УДК 004.93'1: 004.852

И.Н. ПАЛАМАРЬ, С.С. ЮЛИН

ПОРОЖДАЮЩАЯ ГРАФИЧЕСКАЯ ВЕРОЯТНОСТНАЯ МОДЕЛЬ НА ОСНОВЕ ГЛАВНЫХ МНОГООБРАЗИЙ

Паламарь И.Н., Юлин С.С. **Порождающая графическая вероятностная модель на основе главных многообразий.**

Аннотация. В статье предлагается порождающая графическая вероятностная модель со скрытыми состояниями на основе нелинейных главных многообразий, заданных в виде сетки узлов, для решения задачи классификации временных последовательностей. В качестве метода аппроксимации обучающих данных сеткой узлов использован алгоритм самоорганизующихся карт Кохонена. Модель представлена в виде фактор-графа с описанием применяемых фактор-функций. Разработан метод обучения и вероятностного вывода на предлагаемой модели. Проведена оценка качества классификации предлагаемой модели в сравнении с существующими моделями (HMM, HCRF) на различных наборах данных из репозитория UCI, в том числе проведена сравнительная оценка при малом количестве обучающих данных.

Ключевые слова: классификация временных последовательностей, самоорганизующаяся карта Кохонена, скрытые Марковские модели, условные случайные поля со скрытыми состояниями.

Palamar I.N., Yulin S.S. **Generative probabilistic graphical model base on the principal manifolds.**

Abstract. The paper deals with a generative probabilistic graphical model with hidden states based on nonlinear principal manifolds specified as a grid of nodes to solve the problem of classification of sequences or time-series data. Kohonen's self-organizing map is used to approximate the training data as the grid nodes. The model is presented in factor-graph form the used factor-functions description. Method of learning and probabilistic inference is developed on the proposed model. Evaluated quality of the classification of the proposed model is compared with existing models (HMM, HCRF) on different sets of data from the UCI repository, including a comparative evaluation in the case when small amount of the training data is available.

Keywords: sequence classification, self-organizing map, HMM, HCRF.

1. Введение. Графические вероятностные модели со скрытыми состояниями используются в основном для решения задач классификации временных последовательностей [1]. К таким задачам относятся: распознавание речи, рукописного текста, жестов рук или головы, распознавание неисправностей (аварий) промышленных установок. Задача классификации формулируется как нахождение соответствия между наблюдаемой последовательностью данных и классом, к которому эти данные можно отнести. Для её решения хорошо зарекомендовали себя методы машинного обучения, основанные на таких графических вероятностных моделях как скрытые Марковские модели (HMM) [2] и условные случайные поля со скрытыми состояниями (HCRF) [3, 4]. Необходимо отметить, что HMM относится к классу порождающих мо-

делей, HCRF – к классу дискриминантных моделей. Обучение порождающих моделей состоит в нахождении параметров распределения вероятностей $p(x)$ (где x – наблюдаемые данные), наилучшим образом описывающих данные каждого класса, в то время как обучение дискриминантных моделей состоит в нахождении гиперплоскости, наилучшим образом разделяющей данные каждого класса. Основная работа в сравнении порождающих и дискриминантных классификаторов была проведена в 2002 году *Andrew Ng* и *Michael Jordan* [5]. Дискриминантный и порождающий подход к классификации сравнивались на примере наивного Байесовского классификатора и модели линейной логистической регрессии. В результате было теоретически доказано и эмпирически проверено на 15 различных наборах данных из репозитория UCI, что дискриминантному классификатору необходимо p обучающих данных для достижения своей асимптотически минимальной ошибки классификации, тогда как порождающему классификатору необходимо $\log(p)$ обучающих данных. Возможность обучения на малом количестве данных является важной особенностью классификаторов и наиболее актуальна в случае длительного или дорогостоящего процесса получения обучающей выборки.

К недостаткам порождающих моделей следует отнести то, что выбор формы функции плотности распределения вероятностей (нормальное, экспоненциальное и т. д.), накладывает ограничения на применимость порождающих моделей, так как не все данные соответствуют тем или иным известным параметрическим семействам распределений вероятностей. Кроме того, так как наблюдаемые данные представляют собой многомерные вектора признаков, то оценка параметров распределений в случае наличия линейных зависимостей среди признаков вектора является затруднительной или невозможной [6, 7].

Таким образом, цель данной работы – разработка графической вероятностной модели, аппроксимирующей обучающие данные без выполнения оценки параметров функции плотности распределения вероятностей появления наблюдаемых данных и позволяющей улучшить качество классификации в сравнении с классификатором на основе HMM.

В данной работе предлагается графическая вероятностная модель на основе аппроксимации обучающих данных главными многообразиями малой размерности, заданными в виде сетки узлов, методы её обучения и вероятностного вывода. В качестве алгоритма построения сетки узлов используется алгоритм самоорганизующихся карт Кохонена [8]. Предлагаемая в работе модель и методы её обучения основываются на комбинации элементов, используемых в метрическом и

байесовском подходах к классификации. Порождающий подход к классификации подразумевает построение отдельной модели для каждого класса. Частью параметров этой модели являются значения узлов аппроксимирующей сетки. Такой набор, как правило, является уникальным для каждого класса, благодаря чему представляется возможным выполнение классификации на его основе за счет оценки расстояния от классифицируемых данных до сетки узлов, соответствующих классов. Нормированное от 0 до 1 расстояние от наблюдаемых данных до каждого узла сетки может являться аналогом вероятности появления наблюдаемых данных в скрытых состояниях и использоваться в алгоритмах вероятностного вывода на предлагаемой модели.

В настоящее время существуют работы [9, 10], рассматривающие HMM совместно с самоорганизующимися картами Кохонена, которые используются для решения задачи инициализации математического ожидания перед оценкой параметров распределения вероятностей $p(x)$ по обучающим данным алгоритмом Баума-Велша. Также известны работы, например [11], использующие сети векторного квантования как способ сокращения размерности пространства признаков для классификатора HMM.

