В.Е. Уваров, А.А. Попов, Т.А. Гультяева

Анализ неполных последовательностей, описываемых скрытыми марковскими моделями

Аннотация. Работа посвящена исследованию методов анализа неполных последовательностей, описываемых скрытыми марковскими моделями (СММ). Предложен алгоритм маргинализации пропущенных наблюдений, который может применяться как для обучения СММ по неполным последовательностям, так и для распознавания неполных последовательностей, описываемых СММ. Предложена модификация алгоритма Витерби, позволяющая производить декодирование, а также восстановление неполных последовательностей, описываемых СММ. Произведено сравнение предложенных алгоритмов со стандартными методами обработки пропусков методом исключения их из последовательности и склеивания оставшихся подпоследовательностей воедино, а также методом восстановления пропусков по среднему арифметическому соседних с пропуском наблюдений. На основе проведенных вычислительных экспериментов был сделан вывод, что предложенные алгоритмы превосходят другие рассмотренные методы анализа неполных последовательностей.

Ключевые слова: скрытые марковские модели, машинное обучение, последовательности, алгоритм Баума-Велша, пропущенные наблюдения, неполные данные, алгоритм Витерби, классификация.

Введение

Концепция СММ была предложена еще в 1970-х гг. коллективом ученых во главе с Л. Баумом [1, 2]. Традиционно СММ применялись для распознавания речи [3]. Начиная с 1980-х годов СММ стали применять в биоинформатике, например, при анализе цепочек ДНК. Однако наибольшей популярностью СММ стали пользоваться после 1990-х гг. [4]. Данная тенденция сохранилась вплоть до настоящего времени, что можно подтвердить частотой упоминания термина “hidden Markov model” в публикациях [5].

Тем не менее, в теории СММ имеется практически неизученная область, которая касается способов применения СММ в случае неполных данных. В данной работе рассматривается такой случай неполных данных, как присутствие пропусков в распознаваемых последовательностях. Такие последовательности с пропусками будем называть неполными. В рассматриваемой нами ситуации пропуски не генерируются самим случайным процессом, описываемым СММ, а возникают в произвольных местах последовательностей за счёт внешних условий. В этой работе рассматривается ряд задач анализа последовательностей, а именно: обучение, распознавание, декодирование и восстановление. Обучение заключается в нахождении такой оценки параметров СММ, которая наилучшим образом позволит описать имеющиеся последовательности. Задачей распознавания будем называть классификацию последовательностей. Классы последовательностей различаются описываемыми их СММ. Декодирование последовательности предполагает определение наиболее вероятной последовательности скрытых состояний, в которых находился описываемый случайный процесс при генерации данной последовательности. Восстановление последовательности производится путем замещения пропусков в последовательности наиболее подходящими в некотором смысле значениями.

В случае использования неполных последовательностей известные алгоритмы решения перечисленных выше задач требуют корректировки и уточнения. Частично данная проблема затрагивается в статье [6], где с помощью СММ

решалась задача распознавания зашумленной речи. В цитируемой работе анализировались спектрограммы, которые были получены с помощью оконного преобразования Фурье на основе записей речи, содержащих помехи. Авторы предложили в дополнение к классическим методам фильтрации шума, использовать подход, который основан на том, что отдельные сильнозашумленные участки спектрограммы считаются утерянными. Распознавание подобных последовательностей проводилось с использованием двух подходов: маргинализации пропущенных наблюдений и предварительного восстановления последовательностей. Маргинализация пропущенных наблюдений заключается в нахождении маргинального распределения непропущенных наблюдений путем интегрирования совместного распределения пропущенных и непропущенных наблюдений по всем возможным значениям пропущенных наблюдений. Авторы показали, что подобные подходы показывают лучший результат при распознавании зашумлённой речи, чем классические методы фильтрации шумов.

Результаты другого исследования, в котором проводилось распознавание неполных последовательностей с помощью СММ, представлены в [7]. В данной работе рассматривалась задача распознавания движений человека по видеоряду и их воспроизведения виртуальной моделью, изображающей человека. Пропуск наблюдений в этом случае обуславливался тем, что часть тела человека, движения которого повторяет модель, могла быть невидима, к примеру, закрыта препятствием. Для распознавания неполных последовательностей также задействовался подход маргинализации пропусков, а для определения последовательности движений человека использовался алгоритм декодирования неполных последовательностей.

В упомянутых выше работах авторы не затрагивают вопросы обучения СММ по неполным последовательностям. Задача этой работы состоит в исследовании и разработке алгоритмов обучения скрытых марковских моделей на неполных последовательностях, а также их декодирования, восстановления и распознавания.

Данная работа продолжает исследования в области методов использования СММ [8-13], проводимые на кафедре теоретической и прикладной информатики Новосибирского государственного технического университета.

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: $B = \left\{b_i\left(\boldsymbol{o}\right) = f\left(\boldsymbol{o} \mid q = s_i\right), i = \overline{1, N}, \boldsymbol{o} \in \mathbb{R}^Z\right\}$ [4]. В данной работе в качестве функций условной плотности распределения наблюдений рассматривается смесь многомерных нормальных распределений:

$$b_i(\boldsymbol{o}) = \sum_{m=1}^{M} \tau_{im} g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}), \;\; i = \overline{1,N}, \boldsymbol{o} \in \mathbb{R}^Z,$$

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; $g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}), \boldsymbol{o} \in \mathbb{R}^Z$ – функция плотности многомерного нормального распределения, т.е.

$$g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}) = \frac{1}{\sqrt{(2\pi)^Z |\Sigma_{im}|}} e^{-0.5(\boldsymbol{o}-\mu_{im})^T \Sigma_{im}^{-1} (\boldsymbol{o}-\mu_{im})},$$

$\boldsymbol{o} \in \mathbb{R}^Z$.

Таким образом, некоторую конкретную СММ будем задавать в виде набора определяющих ее параметров $\lambda = \{\Pi, A, B\}$.

1.2. Распознавание последовательностей с использованием скрытых марковских моделей

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. В этом случае последовательность O относят к тому классу r^*, для которого значение функции правдоподобия является максимальным: $r^* = \arg\max_{r \in 1,...D} \left(p\left(O \mid \lambda_r\right)\right)$.

