УДК 519.217.2

Т.А. Гультяева, А.А. Попов

КЛАССИФИКАЦИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ В УСЛОВИЯХ НЕТОЧНОГО ЗАДАНИЯ ИХ СТРУКТУРЫ

Рассматривается задача классификации последовательностей с использованием методологии скрытых марковских моделей. Классификация проводится как с использованием стандартного подхода, так и с использованием классификатора *k* ближайших соседей и метода опорных векторов в пространстве признаков, инициированных скрытыми марковскими моделями. Исследовалось поведение классификаторов при ошибках в спецификации структуры марковской модели.

Ключевые слова: *скрытые марковские модели, производные от логарифма функции правдоподобия, классификатор k ближайших соседей, метод опорных векторов.*

Скрытые марковские модели (СММ) широко используются в задачах моделирования различных процессов, демонстрируя при этом хорошие описательные способности. Построенные на обучающих выборках СММ используются также и для задач классификации. Однако на этих задачах СММ не всегда показывают необходимый уровень дискриминирующих свойств.

В работе в качестве объектов классификации рассматривается множество смоделированных последовательностей, порожденных двумя близкими по своим параметрам СММ. Параметризация СММ проводится в соответствии с выбранной их структурой, под которой будем понимать число скрытых состояний и размер словаря наблюдаемых символов. В реальных ситуациях априорная информация о структуре СММ, как правило, отсутствует. Задача структурной идентификации в этом случае может быть поставлена, но на практике, как правило, она не решается в полном объеме. Рассматривается возможность использования традиционного классификатора на основе СММ, классификатора *k* ближайших соседей (*kNN*) и метода опорных векторов (*SVM*) в пространстве признаков, инициированных обученными скрытыми марковскими моделями в условиях их структурной неопределенности.

1. Постановка задачи

СММ – это случайный процесс с ненаблюдаемой стационарной марковской цепью. СММ описывается следующими параметрами [1]:

1. Вектор вероятностей начальных состояний

$$\Pi = \{\pi_i\},\ i = \overline{1, N},$$

где $\pi_i = P\{q_1 = i\}$, q_1 – скрытое состояние в начальный момент времени $t = 1$; N – количество скрытых состояний в модели.

2. Матрица вероятностей переходов

$$A = \{a_{ij}\},\ i, j = \overline{1, N},$$

где $a_{ij} = P\{q_t = j | q_{t-1} = i\}$, $t = \overline{2,T}$, где T – длина наблюдаемой последовательности.

3. Матрица вероятностей наблюдаемых символов выглядит следующим образом:

$$B = \{b_i(t)\},\ i = \overline{1,N},$$

где $b_i(t) = P\{o_t | q_t = i\}$, o_t – символ, наблюдаемый в момент времени $t = \overline{1,T}$. Рассматривается случай, когда функция распределения вероятностей наблюдаемых символов описывается смесью нормальных распределений

$$b_i(t) = \sum_{j=1}^{M} \tau_{ij} \left(\sqrt{2\pi}\sigma_{ij}\right)^{-1} e^{-\left(o_t - \mu_{ij}\right)^2 / 2\sigma_{ij}^2},\ i = \overline{1,N},\ t = \overline{1,T},$$

где τ_{ij} – это вес j-й компоненты смеси в i-м скрытом состоянии, $i = \overline{1,N}$, $j = \overline{1,M}$, M – это количество смесей. Параметры μ_{ij} и σ_{ij}^2 являются соответственно математическим ожиданием и дисперсией j-й компоненты смеси в i-м скрытом состоянии, $i = \overline{1,N}$, $j = \overline{1,M}$.

Таким образом, СММ полностью описывается матрицей вероятностей переходов, а также вероятностями наблюдаемых символов и вероятностями начальных состояний: $\lambda = (A,\ B,\ \pi)$.

В работе рассматривает поведение традиционного классификатора, основанного на отношении логарифмов функций правдоподобия, классификаторов *kNN* и *SVM* в условиях структурной неопределенности СММ. Один из возможных вариантов такой неопределенности заключатся в том, что исследователю точно не известно число скрытых состояний модели.

Метод ближайших соседей основан на оценивании сходства объектов. Классифицируемый объект относится к тому классу, которому принадлежит большинство из его соседей – k ближайших к нему объектов обучающей выборки. Этот классификатор достаточно подробно описан, например в [2].