Кроме непрерывных HMM с гауссовой функцией плотности вероятности существуют дискретные HMM с дискретным распределением и аппроксимацией распределения с помощью сетей векторного квантования [12, 13], но данный подход показал свою несостоятельность в связи с низким качеством классификации (уступающим непрерывным HMM) и требованием наличия большого объема обучающих данных.

В отличие от описанных выше работ, в данной работе предлагается порождающая графическая вероятностная модель, не требующая оценки вероятности появления наблюдаемых данных.

Широко используются модели ANN/HMM (*Artificial Neural Network*/*Hidden Markov Model*) для классификации временных последовательностей на основе объединения нейронных сетей, таких как нейронные сети с задерживанием времени (TDNN) или рекуррентные нейронные сети (RNN), и скрытых Марковских моделей (HMM) [14]. Эти модели и модель, описанную в данной работе, объединяет общая идея выполнения аппроксимации обучающих данных без оценки параметров функции плотности распределения вероятностей появления наблюдаемых данных. В модели ANN/HMM наблюдаемые данные аппроксимируются параметрами нейронной сети. Каждое скрытое состояние HMM связывается с соответствующим нейроном выходного слоя. Вероятность появления некоторого наблюдения в текущем со-

стоянии оценивается как частота активации соответствующего нейрона выходного слоя.

В отличие от модели ANN/HMM, в данной работе описана модель, в которой каждое скрытое состояние графической вероятностной модели является узлом самоорганизующейся карты, а вместо оценки частоты активации соответствующего нейрона выходного слоя производится оценка расстояния между наблюдением и узлом карты.

2. Аппроксимация данных главными многообразиями. Задача аппроксимации данных нелинейными многообразиями формулируется следующим образом. Дано конечное множество n-мерных векторов $\overline{x}_1, \overline{x}_2, ..., \overline{x}_m \in R^n$. Для каждого $k = 0,1,...,n-1$ среди всех k-мерных нелинейных многообразий в R^n найти такое $M_k \subset R^n$, что сумма квадратов уклонений $\overline{x}_i$ от M_k минимальна:

$$\sum_{i=1}^{m} dist^2(\overline{x}_i, M_k) \to \min, \qquad (1)$$

где $dist(\overline{x}_i, M_k)$ – расстояние от точки в R^n до нелинейного многообразия.

Так как вид параметрической зависимости координат точек многообразия неизвестен (нет априорных соображений о структуре данных), то задача построения неограниченного нелинейного многообразия, то есть вычисления координат каждой из его точек, является весьма трудоемкой с точки зрения вычислений. Исходя из этого, многообразие считают ограниченным и задают конечным числом точек, то есть строят точечную аппроксимацию многообразия, называемую сеткой узлов, с заданным на ней отношением соседства [15, 16]. Тогда выражение (1) можно переписать в виде минимизации евклидового расстояния от точки данных до узлов сетки лежащих на нелинейном многообразии:

$$\sum_{i=1}^{m} \sum_{j=1}^{p} \left\| \overline{x}_i - \overline{w}_j \right\|_2 \to \min,$$

где $\overline{w}_j$ – координата j-ого узла сетки в R^n, $\overline{w}_j \in M_k$; p – количество узлов в сетке.

Методы построения нелинейных многообразий малой размерности хорошо справляются с задачей нахождения характерных особенностей данных, отбрасывая при этом шумовые составляющие. Так как неотъемлемой характеристикой любого классификатора является

способность к обобщению, то это свойство нелинейных многообразий будет использовано для построения на его основе предлагаемого классификатора временных последовательностей. Сформулировать основную задачу методов построения нелинейных многообразий в предлагаемой модели можно как задачу аппроксимации обучающих данных для каждого класса соответствующими главными многообразиями, заданными в виде сетки узлов. Так как точность аппроксимации напрямую зависит от количества узлов в сетке и влияет на обобщающую способность классификатора, то данная особенность позволяет добиться приемлемого качества классификации даже при малом наборе обучающих данных, путем подбора оптимального количества узлов в сетке. Следовательно, недостаток обучающих данных, может быть компенсирован введением дополнительных узлов аппроксимирующей сетки.

3. Описание предлагаемой графической вероятностной модели. За основу предлагаемой модели взята структура скрытой Марковской модели, которая представляет собой граф из двух типов вершин, соответствующих двум типам случайных величин (наблюдаемые и скрытые), и ребер между ними. Ребра характеризуют статистическую зависимость вершин и соединяют их, исходя из следующих двух правил:

– текущая наблюдаемая случайная величина связана со скрытой случайной величиной, соответствующей текущему моменту времени;

– скрытая случайная величина, соответствующая текущему моменту времени, связана со скрытой случайной величиной, соответствующей предыдущему моменту времени.

Структура скрытой Марковской модели изображена на рисунке 1.

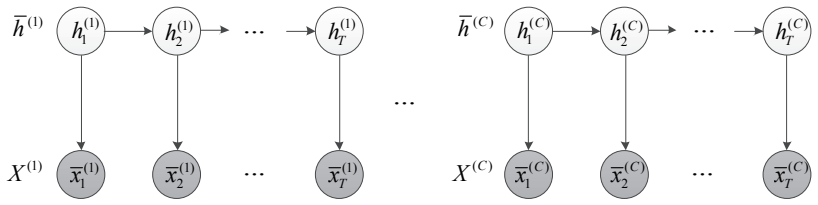

Рис. 1. Структура скрытой Марковской модели (одна модель для каждого класса $y = 1..C$)

Совместное распределение переменных скрытой Марковской модели выглядит как:

$$p(\overline{h}, X) = \prod_{t=1}^{T} p(h_t \mid h_{t-1}) \cdot p(\overline{x}_t \mid h_t), \qquad (2)$$

где $\overline{h}$ – вектор скрытых случайных величин (вектор состояний);

$X = \{\overline{x}_1, \overline{x}_2 ..., \overline{x}_T\}$ – упорядоченное множество векторов, характеризующих наблюдаемую последовательность;

$\overline{x}_t$ – вектор наблюдаемых случайных величин;

t – номер текущего наблюдения;

T – длина временной последовательности;

$p(h_t \mid h_{t-1})$ – распределение вероятностей перехода между состояниями;

$p(\overline{x}_t \mid h_t)$ – распределение вероятностей появления наблюдения $\overline{x}$ в состоянии h_t.