Для расчета значения функции правдоподобия того, что последовательность O была сгенерирована процессом, описываемым СММ λ, т. е. $p\left(O \mid \lambda\right) = \sum_{q_1, q_2, ..., q_T} p\left(\{\boldsymbol{o}_1, ..., \boldsymbol{o}_T\}, \{q_1, q_2, ..., q_T\} \mid \lambda\right)$ обычно применяют алгоритм forward-backward (прямой-обратный) [2]. Для вычисления самого значения $L = p\left(O \mid \lambda\right)$ необходима лишь прямая часть forward-backward алгоритма, однако для полноты далее приводится и обратная часть алгоритма, так как она пригодится в дальнейшем для описания алгоритма обучения [14].

Первая часть forward-backward алгоритма производит вычисление прямых вероятностей $\alpha_t(i) = p(\{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t\}, q_t = s_i \mid \lambda), \;\; t = \overline{1,T}, \;\; i = \overline{1,N}$, 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 расчета прямых вероятностей и значения функции правдоподобия:

1) инициализация

$$\alpha_1(i) = \pi_i b_i(\boldsymbol{o}_1), \; i = \overline{1,N}; \qquad (1)$$

2) индукция

$$\alpha_{t+1}(i) = b_i(\boldsymbol{o}_{t+1}) \left[\sum_{j=1}^{N} \alpha_t(j) a_{ji} \right], \qquad (2)$$
$$i = \overline{1,N}, \;\; t = \overline{1,T-1};$$

3) завершение

$$p(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i). \qquad (3)$$

Вторая часть forward-backward алгоритма позволяет рассчитать обратные вероятности (backward - variables) $\beta_t(i) = p(\{\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, ..., \boldsymbol{o}_T\} \mid q_t = s_i, \lambda), \qquad t = \overline{1,T}, \;\; i = \overline{1,N}$, т. е. вероятности того, что модель λ в момент времени t находилась в состоянии s_i, а затем описываемым ей процессом была порождена последовательность наблюдений $\{\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+1}, ..., \boldsymbol{o}_T\}$. Алгоритм вычисления обратных вероятностей:

1) инициализация

$$\beta_T(i) = 1, \quad i = \overline{1,N};$$

2) индукция

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) b_j(\boldsymbol{o}_{t+1}) a_{ij}, \qquad (4)$$
$$i = \overline{1,N}, \qquad t = \overline{1,T-1}.$$

Таким образом, после рекурсивного вычисления прямых вероятностей по формулам (1)-(2), формула (3) позволяет вычислить искомое значение функции правдоподобия того, что

последовательность O была порождена процессом, описываемым СММ λ.

1.3. Обучение скрытой марковской модели

Для представления по имеющимся последовательностям изучаемого случайного процесса в виде СММ необходимо найти оценку параметров этой модели. Для этого нужно решить задачу обучения, которая состоит в настройке параметров модели λ по последовательности наблюдений $O^* = \{O^1, O^2, ..., O^K\}$, где K – число наблюдаемых последовательностей. Для решения этой задачи, как правило, применяется метод обучения, использующий процедуру максимизации функции правдоподобия $L(O^* \mid \lambda) = \prod_{k=1}^{K} p(O^k \mid \lambda)$.

Для этой процедуры известен эффективный алгоритм Баума-Велша [15], который является частным случаем алгоритма EM (EM – expectation-maximization; ожидание-максимизация). Так как алгоритм является итеративным, то перед началом его работы необходимо задать некоторое начальное приближение $\hat{\lambda}$ параметров СММ.

Для более краткого описания алгоритма Баума-Велша введем вероятности γ, ξ:

$$\gamma_t(i) = p(q_t = s_i \mid O, \hat{\lambda}) = \frac{\alpha_t(i)\beta_t(i)}{p(O \mid \hat{\lambda})}, \quad i = \overline{1,N},\ t = \overline{1,T-1}, \tag{5}$$

$$\xi_t(i,j) = p(q_t = s_i, q_{t+1} = s_j \mid O, \hat{\lambda}) = \frac{\alpha_t(i)a_{ij}b_j(\boldsymbol{o}_{t+1})\beta_{t+1}(j)}{p(O \mid \hat{\lambda})}, \quad i,j = \overline{1,N},\ t = \overline{1,T-1}, \tag{6}$$

$$\gamma_t(i,m) = p(q_t = i, \omega_{it} = m \mid O, \lambda) = \gamma_t(i)\left[\frac{\tau_{im} g(\boldsymbol{o}_t, \mu_{im}, \Sigma_{im})}{b_i(\boldsymbol{o}_t)}\right], \tag{7}$$

где $\hat{\lambda}$ – текущая оценка параметров модели, а ω_{it} – компонента смеси нормальных распределений в момент времени t для состояния i.

Отметим, что в формулах (5)-(6) задействуются прямые и обратные вероятности, которые вычисляются с использованием алгоритма forward-backward по формулам (1)-(4). Кроме того, следует заметить, что для каждой обучающей последовательности под индексом $k = \overline{1,K}$ вычисляется свой набор значений прямых и обратных вероятностей, а также вероятностей γ, ξ. Для их обозначения используется соответствующий индекс – $\alpha^{(k)}, \beta^{(k)}, \gamma^{(k)}, \xi^{(k)}$.

С учетом введенных обозначений для СММ с непрерывным распределением наблюдений новое приближение оценок параметров будет находиться в точке $\hat{\lambda}'$ с координатами [16]:

$$\hat{\pi}'_i = \frac{1}{K}\sum_{k=1}^{K}\gamma_1^{(k)}(i), \quad \hat{a}'_{ij} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\xi_t^{(k)}(i,j)}{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i)}, \tag{8}$$

$$\hat{\tau}'_{im} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)}{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i)}, \quad \hat{\mu}'_{im} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)\,\boldsymbol{o}_t^k}{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)}, \tag{9}$$

$$\hat{\Sigma}'_{im} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)(\boldsymbol{o}_t^{(k)} - \hat{\mu}'_{im})(\boldsymbol{o}_t^{(k)} - \hat{\mu}'_{im})^T}{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)}, \tag{10}$$

$i,j = \overline{1,N}, \quad m = \overline{1,M}$.

С помощью выражений (8)-(10) выполняется итерационное улучшение оценок параметров СММ. При этом на каждой новой итерации производится перерасчет переменных γ, ξ по формулам (5)-(7) с параметрами $\hat{\lambda} = \hat{\lambda}'$. Л. Баум и его коллеги доказали, что получаемая оценка модели $\hat{\lambda}'$ будет более правдоподобной, т.е., что $L\left(O^* \middle| \hat{\lambda}'\right) \geq L\left(O^* \middle| \hat{\lambda}\right)$ [4]. Алгоритм Баума-Велша в общем случае не обязательно сводится к глобальному максимуму, поэтому рекомендуется запускать его поочередно на нескольких различных начальных приближениях параметров, выбирая в итоге наилучшую оценку [17].