Основная идея метода опорных векторов – это перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Этот классификатор описан в [3].

2. Пространства признаков

В качестве пространств признаков, в которых производится классификация, рассматриваются пространства первых производных от логарифма функции правдоподобия по различным параметрам СММ.

Приведем формулы для вычисления первых производных от логарифма функции правдоподобия по некоему параметру η (более подробно вывод формул приведен в [4–6]):

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

где $O = \{O^1,\ O^2,\ \ldots,\ O^K\}$, O^k – k-я наблюдаемая последовательность длиной T, K – количество последовательностей, c_t^k – параметр масштаба для последовательности O^k. В дальнейшем индекс k будем опускать для удобства.

Учитывая, что $c_t = \left(\sum_{i=1}^{N} \tilde{\alpha}_t(i) \right)^{-1}$, формула для вычисления производной от параметра масштаба по параметру η имеет вид

$$\frac{\partial c_t}{\partial \eta} = -c_t^2 \sum_{i=1}^{N} \frac{\partial \tilde{\alpha}_t(i)}{\partial \eta},$$

где $\tilde{\alpha}_t(i)$, $t = \overline{1,T}$, $i = \overline{1,N}$, – так называемая прямая переменная с масштабом [6].

Для вычисления производных

$$\frac{\partial \tilde{\alpha}_t(i)}{\partial a_{xz}}, \frac{\partial \tilde{\alpha}_t(i)}{\partial \mu_{xy}} \text{ и } \frac{\partial \tilde{\alpha}_t(i)}{\partial \sigma_{xy}^2}, \; x, z = \overline{1,N}, \; y = \overline{1,M},$$

используются следующие формулы:

1 шаг. $\frac{\partial \tilde{\alpha}_1(i)}{\partial \eta} = \pi_i \frac{\partial b_i(1)}{\partial \eta}$, $i = \overline{1,N}$;

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

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

Производные $\frac{\partial b_i(t)}{\partial \eta}$ в зависимости от аргумента η имеют вид

$$\frac{\partial b_i(t)}{\partial a_{xz}} = 0; \quad \frac{\partial b_i(t)}{\partial \mu_{xy}} = \tau_{im} \frac{o_t - \mu_{xy}}{\sigma_{xy}^2} \left(\sqrt{2\pi} \sigma_{im} \right)^{-1} e^{-(o_t - \mu_{im})^2 / 2\sigma_{im}^2} \delta_{im}(x, y),$$

где $\delta_{im}(x, y) = \begin{cases} 1, & \textit{если } i = x \textit{ и } m = y, \\ 0, & \textit{иначе}; \end{cases}$

$$\frac{\partial b_i(t)}{\partial \sigma_{xy}^2} = \frac{\tau_{im}}{2\sqrt{2\pi} \sigma_{im}} \frac{\left(o_t - \mu_{xy} \right)^2 - \sigma_{xy}^2}{\sigma_{xy}^4} e^{-(o_t - \mu_{im})^2 / 2\sigma_{im}^2} \delta_{im}(x, y).$$

3. Результаты

Исследования проводились при следующих условиях. Для моделирования последовательностей была выбрана цепь с $N = 4$ скрытыми состояниями. Рассматривалась задача двухклассовой классификации с моделями λ_1 и λ_2, определенными на одинаковых по структуре скрытых марковских цепях и различающимися только в матрицах переходных вероятностей. Параметры моделей будем отличать по их верхнему индексу.

$$A^{\lambda_1} = \begin{pmatrix} 0.2 & 0.5 & 0.3 & 0.05 \\ 0.2 & 0.25 & 0.5 & 0.05 \\ 0.5 & 0.25 & 0.2 & 0.05 \\ 0.2 & 0.25 & 0.5 & 0.05 \end{pmatrix}, \; A^{\lambda_2} = \begin{pmatrix} 0.2 + d^A & 0.5 - d^A & 0.3 & 0.05 \\ 0.2 & 0.25 + d^A & 0.5 - d^A & 0.05 \\ 0.5 - d^A & 0.25 & 0.2 + d^A & 0.05 \\ 0.2 & 0.25 & 0.5 & 0.05 \end{pmatrix}.$$

Параметры гауссовских распределений для модели λ_1 и λ_2 выбирались одинаковыми для каждой модели. Параметр d^A, который можно варьировать в определенных пределах, определял степень близости конкурирующих моделей.