Выражение (2) представляет собой совместное распределение вероятностей всех случайных величин в графе, факторизованное в произведение двух условных распределений вероятности, составленных в соответствии с двумя правилами, описанными выше.

Так как предлагаемая модель не предполагает оценки параметров распределения $p(x)$, то целесообразным является описать структуру модели с помощью фактор-графа. В теории графических вероятностных моделей понятие фактора $\Psi(V_1, ..., V_k)$ определяется как некоторая функция $\Psi : Val(V_1), ..., Val(V_k) \to R$ от значений случайных величин $V_1, ..., V_k$, где $Val(V_k)$ – значения, принимаемые случайной величиной V_k. Фактор может представлять как условное, так и совместное распределение вероятностей случайных величин, а в общем случае любую функциональную зависимость.

Структура предлагаемой графической вероятностной модели изображена на рисунке 2.

Совместное распределение переменных предлагаемой модели представляется как произведение факторов, описываемое следующим выражением:

$$p(\overline{u}, X) = \prod_{t=1}^{T} p(u_t \mid u_{t-1}) \cdot \Psi(u_t, \overline{x}_t),$$

где $\overline{u}$ – вектор случайных величин, соответствующих номерам узлов аппроксимирующей сетки, каждая случайная величина принимает значение от 1 до количества узлов;

$p(u_t \mid u_{t-1})$ – распределение вероятностей перехода между состояниями (между узлами аппроксимирующей сетки);

$\Psi(u_t, \bar{x}_t)$ – нормированная фактор-функция, определяющая связь между текущим узлом аппроксимирующей сетки и текущим наблюдаемым отсчетом временной последовательности.

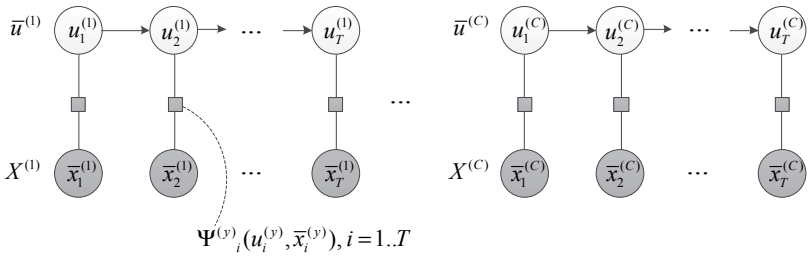

Рис. 2. Структура предлагаемой модели в виде вероятностного фактор-графа (одна модель для каждого класса $y = 1..C$)

Построение порождающего классификатора на основе графической вероятностной модели классификации сводится к оценке параметров совместного распределения всех случайных величин, заданных на структуре такой модели методом максимального правдоподобия (*Maximum Likelihood* (ML)). Для каждого класса образов (временных последовательностей) необходимо найти собственное совместное распределение. По факту наибольшего правдоподобия параметров той или иной модели наблюдаемым данным и выполняется процедура классификации.

4. Обучение модели и выполнение классификации. Рассмотрим этапы обучения модели. Пусть дано обучающее множество $Z^{(y)} = \{X_1^{(y)}, X_2^{(y)}, ..., X_m^{(y)}\}$ – неупорядоченный набор последовательностей наблюдений длины m соответствующих классу y, $y = 1..C$, где C – количество классов, а $X_i^{(y)} = \{\bar{x}_1^{(y)}, \bar{x}_2^{(y)}, ..., \bar{x}_T^{(y)}\}$ – i-ая последовательность наблюдений длины T, представляющая собой упорядоченный набор векторов-признаков $\bar{x}_t^{(y)}$, где $t = 1..T$.

Обучение представляет собой оптимизацию целевой функции:

$$L(W, \mathrm{A}) = \prod_{\forall X \in Z} \prod_{t=1}^{T} p(u_t \mid u_{t-1}) \cdot \Psi(u_t, \bar{x}_t) \to \max,$$

где W – множество значений узлов аппроксимирующей сетки; A – матрица вероятностей переходов между узлами сетки.

В результате обучения получим модели $M_y = \{W, \mathrm{A}\}$ для каждого класса y. Для обучения предлагаемой модели разработан алгоритм, состоящий из трех этапов.

Первый этап. Этап заключается в нахождении значений узлов аппроксимирующей сетки и соответствует стандартному алгоритму обучения самоорганизующейся карты Кохонена путем выполнения одной эпохи обучения карты. Для каждого класса y строится отдельная аппроксимирующая сетка узлов на наборе данных $Z^{(y)}$. В результате обучения каждому узлу $u_i, i = 1..N$, где N – количество узлов, будет соответствовать весовой вектор $\overline{w}_i$.

Второй этап. Этап состоит в оценке условного распределения вероятностей $p(u_t \mid u_{t-1})$.

Оценка распределения дискретной случайной величины по методу максимального правдоподобия сводится к подсчету частоты наблюдения того или иного события. В данном случае таким событием является переход из одного узла сетки в другой, а оценка условного распределения вероятностей $p(u_t \mid u_{t-1})$ выполняется как вычисление выражения:

$$p(u_t \mid u_{t-1}) = \mathrm{A},\ \mathrm{A}(i,j) = \alpha_{ij} = \frac{\gamma_{ij}}{\sum_{j=1}^{N} \gamma_{ij}},$$

$$\gamma_{ij} = \begin{cases} \gamma_{ij} + 1, \textit{если } i = u_{t-1} = \arg\min\limits_{u} \left\| \overline{w} - \overline{x}_{t-1} \right\| \\ \qquad\qquad u \\ \qquad\qquad j = u_t = \arg\min\limits_{u} \left\| \overline{w} - \overline{x}_t \right\| \\ \gamma_{ij}, \textit{в противном случае} \end{cases},$$

где γ_{ij} – количество переходов от одного узла карты с номером i к другому – с номером j в рамках одной последовательности, $i, j = 1..N$; u_t – номер узла карты, соответствующий вектору $\overline{x}_t$.