2. Проблема неполных последовательностей и методы ее решения

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

2.1. Маргинализация пропусков и склеивание неполных последовательностей

Для получения алгоритма распознавания неполных последовательностей с помощью СММ необходимо, прежде всего, обратиться к формулам (1)-(4), 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 (1)-(4), которые используются как в алгоритме обучения СММ, так и в алгоритме распознавания последовательностей, невозможно, если $\boldsymbol{o}_t=\varnothing$, 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 $\mathbb{R}^Z$ [6]. 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 [6]:

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

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 (1)-(4) расчета прямых и обратных вероятностей можно расширить на случай неполных последовательностей.

Модифицированный алгоритм вычисления прямых вероятностей, используемый как при обучении СММ, так и при распознавании неполных последовательностей:

1) инициализация

$$\alpha_1(i)=\begin{cases}\pi_i, & \boldsymbol{o}_1=\varnothing\\ \pi_i b_i(\boldsymbol{o}_1), & \textit{иначе}\end{cases},\qquad i=\overline{1,N};$$

2) индукция

$$\alpha_{t+1}(i)=\begin{cases}\sum_{j=1}^{N}\alpha_t(j)a_{ji}, & \boldsymbol{o}_{t+1}=\varnothing\\ b_i(\boldsymbol{o}_{t+1})\left[\sum_{j=1}^{N}\alpha_t(j)a_{ji}\right], & \textit{иначе}\end{cases},$$
$$i=\overline{1,N},\ t=\overline{1,T-1}.$$

Модифицированный алгоритм вычисления обратных вероятностей, используемый при обучении:

1) инициализация

$$\beta_T(i)=1,\quad i=\overline{1,N};$$

2) индукция

$$\beta_t(i)=\begin{cases}\sum_{j=1}^{N}\beta_{t+1}(j)a_{ij}, & \boldsymbol{o}_{t+1}=\varnothing\\ \sum_{j=1}^{N}\beta_{t+1}(j)b_j(\boldsymbol{o}_{t+1})a_{ij}, & \textit{иначе}\end{cases},$$
$$i=\overline{1,N},\ t=\overline{1,T-1}.$$

Также необходимо внести изменения в формулу (7):

$$\gamma_t(i,m)=\begin{cases}\gamma_t(i)\,\tau_{im}, & \boldsymbol{o}_t=\varnothing\\ \gamma_t(i)\left[\dfrac{\tau_{im}g(\boldsymbol{o}_t,\mu_{im},\Sigma_{im})}{b_i(\boldsymbol{o}_t)}\right], & \textit{иначе}\end{cases}.$$

Далее, формулы оценивания матриц ковариаций нормальных распределений (10), входящих в смеси, изменятся следующим образом:

$$\hat{\Sigma}'_{im} = \frac{\sum_{k=1}^{K}\sum_{\substack{t=1 \\ \boldsymbol{o}_t \neq \varnothing}}^{T^k-1} \gamma_t^{(k)}(i,m)(\boldsymbol{o}_t^{(k)} - \hat{\mu}'_{im})(\boldsymbol{o}_t^{(k)} - \hat{\mu}'_{im})^T}{\sum_{k=1}^{K}\sum_{\substack{t=1 \\ \boldsymbol{o}_t \neq \varnothing}}^{T^k-1} \gamma_t^{(k)}(i,m)}.$$

Как можно заметить, отличие состоит в том, что в данной формуле суммируются только те компоненты, которым соответствуют наблюдения, не являющиеся пропусками.

Назовем описанный выше прием доопределения неизвестных величин «маргинализацией пропущенных наблюдений». Так здесь вычисляется маргинальное распределение $b_i(\varnothing)$, $i = \overline{1,N}$ для случайной величины $\varnothing$, которая может принимать любое значение из множества R^*. Легко видеть, что с помощью процедуры маргинализации можно решать как задачу обучения СММ по неполным последовательностям, так и задачу распознавания неполных последовательностей, поскольку необходимые формулы доопределены на случай пропущенных наблюдений. Восстановления пропусков алгоритм маргинализации не предполагает.

Также опишем подход «склеивания» неполных последовательностей. Он предполагает исключение пропущенных наблюдений из исходной неполной последовательности и склеивание оставшихся подпоследовательностей в единую последовательность без пропусков. После очистки наблюдаемых последовательностей таким способом от пропусков можно использовать стандартную процедуру обучения, например, с помощью алгоритма Баума-Велша, как в разделе 1.3 или стандартную процедуру распознавания последовательности, например, как в разделе 1.2.

2.2. Декодирование последовательностей с пропусками

Для декодирования последовательностей, порожденных процессами, описываемыми СММ, т.е. формирования наиболее вероятной последовательности скрытых состояний $\hat{Q} = \{\hat{q}_1, \ldots, \hat{q}_T\}$ по наблюдаемой последовательности $O = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$, традиционно используется эффективный алгоритм Витерби [18]. Воспользовавшись идеей маргинализации пропущенных наблюдений, модифицируем алгоритм Витерби таким образом, чтобы его можно было применить для декодирования неполных последовательностей.

Модифицированный алгоритм Витерби:

1) инициализация

$$\delta_1(i) = \begin{cases} \pi_i, & \boldsymbol{o}_1 = \varnothing \\ \pi_i b_i(\boldsymbol{o}_1), & \textit{иначе} \end{cases}, \quad i = \overline{1,N};\ \psi_1(i) = 0;$$

2) индукция

$$\delta_t(j) = \begin{cases} \max\limits_{1 \le i \le N}\left[\delta_{t-1}(i)a_{ij}\right], & \boldsymbol{o}_t = \varnothing \\ \max\limits_{1 \le i \le N}\left[\delta_{t-1}(i)a_{ij}\right] b_j(\boldsymbol{o}_t), & \textit{иначе} \end{cases};$$

$$j = \overline{1,N}, \quad t = \overline{2,T}$$

$$\psi_t(j) = \arg\max_{1 \le i \le N}\left[\delta_{t-1}(i)a_{ij}\right], \quad j = \overline{1,N},\ t = \overline{2,T};$$

3) завершение

$$\hat{q}_T = \arg\max_{1 \le i \le N}\left[\delta_T(i)\right];$$

4) рекурсивное определение наиболее вероятной последовательности скрытых состояний

$$\hat{q}_t = \psi_{t+1}\left(\hat{q}_{t+1}\right), \quad t = \overline{T-1,\ 1}.$$

После завершения алгоритма получим сформированную наиболее вероятную последовательность скрытых состояний: $\hat{Q} = \{\hat{q}_1, \cdots, \hat{q}_T\}$.

