing and mathematical modelling]. Proc. of Russian Conference. Novosibirsk. April 21-22, 2016. pp. 125-139. (In Russian).
7. Popov, A., Gultyaeva, T. & Uvarov, V. (2016) A Comparison of Some Methods for Training Hidden Markov Models on Sequences with Missing Observations. *Proc. of 11th Int. Forum on Strategic Technology IFOST-2016.* 1. pp. 431-435. DOI: 10.1109/IFOST.2016.7884147
8. Popov, A., Gultyaeva, T. & Uvarov, V. (2016) [Training hidden Markov models on sequences with missing observations]. *Proc. of 13th Int. Conference on Actual Problems of Electronic Instrument Engineering (APEIE 2016)*. Vol. 1. pp. 317-320.
9. Cooke, M., Green, P., Josifovski, L. & Vizinh, A. (2001) Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication.* 34(3). pp. 267-285.
10. Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *Proc. of the Fifth Annual Workshop on Computational Learning Theory. COLT '92*. pp. 144. DOI: 10.1145/130385.130401