При оценке $p(u_t \mid u_{t-1})$ по методу максимального правдоподобия при большом количестве узлов сетки и при малом количестве обучающих данных не все вероятности переходов будут оценены. Исходя из этого, а также с целью внести регуляризацию в модель, мы произ-

водим оценку по методу апостериорного максимума (*Maximum a posteriori* (MAP)). В качестве априорного распределения выберем распределение Дирихле. Тогда выражение для оценки $p(u_t \mid u_{t-1})$ по методу апостериорного максимума запишем как:

$$\mathrm{A}(i,j) = \alpha_{ij} = \frac{\gamma_{ij} + \theta_i}{\sum_{j=1}^{N} \gamma_{ij} + \sum_{j=1}^{N} \theta_i},$$

где θ_i – параметр распределения Дирихле.

Так как в большинстве случаев какая-либо дополнительная информация о наиболее или наименее вероятных переходах отсутствует, то будем предполагать, что θ_i принимает одно и то же значение, одинаковое для всех i. Параметр θ_i оценивается на проверочной выборке (*validation set*).

Третий этап. Этап заключается в вычислении правдоподобия параметров полученной модели обучающим данным и последующем анализе изменения значения правдоподобия:

– если правдоподобие модели после текущей эпохи обучения самоорганизующейся карты не изменилось или уменьшилось, то процедура обучения останавливается и полученная модель признается оптимальной;

– если правдоподобие возросло, то выполняется переход к первому, а затем второму этапам обучения, и выполняется следующая эпоха обучения карты.

Сходимость такой процедуры обучения обеспечена сходимостью алгоритма обучения самоорганизующейся карты Кохонена, который гарантирует прекращение изменения значений узлов карты при большом количестве эпох обучения, что приведет к прекращению изменения значений правдоподобия в описанном выше алгоритме обучения и остановке процедуры обучения.

Непосредственно сама процедура классификации выполняется следующим образом. Пусть дано тестовое множество $B = \{b_1, b_2, ..., b_m\}$ – неупорядоченный набор последовательностей наблюдений длины m без указания принадлежности к классу, где $b_i = \{\overline{x}_1, \overline{x}_2, ..., \overline{x}_T\}$ – i-ая последовательность наблюдений длины T, представляющая собой упорядоченный набор векторов-признаков $\overline{x}_t$,

где $t = 1..T$. Тогда классификация последовательности b_i выполняется как вычисление выражения:

$$y_0 = \arg\max_{y=1..C}(L(b_i \mid M_y = \{W, \mathrm{A}\})),$$

где $L(b_i \mid M_y = \{W, \mathrm{A}\})$ – правдоподобие модели M_y класса y наблюдаемой последовательности b_i .

Учитывая то, что предложенная модель для классификации основана на комбинации теории построения нелинейных главных многообразий и теории вероятностных графических моделей в дальнейшем для краткости наименования будем использовать аббревиатуру NPM-PGM (*Nonlinear Principal Manifolds - Probabilistic Graphical Model*).

5. Вероятностный вывод на графической модели. Вычисление правдоподобия параметров графической вероятностной модели со скрытыми состояниями является NP-полной задачей. Но благодаря использованию Марковского предположения о зависимости текущего состояния модели только от предыдущего, можно воспользоваться эффективным алгоритмом динамического программирования, называемым алгоритмом «прямого-обратного хода», применяемым для вычисления правдоподобия, как в скрытых Марковских моделях, так и в условных случайных полях со скрытыми состояниями. Для вычисления правдоподобия используется только «прямой ход» данного алгоритма.

Логарифм правдоподобия модели M_y класса y наблюдаемой последовательности X вычисляется как:

$$L(X \mid M_y = \{W, \mathrm{A}\}) = \ln \sum_{u_1, u_2 \ldots u_T} \prod_{t=1}^{T} p(u_t \mid u_{t-1}) \cdot \tilde{\Psi}(u_t, \overline{x}_t).$$

Для того чтобы вычислить значение правдоподобия необходимо получить значение ненормированного фактора $\tilde{\Psi}(u_t, \overline{x}_t)$, которое определим как:

$$\tilde{\Psi}(u_t, x_t) = \tilde{\mathrm{B}},$$
$$\tilde{\mathrm{B}}(i,t) = \tilde{\beta}_{it} = -d(\overline{w}_i, \overline{x}_t),$$

где $d(\overline{w}_i, \overline{x}_t)$ – расстояние от весового вектора $\overline{w}_i$ узла i до наблюдаемого вектора $\overline{x}_t$, $t = 1..T$, $i = 1..N$ в заданном метрическом пространстве.

Метрика $d(\overline{w}_i, \overline{x}_t)$ выбирается, исходя из особенностей данных. В данной работе рассматривается только евклидово пространство, в котором $d(\overline{w}_i, \overline{x}_t) = \left\| \overline{w}_i - \overline{x}_t \right\|_2$.

Взятие расстояния со знаком «минус» необходимо для создания аналогии с вероятностью. Чем больше значение вероятности, тем более вероятно соотнесение наблюдения $\overline{x}$ к состоянию i. В случае замены понятия вероятности на понятие расстояния получаем обратную зависимость – чем больше значение расстояния, тем менее вероятно соотнесение наблюдения $\overline{x}$ к узлу карты i. Это несоответствие позволяет устранить смена знака значения расстояния.

Для того чтобы значения такого фактора можно было использовать для вычисления правдоподобия необходимо нормировать значения элементов матрицы $\tilde{\mathrm{B}}$ от 0 до 1, так чтобы сумма всех элементов матрицы $\tilde{\mathrm{B}}$ по столбцам составляла единицу, то есть сформировать матрицу B удовлетворяющую двум условиям:

1. $0 < \mathrm{B}(i,t) \le 1$;
2. $\sum_{i=1}^{N} \mathrm{B}(i,t) = 1$.

Нормирование значений распределения вероятностей является стандартной процедурой математической статистики зачастую использующейся для нормирования гистограмм распределений. Применим эту процедуру для нормирования расстояний.