2.3. Восстановление неполных последовательностей с использованием модифицированного алгоритма Витерби

Алгоритм декодирования неполных последовательностей, можно применить для восстановления последовательностей, содержащих пропуски. Пусть имеется СММ λ, а также сгенерированная соответствующим ей процессом последовательность O, в которой образовались случайным образом пропуски.

Для восстановления пропусков в последовательности O применим сначала к ней метод декодирования последовательностей с пропусками, описанный в разделе 2.2. Таким образом, можно найти наиболее вероятную последовательность скрытых состояний $\hat{Q} = \{\hat{q}_1, \cdots, \hat{q}_T\}$. После декодирования восстановить каждый пропуск можно используя найденное скрытое состояние. Заместим пропуск наиболее вероятным наблюдением, соответствующим этому скрытому

состоянию. Таким образом, пропуск в момент t с найденным скрытым состоянием $\hat{q}_t = s_{i^*}$ замещается наблюдением $\hat{\boldsymbol{o}}_t = \arg\max_{\boldsymbol{x}\in R^*} b_{i^*}(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}\in R^*} \sum_{m=1}^{M} \tau_{im} g(\boldsymbol{x}; \mu_{im}, \Sigma_{im})$. Очевидно, что в том случае, когда распределение наблюдений смесями нормальных распределений, максимум в данном выражении будет достигаться при $\boldsymbol{x} = \mu_{im^*}$, где $m^* = \arg\max_m (\tau_{im})$, т.е. наиболее вероятным значением $\boldsymbol{x}$ будет математическое ожидание компоненты смеси нормальных распределений, имеющей наибольший вес. К сожалению, такой выбор значения для замещения пропуска приводит к сильному переобучению СММ, что вызывает вырождение ковариационных матриц, как показали наши эксперименты. Более целесообразно заменять пропуск реализацией непрерывной случайной величины, соответствующей i^* -му состоянию скрытой марковской модели, т.е. имеющей распределение $b_{i^*}(\boldsymbol{x})$. В описанных далее экспериментах использовался последний подход.

2.4. Восстановление неполных последовательностей с помощью среднего арифметического соседних наблюдений

Для оценки эффективности разработанного алгоритма восстановления было произведено его сравнение со стандартным метод восстановления пропусков по среднему арифметическому k соседних наблюдений [19]. После восстановления последовательности таким способом некоторые пропуски могут остаться невосстановленными (к примеру, такие пропуски, у которых k соседних наблюдений тоже пропуски). Поэтому данная процедура выполняется повторно, но число рассматриваемых соседей k при этом увеличивается до размера всей последовательности T.

В этой работе пропуск замещался средним арифметическим 10 ближайших соседей (5 соседей слева и 5 справа). Данное число соседей было выбрано экспериментально: при таком значении параметра алгоритм восстановления последовательностей по среднему арифметическому соседей показал наилучшее качество восстановления.

2.5. Обучение и распознавание путем условного восстановления неполных последовательностей

Задача распознавания в данном случае имеет ту же постановку, что и в разделе 1.2 за исключение того, что распознаваемая последовательность является неполной. Предлагаемый подход заключается в том, чтобы сначала заместить пропуски в последовательности наиболее подходящими значениями с помощью процедуры из раздела 2.3, а затем распознать ее с помощью критерия максимума правдоподобия. Однако этот подход требует уточнения, поскольку для применения процедуры восстановления из раздела 2.3 требуется знание модели. Так как для последовательности неизвестна истинная метка класса, то имеет смысл условно восстанавливать неполную последовательность O с помощью той же СММ λ, по которой будет затем рассчитываться значение $P(O|\lambda)$.

В свою очередь, обучение СММ по последовательностям с пропусками можно осуществить, используя стандартные методы (например, алгоритм Баума-Велша), если предварительно восстановить данные последовательности. Для восстановления по методу из пункта 2.3 требуется знание модели. Если априорные знания отсутствуют, то модель нужно получить через процедуру обучения, например, используя подход обучения с маргинализацией из пункта 2.1, а уже после восстановления можно попытаться уточнить модель, проводя ее переобучение на восстановленных последовательностях. Эффективность подобного подхода необходимо проверить экспериментально. Очевидный недостаток такого подхода заключается в том, что обучение СММ необходимо проводить два раза.

3. Результаты экспериментов

3.1. Обучение СММ по неполным последовательностям

В первом вычислительном эксперименте проводилось сравнение различных подходов к обучению СММ по неполным последовательностям. В качестве истинной СММ была взята модель λ 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: $A = \begin{bmatrix} 0.1 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.6 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$, веса компонент смесей $\left\{\tau_{im}, i = \overline{1,N}, m = \overline{1,M}\right\} = \begin{pmatrix} 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), вектора математических ожиданий компонент смесей $\left\{\mu_{im}, i = \overline{1,N}, m = \overline{1,M}\right\} = \begin{pmatrix} (0\ \ 0)^T & (1\ \ 1)^T & (2\ \ 2)^T \\ (3\ \ 3)^T & (4\ \ 4)^T & (5\ \ 5)^T \\ (6\ \ 6)^T & (7\ \ 7)^T & (8\ \ 8)^T \end{pmatrix}$ (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 единичными. По заданной СММ было сгенерировано $K = 100$ обучающих последовательностей $\left\{O^1, O^2, ..., O^K\right\}$ длиной $T = 100$.

Генерация последовательностей, описываемых СММ с параметрами $\lambda = (\Pi, A, B)$, производилась по следующему алгоритму:

1) в качестве номера скрытого состояния q_1^*, в котором находился случайный процесс в первый момент времени генерации последовательности, принимается реализация дискретной случайной величины, имеющей распределение, заданное с помощью вектора начального распределения вероятностей Π;

2) в качестве номера скрытого состояния $q_t^*, t = \overline{2,T}$, в котором находился случайный процесс в момент времени t генерации последовательности, принимается реализация дискретной случайной величины x, имеющей распределение $p\left(q_t = x \mid q_{t-1} = q_{t-1}^*\right)$, которое задается строкой матрицы вероятностей переходов A, соответствующей скрытому состоянию q_{t-1}^*;

3) наблюдение $\boldsymbol{o}_t, t = \overline{1,T}$, сгенерированное случайным процессом в момент времени t, принимается равным реализации многомерной непрерывной случайной величины $\boldsymbol{x}$, имеющей распределение, представленное смесью многомерных нормальных распределений $f(\boldsymbol{x}) = \sum_{m=1}^{M} \tau_{q_t^* m} g(\boldsymbol{x}; \mu_{q_t^* m}, \Sigma_{q_t^* m})$, которая соответствует условной плотности распределения наблюдений из B в скрытом состоянии q_t^*.