Нормирование можно произвести следующим способом:

$$\mathrm{B}(i,t) = \frac{\tilde{\mathrm{B}}(i,t)}{\sum_{i=1}^{N} \tilde{\mathrm{B}}(i,t)}.$$

Следует отметить, что расстояние от наблюдения $\overline{x}_t$ до узлов карты $i, i = 1..N$ носит сильно неравномерный характер, в том смысле, что величина значения расстояния до большинства узлов карты будет иметь большое значение, и лишь для нескольких или одного узла («узла победителя») она будет иметь достаточно малое значение. В связи с этим предложено использовать нормирование в логарифмиче-

ском пространстве с вычитанием максимальных значений. Итоговое выражение для вычисления правдоподобия выглядит как:

$$L(X \mid M_y = \{W, \mathrm{A}\}) = \ln \sum_{u_1, u_2 \ldots u_T} \prod_{t=1}^{T} p(u_t \mid u_{t-1}) \cdot \Psi(u_t, \overline{x}_t) + \sum_{t=1}^{T} l_t = \\ = \ln \sum_{u_1, u_2 \ldots u_T} \beta_{11} \cdot \alpha_{12} \cdot \beta_{22} \cdot \ldots \cdot \alpha_{T-1,T} \cdot \beta_{TT} + \sum_{t=1}^{T} l_t. \quad (3)$$

6. Эксперименты и анализ результатов. Проведем сравнительную оценку моделей HMM, HCRF и предлагаемой модели (NPM-PGM) в решении задачи классификации на наборах данных из репозитория машинного обучения UCI [17, 18]. Описание данных приведено в таблице 1. Параметры тестируемых классификаторов приведены в таблице 2.

Приведем характеристики используемых наборов данных:

1) набор данных «*Spoken Arabic Digit Data Set*».Набор данных из 8800 (10 слов × 10 повторении × 88 говорящих) произнесенных цифр на арабском языке, представляющих собой временные последовательности из 13 MFCC коэффициентов, полученных путем цифровой обработки слов, произносимых носителями арабского языка в возрасте от 18 до 40 лет в составе 44 женщин и 44 мужчин. Частота дискретизации речевого сигнала – 11025 Гц;

2) набор данных «*Character Trajectories Data Set*».Траектории движения пера полученные при написании букв английского алфавита на планшете «*Wacom*». Данные состоят из трех параметров: координата точки по оси абсцисс, оси ординат и сила нажима. Данные численно дифференцированы, сглажены и нормированы. Частота считывания координат – 200 Гц.

Таблица 1. Описание данных для оценки качества классификаторов

Номер набора данных	1	2
Название	Spoken Arabic Digit Data Set	Character Trajectories Data Set
Описание	Слова, произнесённые на арабском языке	Траектории рукописных букв английского алфавита
Количество классов	10	20
Размерность пространства признаков	13	3
Количество экземпляров каждого класса	880	100
Источник	Репозиторий UCI [17]	Репозиторий UCI [18]

Таблица 2. Параметры тестируемых классификаторов

Параметры классификаторов	HMM	HCRF	NPM-PGM
Метод инициализации	Первоначальная инициализация центров распределения вероятности в каждом состоянии производится алгоритмом k -средних	Инициализация случайными значениями	Инициализация случайными значениями
Метод оптимизации	Алгоритм Баума-Велша	Квазиньютоновский алгоритм оптимизации – BFGS и метод сопряженных градиентов (CG)	Алгоритм, описанный в работе с использованием евклидовой метрики
Метод регуляризации	Без регуляризации	L_2 -регуляризация	Априорное распределение Дирихле

Оценка качества классификации проводилась в соответствии с методом k -блочной перекрёстной проверки (*k-fold cross-validation*), при которой размеченная выборка разбивается на k блоков: $(k-1)$ блок используется для обучения и один блок используется для тестирования. Процедура выполняется итеративно для всех k блоков, результатом является средняя оценка качества на тестовом наборе по k блокам. Эксперименты проводились при значении k равном 10. В качестве меры оценки качества классификации используем сбалансированную F-меру с усреднением по всем классам (*macro-average F-measure*).

Оценка параметра L_2 -регуляризации для модели HCRF, параметра распределения Дирихле для модели NPM-PGM, а также количества скрытых состояний для моделей HCRF и HMM и количества узлов аппроксимирующей сетки для модели NPM-PGM выполнялась на проверочной выборке, составляющей 10 % от тестовой. Значение параметра регуляризации λ было выбрано равным 1. Значение параметра распределения Дирихле θ было выбрано равным 0.0001. В качестве функции $p(\overline{x}_t \mid h_t)$ при обучении HMM выбрана конечная смесь гаус-

совских плотностей вероятности. Эксперименты проводились с одной и шестнадцатью компонентами смеси. В экспериментах использовалась двумерная квадратная карта Кохонена с гексагональной формой узлов с функцией соседства на основе вычисления евклидового расстояния между двумя узлами карты.

Результаты оценки качества классификации по методу k-блочной перекрёстной проверки приведены в таблице 3.

Выполним оценку качества классификации при малом количестве обучающих данных. Для этого обучение будем проводить на одном блоке, а тестирование на $(k-1)$ блоке. Результаты оценки качества классификации при малом количестве обучающих данных приведены в таблице 4.

Таблица 3. Результаты оценки качества классификации

	HMM (1 компонента)	HMM (16 компонент)	HCRF (BFGS)	HCRF (CG)	NPM-PGM
Spoken Arabic Digit Data Set (792 обучающих и 88 тестовых экземпляров каждого класса)					
Параметры	5 состояний	5 состояний	5 состояний	5 состояний	1024 узла
F-мера на обучающей выборке	0.9313	0.9722	0.9587	0.9353	0.9361
F-мера на тестовой выборке	0.8778	0.8368	0.9525	0.9344	0.9349
Переобучение	0.0535	0.1354	0.0062	0.0009	0.0012
Character Trajectories Data Set (90 обучающих и 10 тестовых экземпляров каждого класса)					
Параметры	7 состояний	7 состояний	7 состояний	7 состояний	256 узлов
F-мера на обучающей выборке	0.9568	0.9878	0.9960	0.9829	0.9999
F-мера на тестовой выборке	0.9329	0.9288	0.9651	0.9516	0.9809
Переобучение	0.0239	0.0590	0.0309	0,0313	0.0190

Таблица 4. Результаты оценки качества классификации при малом количестве обучающих данных

	HMM (1 компонента)	HMM (16 компонент)	HCRF (BFGS)	HCRF (CG)	NPM-PGM
Spoken Arabic Digit Data Set (88 обучающих и 792 тестовых экземпляров каждого класса)					
Параметры	5 состояний	5 состояний	5 состояний	5 состояний	1024 узла
F-мера на обучающей выборке	0.7421	1.0000	0.9977	0.9912	1.0000
F-мера на тестовой выборке	0.6018	0.5374	0.6420	0,6274	0.8043
Переобучение	0.1403	0.4626	0.3557	0,3638	0.1957
Character Trajectories Data Set (10 обучающих и 90 тестовых экземпляров каждого класса)					
Параметры	7 состояний	7 состояний	7 состояний	7 состояний	256 узлов
F-мера на обучающей выборке	0.9999	0.9952	1.0000	1.0000	1.0000
F-мера на тестовой выборке	0.7397	0.7243	0.6394	0.6053	0.8576
Переобучение	0.2602	0.2709	0.3606	0.3977	0.1424

В результате анализа оценки качества классификации можно отметить следующее:

– на всех используемых наборах данных предлагаемая порождающая графическая вероятностная модель (NPM-PGM) показала лучшее качество классификации, чем модель HMM, при оценке качества классификации по методу k -блочной перекрестной проверки при большом количестве обучающих данных;

– на одном из наборов данных (*Character Trajectories Data Set*) модель NPM-PGM показала лучшее качество классификации, чем модель HCRF, при оценке качества классификации по методу k -блочной перекрестной проверки при большом количестве обучающих данных;

– на всех используемых наборах данных модель NPM-PGM показала лучшее качество классификации, чем модели HMM и HCRF, при оценке качества классификации при малом количестве обучающих данных;

– количество узлов аппроксимирующей сетки значительно влияет на качество классификации и на эффект переобучения;

– модель HMM с шестнадцатью компонентами смеси показала худшее качество классификации, чем модель с одной компонентой в результате сильного переобучения, обусловленного возрастанием количества параметров модели;

– BFGS метод оптимизации, используемый при обучении HCRF, показал лучшее качество классификации, чем метод сопряженных градиентов. Это может быть объяснено спецификой работы алгоритмов на конкретных данных, а также наибольшей склонностью метода сопряженных градиентов, в отличие от BFGS, сходиться к локальным экстремумам. Целевая функция оптимизации модели HCRF является невыпуклой, что потребовало множества повторов выполнения процедуры обучения для получения результата близкого к глобально оптимальному.

Учитывая тот факт, что вычислительная сложность алгоритма «прямого-обратного хода» квадратично зависит от размера сетки, необходимо выбирать оптимальное количество узлов, так как с увеличением их числа время процедуры классификации путем вычисления выражения (3) будет квадратично расти, при этом приводя к незначительным увеличениям показателя F-меры.

К достоинствам предлагаемой модели следует отнести:

– возможность использования различных функций расстояния в зависимости от конкретных данных;

– высокое качество классификации при малом количестве обучающих данных.

К недостаткам предлагаемой модели следует отнести: сложность выбора оптимального количества узлов аппроксимирующей сетки, которое не приводит к переобучению и обеспечивает приемлемое качество классификации, а также значительно большее количество скрытых состояний, чем в моделях HMM и HCRF, что сказывается на увеличении времени выполнения классификации.

7. Выводы и направления дальнейших исследований. В данной работе предложена порождающая графическая вероятностная модель со скрытыми состояниями на основе главных многообразий для классификации временных последовательностей. Проведена сравнительная оценка качества классификации на различных наборах данных. Предлагаемая модель показала лучшие среди тестируемых моделей результаты классификации при обучении на малом количестве обучающих данных. Улучшение качества классификации по показате-

лю F-меры при тестировании методом k-блочной перекрёстной проверки при большом количестве обучающих данных составило:

– в сравнении с классификатором HMM – на 5.7 % на тестовом наборе 1, на 4.8 % на тестовом наборе 2;

– в сравнении с классификатором HCRF – на 1.6 % на тестовом наборе 2.

Используемый метод обучения модели направлен на увеличение правдоподобия модели для каждого класса своим обучающим данным. При таком подходе непосредственно сама дискриминантная способность модели не максимизируется. В дальнейшем предполагается улучшение данного метода за счет применения дискриминантного обучения с целью выбора оптимального количества узлов аппроксимирующей сетки, а также оптимального значения весовых векторов узлов.