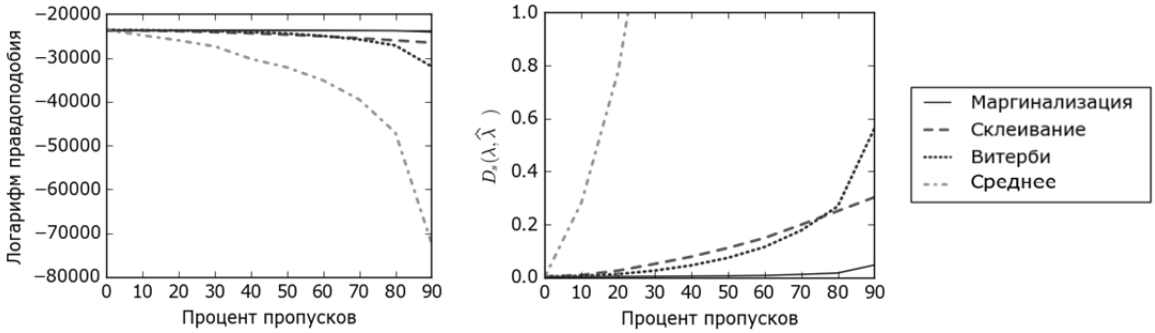
В ходе исследования изменялось количество пропусков в обучающих последовательностях. Пропуски распределялись случайным образом в каждой последовательности. Выход из итерационного процесса обучения осуществлялся по сходимости.

При изменении количества пропусков фиксировалось изменение следующих величин. Во-первых, фиксировалось значение логарифма функции правдоподобия того, что обученная модель сгенерировала целые обучающие последовательности, т.е. $\ln p\left(\left\{O^1, O^2, ..., O^K\right\} \mid \hat{\lambda}\right)$. Во-вторых, фиксировалось расстояние, основанное на симметричной разности логарифмов правдоподобия, между истинной и обученной моделью. Это расстояние вычисляется по формуле:

$$D_s = \frac{D\left(\lambda, \hat{\lambda}\right) + D\left(\hat{\lambda}, \lambda\right)}{2}, \qquad (11)$$

где $D\left(\lambda_1, \lambda_2\right) = \frac{1}{T}\left| \ln p\left(O^2 \middle| \lambda_1\right) - \ln p\left(O^2 \middle| \lambda_2\right) \right|$, а O^2 – последовательность, порожденная процессом, описываемым λ_2.

Данная метрика позволяет более адекватным образом сравнить две СММ, нежели норма разности параметров [4]. Для расчетов по формуле (11) генерировалось $K_D = 100$ последовательностей длиной $T_D = 500$ для каждой СММ и брался средний результат. Результаты эксперимента представлены на Рис. 1. Приведены средние значения после 10 проведенных экспериментов. Начертание линии обозначает использованный метод обучения: сплошная – алгоритм Баума-Велша с использованием маргинализации пропущенных наблюдений (раздел 2.1), штриховая – склеивание последовательностей с пропусками (раздел 2.2) и затем использование стандартного алгоритма Баума-Велша (раздел 1.3), пунктирная – восстановление



а) б)

Рис. 1. Средние значения: а - логарифма функции правдоподобия, рассчитанного на исходных последовательностях без пропусков для обученной СММ; б - расстояния, основанного на правдоподобии, между истинной моделью и ее оценкой

последовательностей с пропусками с помощью модифицированного алгоритма Витерби (раздел 2.3) и затем использование стандартного алгоритма Баума-Велша (раздел 1.3), штрихпунктирная – восстановление последовательностей с пропусками по среднему арифметическому соседних наблюдений (раздел 2.4) и затем использование стандартного алгоритма Баума-Велша (раздел 1.3).

На Рис. 1 видно, что алгоритм, использующий маргинализацию пропусков, показывает наилучшие результаты. Алгоритм, использующий восстановление пропусков по модифицированному алгоритму Витерби и алгоритм обучения, основанный на склеивании последовательностей с пропусками, очень близки по эффективности. Метод, основанный на восстановлении пропусков по среднему арифметическому ближайших соседей, показывает неудовлетворительные результаты.

Важнейшим показателем работоспособности алгоритмов обучения является использование их для построения классификаторов на основе полученных моделей. В этом случае в качестве метрики для сравнения качества обучающих алгоритмов можно использовать процент верно распознанных последовательностей. Затрудним условия распознавания, выбрав достаточно близкие по параметрам две модели СММ. Для этого рассмотрим модели λ_1 и λ_2, различающиеся только матрицами вероятностей переходов:

$$A = \begin{bmatrix} 0.1+\Delta A & 0.7-\Delta A & 0.2 \\ 0.2 & 0.2+\Delta A & 0.6-\Delta A \\ 0.8-\Delta A & 0.1+\Delta A & 0.1 \end{bmatrix}.$$

У первой модели λ_1 параметр $\Delta A = 0$, а у второй λ_2 $\Delta A = 0.3$. Все остальные параметры у исходных моделей совпадают и равны параметрам модели, использованной в предыдущем эксперименте. Нахождение оценок параметров каждой из двух моделей проводилось по набору из $K = 100$ обучающих последовательностей длиной $T = 100$, сгенерированному соответствующей истинной моделью. После нахождения оценок проводилось распознавание двух наборов, состоящих из $K_C = 100$ тестовых последовательностей длиной $T_C = 100$ без пропусков, сгенерированных каждой из двух исходных моделей соответственно. В качестве классификатора применялся алгоритм максимума логарифма правдоподобия (раздел 1.2). Результаты данного эксперимента представлены на Рис. 2. На графике приведены средние значения после 10 запусков. Присутствует также утолщенная сплошная линия, которая соответствует проценту последовательностей, верно распознанных с помощью истинных моделей.

Как видно из графика, обучение с помощью маргинализации пропущенных наблюдений обеспечивает наилучшие дискриминационные свойства полученных моделей. Алгоритм, использующий восстановление пропусков по модифицированному алгоритму Витерби и алгоритм обучения, основанный на склеивании последовательностей с пропусками очень близки по эффективности. Метод, основанный на восстановлении пропусков по среднему арифметическому ближайших соседей показывает неудовлетворительные результаты, даже несмотря на то, что он был оптимизирован по числу соседей.