Литература

1. *Koller D., Friedman N.* Probabilistic Graphical Models: Principles and Techniques // USA: MIT Press. 2009. 1265 p.
2. *Рабинер Л.Р.* Скрытые Марковские модели и их применение в избранных приложениях при распознавании речи: Обзор // ТИИЭР. 1989. Т. 77. № 2. С. 86–120.
3. *Gunawardana A., Mahajan M., Acero A., Platt J.C.* Hidden conditional random fields for phone classification. // International Conference on Speech Communication and Technology. 2005. pp. 1117–1120.
4. *Sutton C., McCallum A*. An Introduction to Conditional Random Fields for Relational Learning // USA: MIT Press, 2006. 35 p.
5. *Ng A., Jordan M*. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes // In Advances in Neural Information Processing Systems 14. 2002. pp. 841– 848.
6. *Sung Y.-H., Boulis C., Manning C., Jurafsky D.* Regularization, adaption, and non-independent features improve hidden conditional random fields for phone classification // Automatic Speech Recognition & Understanding. Kyoto. 2007. pp. 347 – 352.
7. *Ширяев А. Н*. Вероятность: В 2-х кн. Кн. 1 // Москва: МЦНМО. 2007. 551 с.
8. *Кохонен Т*. Самоорганизующиеся карты // Москва: Бином. 2008. 655 с.
9. *Kurimo M.* Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models. Thesis for the degree of Doctor of Technology // Finland: Helsinki University of Technology. 1997.
10. *Somervuo P.* Competing Hidden Markov Models on the Self-Organizing Map // IJCNN. 2000. vol. 3. pp. 169–174.
11. *Calinon S., Billard A.* Recognition and Reproduction of Gestures using a Probabilistic Framework combining PCA, ICA and HMM // ICML. 2005. pp. 105–112.
12. *Neukirchen C., Rottland J., Willett D., Rigoll G.* A continuous density interpretation of discrete HMM systems and MMI-neural networks // IEEE Transactions on Speech and Audio Processing. 2001. vol. 9. Iss. 4. pp. 367–377.
13. *Osterndorf M., Rohlicek J. R.* Joint quantizer design and parameter estimation for discrete hidden Markov models // Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing. 1990. pp. 705 – 708.
14. *Маковкин К.А*. Гибридные модели: скрытые марковские модели и нейронные сети, их применение в системах распознавания речи // Модели, методы, алго-

ритмы и архитектуры систем распознавания речи. Москва: Вычислительыний центр имени А.А. Дородницына, 2006. С. 40 – 96.
15. *Горбань А. Н*., *Россиев А.А.* Итерационное моделирование неполных данных с помощью многообразий малой размерности // Нейрокомпьютеры. 2002. Т. 4. С. 40–44.
16. *Gorban A., Kegl. B., Wunsch D., Zinovyev A.* Principal Manifolds for Data Visualisation and Dimension Reduction // New York : Springer. 2008. 340 p.
17. UCI Machine Learning Repository. Spoken Arabic Digit Data Set. URL: http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit (дата обращения: 12.05.2014).
18. UCI Machine Learning Repository. Character Trajectories Data Set. URL: http://archive.ics.uci.edu/ml/datasets/Character+Trajectories (дата обращения: 12.05.2014).

References

1. Koller D., Friedman N. Probabilistic Graphical Models: Principles and Techniques. USA: MIT Press, 2009. 1265 p.
2. Rabiner, L. [A tutorial on hidden Markov models and selected applications in speech recognition]. *TIIJeR – Proceedings of the IEEE*. 1989. vol. 77. no 2. pp. 86 – 120. (In Russ.).
3. Gunawardana A., Mahajan M., Acero A., Platt J.C. Hidden conditional random fields for phone classification. International Conference on Speech Communication and Technology. 2005. pp. 1117 – 1120.
4. Sutton C., McCallum A. An Introduction to Conditional Random Fields for Relational Learning. USA: MIT Press, 2006. 35 p.
5. Ng A., Jordan M. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. In Advances in Neural Information Processing Systems 14. 2002. pp. 841– 848.
6. Sung Y.-H., Boulis C., Manning C., Jurafsky D. Regularization, adaption, and non-independent features improve hidden conditional random fields for phone classification. Automatic Speech Recognition & Understanding. Kyoto. 2007. pp. 347 – 352.
7. Shirjaev A.N. *Verojatnost'* [Probability]. Moscow: MCCME. 2007. 551 p. (In Russ.).
8. Kohonen *T*. *Samoorganizujushhiesja karty* [The Self-Organizing Map]. Moscow: Binom, 2008. 655 p. (In Russ.).
9. Kurimo M. Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models. Thesis for the degree of Doctor of Technology. Finland: Helsinki University of Technology. 1997.
10. Somervuo P. Competing Hidden Markov Models on the Self-Organizing Map. IJCNN. 2000. vol. 3. pp. 169 – 174.
11. Calinon S., Billard A. Recognition and Reproduction of Gestures using a Probabilistic Framework combining PCA, ICA and HMM. ICML. 2005. pp. 105 – 112.
12. Neukirchen C., Rottland J., Willett D., Rigoll G. A continuous density interpretation of discrete HMM systems and MMI-neural networks. IEEE Transactions on Speech and Audio Processing. 2001. vol. 9. Iss. 4. pp. 367 – 377.
13. Osterndorf M., Rohlicek J. R. Joint quantizer design and parameter estimation for discrete hidden Markov models. Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing. 1990. pp. 705–708.
14. Makovkin K.A. [Hybrid models: the Hidden Markov Models and Neural Networks and their application in speech recognition systems]. *Modeli, metody, algoritmy i arhitektury sistem raspoznavanija rechi – Models, methods, algorithms and architec-*

ture of speech recognition systems. Moscow: Dorodnicyn Computing Centre of RAS. 2006. pp. 40 – 96. (In Russ.).
15. Gorban, A. N., Rossiev A.A. [Iterative modeling incomplete data using low-dimensional manifolds]. *Nejrokomp'jutery – Neurocomputers*. 2002. vol. 4. pp. 40–44. (In Russ.).
16. Gorban A., Kegl B., Wunsch D., Zinovyev A. Principal Manifolds for Data Visualisation and Dimension Reduction. New York: Springer. 2008. 340 p.
17. UCI Machine Learning Repository. Spoken Arabic Digit Data Set. Available at: http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit (accessed 12.05.2014).
18. UCI Machine Learning Repository. Character Trajectories Data Set. Available at: http://archive.ics.uci.edu/ml/datasets/Character+Trajectories (accessed 12.05.2014).

Паламарь Ирина Николаевна — к-т техн. наук, доцент, профессор кафедры вычислительных систем Рыбинского государственного авиационного технического университета имени П. А. Соловьёва. Область научных интересов: теория анализа и распознавания изображений, речи, искусственные нейронные сети, теория искусственного интеллекта, системный анализ. Число научных публикаций — 57 и 23 запатентованных изобретений. irina.palamar@mail.ru; ФГБОУ ВПО «Рыбинский государственный авиационный технический университет имени П.А. Соловьева» (РГАТУ имени П.А. Соловьева), Пушкина ул., д. 53, Рыбинск, Ярославская обл., 152934, р.т. +7(4855) 21-97-16, факс +7 (4855) 21-39-64.