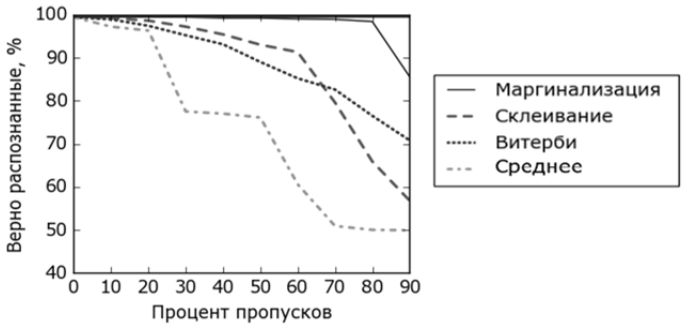


Рис. 2. Средний процент верно распознанных тестовых последовательностей

В реальных ситуациях может возникнуть необходимость решения задачи распознавания не только целых последовательностей, но и последовательностей с пропусками. Вначале посмотрим, как меняется эффективность распознавания таких последовательностей, если в классификаторе использовать исходные модели λ_1 и λ_2, по которым и проводилась генерация тестовых последовательностей. Результаты эксперимента представлены на Рис. 3. Приведены средние значения после 10 запусков. Начертание линии обозначает использованный метод классификации последовательностей с пропусками: сплошная – алгоритм маргинализации пропущенных наблюдений (раздел 2.1), штриховая – склеивание последовательностей с пропусками (раздел 2.2) и затем стандартный алгоритм распознавания (раздел 1.2), пунктирная – алгоритм восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби и дальнейшее распознавание стандартным алгоритмом (раздел 2.5), штрихпунктирная – восстановление последовательностей с пропусками по среднему арифметическому соседних наблюдений (раздел 2.4) и затем стандартный алгоритм распознавания (раздел 1.2).

Как видно, метод распознавания с помощью метода маргинализации пропущенных наблюдений показывает наилучший результат. На втором месте – алгоритм, основанный на склеивании последовательностей с пропусками с последующим стандартным распознаванием. Далее идет алгоритм восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби с последующим стандартным распознаванием. Худший результат показал алгоритм восстановления последовательностей с пропусками по среднему арифметическому соседних наблюдений с последующим стандартным распознаванием.

Наконец, рассмотрим наиболее реалистичный, на наш взгляд, случай, когда СММ, обученные на последовательностях с пропусками будут применяться для классификации подобных «дефектных» последовательностей. Данное исследование было проведено таким же образом, как и описанный выше эксперимент по распознаванию последовательностей без пропусков с помощью моделей, обученных на

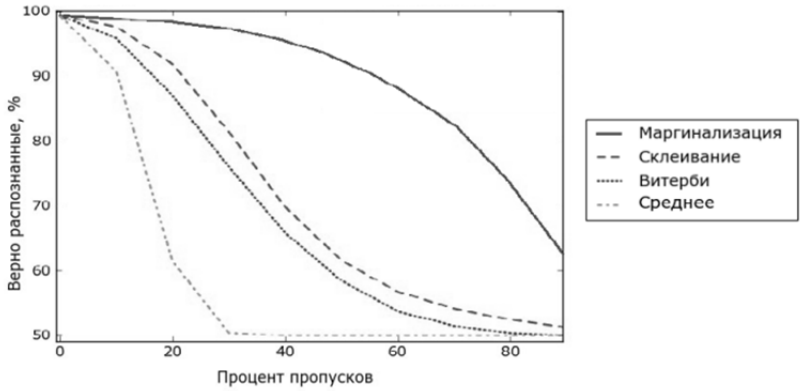


Рис. 3. Средний процент верно распознанных неполных последовательностей

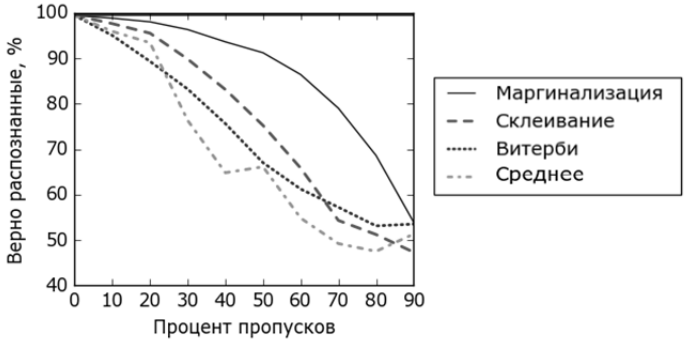


Рис. 4. Средний процент верно распознанных неполных последовательностей по моделям, обученным на неполных последовательностях

последовательностях с пропусками. Единственное отличие состояло в том, что в распознаваемых последовательностях теперь появлялись пропуски, причем процент пропусков в распознаваемых последовательностях равнялся проценту пропусков в обучающих последовательностях. Фиксировался процент верно распознанных последовательностей при изменении процента пропусков в обучающих и распознаваемых последовательностях. Результаты данного эксперимента представлены на Рис. 4. На графике приведены средние значения после 10 проведений эксперимента. Начертание линии обозначает использованный метод обучения и распознавания: сплошная – обучение и распознавание путем маргинализации пропущенных наблюдений (раздел 2.1), штриховая – обучение и распознавание путём склеивания последовательностей с пропусками (раздел 2.2), пунктирная – обучение и распознавание путём восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби (раздел 2.3), штрихпунктирная – обучение и распознавание путём восстановления последовательностей с пропусками по среднему арифметическому соседних наблюдений (раздел 2.4).

Как видно из рисунка, наилучший результат у алгоритма обучения и распознавания с помощью алгоритма маргинализации пропущенных наблюдений. Алгоритмы, использующие восстановление пропусков по модифицированному алгоритму Витерби и алгоритмы, основанные на склеивании последовательностей с пропусками очень близки по эффективности. Алгоритмы обучения и распознавания путем восстановления последовательностей с пропусками по среднему арифметическому соседних наблюдений показывают худший результат.