Palamar Irina Nikolaevna — Ph.D., associate professor, professor, Computing Systems Department, Faculty of Informatics and Radio Electronics, P.A. Solovyov Rybinsk State Aviation Technical University (RSATU). Research interests: image analysis, image and speech recognition, artificial neural networks, artificial intelligence, system analysis. The number of publications — 57. The number of patented invention — 23. irina.palamar@mail.ru; Rybinsk State Aviation Technical University (RSATU), 53, Pushkin Str, Rybinsk, 152934, office phone +7(4855) 21-97-16, fax +7 (4855) 21-39-64.

Юлин Сергей Сергеевич — инженер-программист комплексного тематического отдела по работе с беспилотными летательными аппаратами ОАО «КБ «ЛУЧ». Область научных интересов: классификация временных последовательностей, графические вероятностные модели со скрытыми состояниями, алгоритмы поиска закономерностей в данных. Число научных публикаций — 6. julin.serg@gmail.com; ОАО «КБ «ЛУЧ», б-р Победы, 25, Рыбинск, 152920, РФ; р.т. +7(4855)28-58-22, факс +7(4855)28-58-35.

Yulin Sergey Sergeevich — software engineer integrated thematic department on work with drones of Lutch, JSC. Research interests: sequence classification, probabilistic graphical model with hidden states, data mining. The number of publications — 6. julin.serg@gmail.com; Lutch, JSC, 25, blvd. Pobedi, Rybinsk, 152920, Russia; office phone +7(4855)28-58-22, fax +7(4855)28-58-35.

РЕФЕРАТ

Паламарь И.Н., Юлин С.С. **Порождающая графическая вероятностная модель на основе главных многообразий.**

Графические вероятностные модели со скрытыми состояниями используются для решения задач классификации временных последовательностей. К таким задачам относятся: распознавание речи, рукописного текста, жестов рук или головы.

В данной работе предлагается графическая вероятностная модель на основе аппроксимации обучающих данных главными многообразиями малой размерности, заданными в виде сетки узлов. В качестве алгоритма построения сетки узлов используется алгоритм самоорганизующихся карт Кохонена. Нормированное от 0 до 1 расстояние от наблюдаемых данных до каждого узла сетки может являться аналогом вероятности появления наблюдаемых данных в скрытых состояниях и использоваться в алгоритмах вероятностного вывода на предлагаемой модели. За основу предлагаемой модели взята структуру скрытой Марковской модели, которая представляет собой граф из двух типов вершин, соответствующих двум типам случайных величин (наблюдаемые и скрытые), и ребер между ними. Каждому скрытому состоянию соответствует узел аппроксимирующей сетки. Обучение модели производится итерационной процедурой, направленной на максимизацию правдоподобия. Вероятностный вывод с целью вычисления правдоподобия производится алгоритмом «прямого-обратного хода».

Предлагаемая в работе модель показала лучшие среди тестируемых моделей (HMM, HCRF) результаты классификации при обучении на малом количестве обучающих данных (менее 100 экземпляров каждого класса) на тестовых наборах «*Character Trajectories Data Set» и «Spoken Arabic Digit Data Set»* из репозитория UCI. Улучшение качества классификации по показателю F-меры при тестировании методом k-блочной перекрёстной проверки при большом количестве обучающих данных составило:

- в сравнении с классификатором HMM – на 5.7 % на тестовом наборе *«Spoken Arabic Digit Data Set»*, на 4.8 % на тестовом наборе «*Character Trajectories Data Set»*;
- в сравнении с классификатором HCRF – на 1.6 % на тестовом наборе «*Character Trajectories Data Set»*.

К достоинствам предлагаемой модели следует отнести: возможность использования различных функций расстояния в зависимости от конкретных данных и высокое качество классификации при малом количестве обучающих данных. К недостаткам предлагаемой модели следует отнести: сложность выбора оптимального количества узлов аппроксимирующей сетки, которое не приводит к переобучению и обеспечивает приемлемое качество классификации, а так же значительно большее количество скрытых состояний, чем в моделях HMM и HCRF, что сказывается на увеличении времени выполнения классификации.

SUMMARY

Palamar I.N., Yulin S.S. **Generative probabilistic graphical model base on the principal manifolds.**

Probabilistic graphical models with hidden states are used for sequences or time-series data classification. Sequences classification include: speech, handwriting and gestures classification.

In this paper we propose a probabilistic graphical model based on approximation of training data with principal manifolds of small dimension, given in the form of grid nodes. Kohonen`s self-organizing map algorithm is used as an algorithm for constructing the grid nodes. Normalized from 0 to 1, the distance from the observed data to each node is considered as an analog of the probability of occurrence of the observed data in the hidden states and is used in probabilistic inference algorithms on the proposed model. The basis of the proposed model is taken Hidden Markov Model structure, which is a graph of the two types of nodes corresponding to two types of random values (observed and hidden), and edges between them. Each hidden node corresponds to a node of the approximating grid. Model's learning is performed by an iterative procedure directed to maximizing likelihood. Probabilistic inference to compute the likelihood is performed using «forward–backward» algorithm.

The proposed model showed better results than the HMM and HCRF at training on a small amount of training data (less than 100 copies of each class) on the test sets from the UCI repository: «Character Trajectories Data Set» and «Spoken Arabic Digit Data Set». Classification quality by F-measure by using the method k-folds cross-validation with a large number of training data improvement:

- compared with the HMM classifier was 5.7% on «Spoken Arabic Digit Data Set» and 4.8% on «Character Trajectories Data Set»;
- compared with the HCRF classifier was 1.6% on «Character Trajectories Data Set».

The advantages of the proposed model include: possibility of using different distance functions depending on the specific data; and high quality classification with a small amount of training data. The disadvantages of the proposed model include: the complexity of selecting the optimal number of the approximating grid nodes, which does not lead to overfitting and ensures a good quality of the classification, as well as considerably larger amount of hidden states needed comparing with HMM and HCRF classifiers which leads to longer classification runtimes.