3.2. Декодирование и восстановление последовательностей с пропусками

В данном эксперименте сравнивались алгоритмы декодирования последовательностей с пропусками. С помощью модели λ_1 из предыдущего раздела было сгенерировано $K = 100$ последовательностей наблюдений длиной $T = 100$ с пропусками. Для декодирования использовалась истинная модель λ_1. Фиксировался процент верно декодированных скрытых состояний. Результаты эксперимента представлены на Рис. 5. Приведены средние значения после 10 запусков. Начертание линии обозначает использованный метод декодирования: пунктирная – декодирование с помощью модифицированного алгоритма Витерби (раздел 2.2), штрихпунктирная – восстановление пропусков по среднему

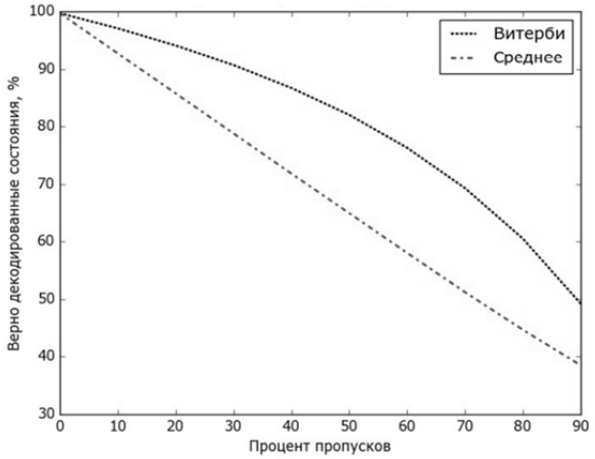


Рис. 5. Средний процент верно декодированных состояний в последовательностях с пропусками

арифметическому ближайших соседей (раздел 2.4) и затем декодирования восстановленной последовательности с помощью стандартного алгоритма Витерби. Как видно, метод декодирования с помощью модифицированного алгоритма Витерби несколько превосходит подход, основанный на восстановлении пропусков по среднему арифметическому ближайших соседей.

Также был проведен эксперимент по сравнению алгоритмов восстановления последовательностей с пропусками. Последовательности с пропусками генерировались таким же образом, как и в предыдущем эксперименте с помощью модели λ_1. Для восстановления использовалась истинная модель λ_1. Фиксировалась разница между исходными и восстановленными наблюдениями, выраженная в средней арифметической норме разностей исходных и восстановленных векторов наблюдений. Результаты эксперимента представлены на Рис. 6. Приведены средние значения после 10 запусков. Начертание линии обозначает использованный метод восстановления: пунктирная – восстановление с помощью модифицированного алгоритма Витерби (раздел 2.3), штрихпунктирная – восстановление пропусков по среднему арифметическому ближайших соседей (раздел 2.4).

Как видно из графика, метод восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби превосходит подход, основанный на восстановлении пропусков по среднему арифметическому ближайших соседей.

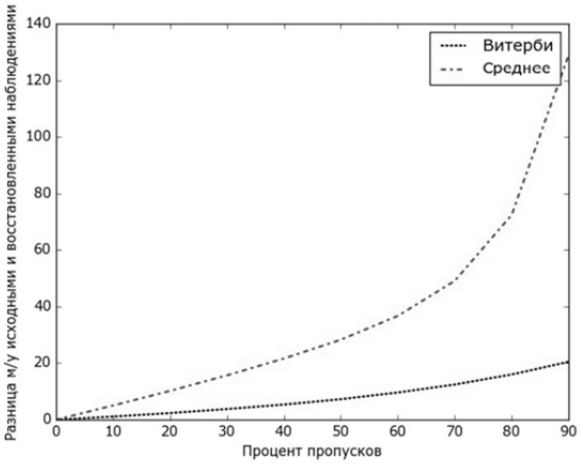


Рис. 6. Значение среднего арифметического нормы разностей исходных и восстановленных векторов наблюдений в последовательностях

Заключение

В результате проделанной работы был предложены алгоритмы обучения скрытых марковских моделей по последовательностям с пропусками, а также распознавания последовательностей с пропусками. Оба алгоритма основаны на маргинализации пропущенных наблюдений. Для декодирования и восстановления последовательностей с пропусками были предложены алгоритмы, основанные на модификации алгоритма Витерби для случая пропущенных наблюдений. Преимущество предложенных алгоритмов по сравнению с ранее известными подходами было подтверждено в ходе проведенных вычислительных экспериментов. В дальнейшем планируется исследовать эффективность распознавания последовательностей с пропусками с помощью классификатора, основанного на производных от логарифма функции правдоподобия по параметрам СММ [8].

Литература

1. Baum L. E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. The Annals of Mathematical Statistics, 1966, vol. 37, pp. 1554-1563.
2. 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, vol. 73, pp. 360-363.
3. Gales M., Young S. The Application of Hidden Markov Models in Speech Recognition. Signal Processing, 2007, vol. 1, no. 3, pp. 195-304.
4. Rabiner L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, 1989, vol. 77, pp. 257-285.
5. Упоминания ключевого слова “hidden Markov models” между 1800 и 2008 годами: данные из Google Ngram Viewer [Электронный ресурс]: – режим доступа: http://tinyurl.com/gmq5snv
6. Cooke M., Green P., Josifovski L., Vizinh A. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication, 2001, vol. 34, no. 3, pp. 267-285.
7. Lee D., Kulic D., Nakamura Y. Missing motion data recovery using factorial hidden Markov models. IEEE International Conference on Robotics and Automation, 2008, pp. 1722-1728.
8. Gultyaeva A., Popov A., Kokoreva V., Uvarov V. Classification of observation sequences described by Hidden Markov Models. Proceedings of the International Work-

shop Applied Methods of Statistical Analysis: Nonparametric approach, Novosibirsk, 2015, pp. 136-143.
9. Gultyaeva A., Popov A., Kokoreva V., Uvarov V. Training Hidden Markov Models on Incomplete Sequences. Proceeding of 13th International Conference on Actual Problems of Electronic Instrument Engineering, Novosibirsk, 2016. Vol. 1. pp. 317-320.
10. Гультяева Т. А., Попов А. А., Саутин А. С. Методы статистического обучения в задачах регрессии и классификации: монография; НГТУ. – Новосибирск, 2016. – 322 с.
11. 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. Новосибирск, 2016. С. 125-139.
12. Popov A., Gultyaeva A., Uvarov V. A Comparison of Some Methods for Training Hidden Markov Models on Sequences with Missing Observations. Proceedings of 11th International Forum on Strategic Technology IFOST-2016, 2016, vol. 1, pp. 431-435.
13. Попов А. А., Гультяева Т. А., Уваров В. Е. Исследование Методов Обучения Скрытых Марковских Моделей при Наличии Пропусков в Последовательностях//Актуальные проблемы электронного приборостроения. Новосибирск, 2016. Т. 8. С. 149-152.
14. Baum L. E., Sell G. R. Growth functions for transformations on manifolds. Pacific Journal of Mathematics, 1968, vol. 27, no. 2, pp. 211-227.
15. Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1977, vol. 39, pp. 1-38.
16. Li X. Training Hidden Markov Models with Multiple Observations – A Combinatorial Method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, vol. PAMI-22, no. 4, pp. 371-377.
17. Baum L. E., Petrie T., Soules G., Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics, 1970, vol. 41, no. 1, pp. 164-171.
18. Viterbi A. J. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Transactions on Information Theory, 1967, no. 13, pp. 260-269.
19. Gelman A., Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, 2006.

Уваров Вадим Евгеньевич. Аспирант кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Количество печатных работ: 18. Область научных интересов: структурные и статистические методы распознавания. E-mail: uvarov.vadim42@gmail.com

Попов Александр Александрович. Доктор технических наук, профессор кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Количество печатных работ: 150, в том числе три монографии. Область научных интересов: статистические методы анализа данных и планирования экспериментов. E-mail: a.popov@corp.nstu.ru

Гультяева Татьяна Александровна. Кандидат технических наук, доцент кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Количество печатных работ: более 70, в том числе одна монография. Область научных интересов: структурные и статистические методы распознавания. E-mail: t.gultyaeva@corp.nstu.ru

Analyzing Incomplete Sequences Using Gaussian Hidden Markov Models

V. E. Uvarov, A. A. Popov, T. A. Gultyaeva

Abstract. This paper studies the methods of incomplete sequence analysis using Gaussian hidden Markov models (HMMs). We present Marginalization algorithm, which can be applied both for training HMM on incomplete sequences and for recognition of incomplete sequences using HMMs. In addition, we present a modification of Viterbi algorithm that can be used for decoding and imputation of incomplete sequences using HMM. Both presented algorithms significantly outperform the standard methods of incomplete sequence analysis, namely: elimination of missing observations in sequences followed by “gluing” of the remaining subsequences into one sequence and imputation of missing observations with the mean of the neighboring observations.

Keywords: hidden Markov models, machine learning, sequences, Baum-Welch algorithm, missing observations, incomplete data, Viterbi algorithm, classification, decoding, imputation.

References

1. Baum L. E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. The Annals of Mathematical Statistics, 1966, vol. 37, pp. 1554-1563.
2. 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, vol. 73, pp. 360-363.

3. Gales M., Young S. The Application of Hidden Markov Models in Speech Recognition. Signal Processing, 2007, vol. 1, no. 3, pp. 195-304.
4. Rabiner L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, 1989, vol. 77, pp. 257-285.
5. Frequencies of "hidden Markov models" keyword in literature published between 1800 and 2008 year provided by Google Ngram Viewer. Available at: https://books.google.com/ngrams/graph?content=hidden+Markov+models&year_start=1800&year_end=2008&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Chidden%20Markov%20models%3B%2Cc0
6. Cooke M., Green P., Josifovski L., Vizinh A. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication, 2001, vol. 34, no. 3, pp. 267-285.
7. Lee D., Kulic D., Nakamura Y. Missing motion data recovery using factorial hidden Markov models. IEEE International Conference on Robotics and Automation, Pasadena, California, 2008, pp. 1722-1728.
8. Gultyaeva A., Popov A., Kokoreva V., Uvarov V. Classification of observation sequences described by Hidden Markov Models. Proceedings of the International Workshop Applied Methods of Statistical Analysis: Nonparametric approach, Novosibirsk, 2015, pp. 136-143.
9. Gultyaeva A., Popov A., Kokoreva V., Uvarov V. Training Hidden Markov Models on Incomplete Sequences. Proceedings of 13th International Conference on Actual Problems of Electronic Instrument Engineering, Novosibirsk, 2016. – Vol. 1. — pp. 317-320.
10. Gultyaeva A., Popov A., Sautin A. Metody statisticheskogo obucheniia v zadachakh regressii i klassifikatsii [Methods of Statistical Learning for the Problems of Regression and Classification]. Novosibirsk, 2016, 322 p.
11. Popov A., Gultyaeva A., Uvarov V. Issledovanie podkhodov k obucheniiu skrytykh markovskikh modelei pri nalichii propuskov v posledovatel'nostiakh [Training Hidden Markov Models on Incomplete Sequences]. Materialy konferentsii «Obrabotka informatsii i matematicheskoe modelirovanie», Rossiiskaia nauchno tekhnicheskaia konferentsiia [Proceeding of Russian Scientific conference "Information processing and mathematical modelling"]. Novosibirsk, 2016, pp. 125-139.
12. Popov A., Gultyaeva A., Uvarov V. A Comparison of Some Methods for Training Hidden Markov Models on Sequences with Missing Observations. Proceedings of 11th International Forum on Strategic Technology IFOST-2016, 2016, vol. 1, pp. 431-435
13. Popov A., Gultyaeva A., Uvarov V. Issledovanie Metodov Obucheniia Skrytykh Markovskikh Modelei pri Nalichii Propuskov v Posledovatel'nostiakh [Training Hidden Markov Models on Sequences with Missing Observations]. Trudy XIII mezhdunarodnoi konferentsii «Aktual'nye problemy elektronnogo priborostroeniia» [Proceeding of 13-th international conference "Actual Problems of Electronic Instrument Engineering"]. Novosibirsk, 2016, vol. 8, pp. 149-152.
14. Baum L. E., Sell G. R. Growth functions for transformations on manifolds. Pacific Journal of Mathematics, 1968, vol. 27, no. 2, pp. 211-227.
15. Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1977, vol. 39, pp. 1-38.
16. Li X. Training Hidden Markov Models with Multiple Observations – A Combinatorial Method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, vol. PAMI-22, no. 4, pp. 371-377.
17. Baum L. E., Petrie T., Soules G., Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics, 1970, vol. 41, no. 1, pp. 164-171.
18. Viterbi A. J. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Transactions on Information Theory, 1967, no. 13, pp. 260-269.
19. Gelman A., Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, 2006.

Vadim Uvarov. Postgraduate student at Novosibirsk State Technical University (20 Karla Marksa, Novosibirsk, 630073, Russia), author of 18 scientific papers, main academic interest – machine learning.

Alexander Popov. D.Sc. in engineering, professor at Novosibirsk State Technical University (20 Karla Marksa, Novosibirsk, 630073, Russia), author of more than 150 scientific papers and 3 monographs, main academic interest – statistical methods of data analysis and experimental design.

Tatyana Gultyaeva. PhD in engineering, associate professor at Novosibirsk State Technical University (20 Karla Marksa, Novosibirsk, 630073, Russia), author of more than 70 scientific papers and 1 monograph, main academic interest – structural and statistical methods of pattern recognition.