Таким образом, последнее скрытое состояние являлось своего рода шумовым: вероятность перейти в него и остаться в нем очень мала.

Обучающие и тестовые последовательности моделировались по методу Монте-Карло. Для проведения экспериментов было сгенерировано по 5 обучающих наборов последовательностей для каждого класса. К каждому набору этих последовательностей моделировалось по 500 тестовых последовательностей. Результаты классификации усреднялись.

Число скрытых состояний и количество компонент гауссовых смесей при моделировании последовательностей будем обозначать параметрами N и M, а параметры рабочих моделей, используемых при обучении и тестировании, – как N_l и M_l.

На рис. 1 – 3 приведены графики, отражающие результаты классификации:

- для *kNN* в пространстве первых производных от логарифма функции правдоподобия по элементам матрицы переходных вероятностей имеют пунктирную линию;
- для *SVM* в этом же пространстве – имеют штриховую линию с короткими штрихами;
- для *SVM* в объединенном пространстве (включаются все пространства по различным параметрам модели) – имеют штриховую линию с длинными штрихами;
- графики для традиционного подхода – сплошную линию.

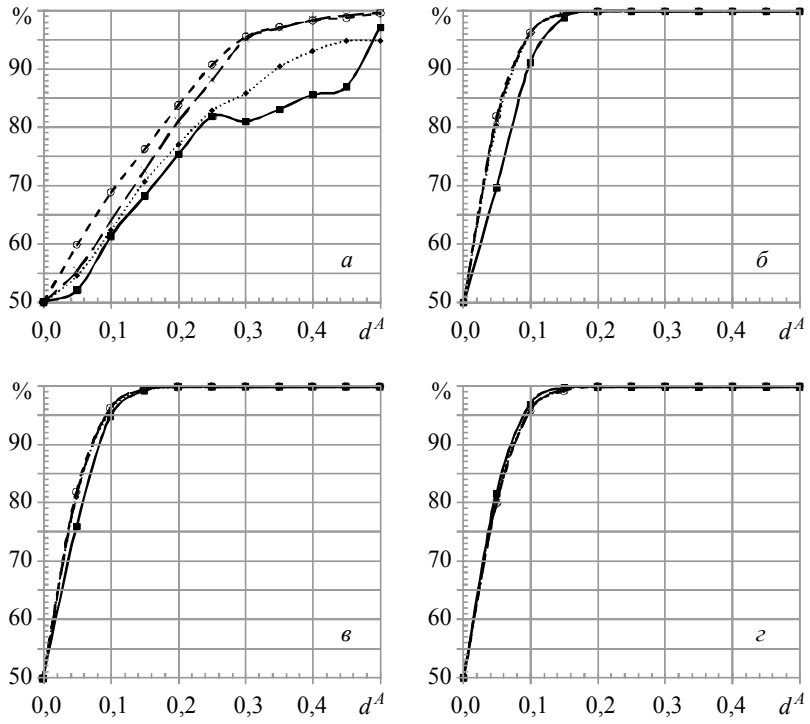

Рис. 1. Зависимость среднего процента верно классифицированных последовательностей от параметра близостей моделей d^A при $M = 2$ и при $N_l = M_l = 2$ (*а*); $N_l = M_l = 3$ (*б*); $N_l = M_l = 4$ (*в*); $N_l = M_l = 6$ (*г*)

На рис. 1 приведены зависимости среднего процента верно классифицированных последовательностей от параметра близостей моделей d^A для случая, когда моделирование производилось при $N = 4$ и $M = 2$. Для рабочих моделей при обучении и тестировании выбирались различные параметры N_l и M_l. На рис. 2 и 3 приведены аналогичные зависимости для $M = 4$ и $M = 6$ соответственно.

По рис. 1 можно отметить, что уже при $N_l = M_l = 3$ у *SVM*- и *kNN*-классификаторов наблюдаются примерно такие же результаты, как $N_l = M_l = 6$, в то время как традиционный классификатор показывает худшие результаты. Аналогичная картина наблюдается и на рис. 2 и 3 при $N_l = M_l = 4$. Таким образом, для классификации по производным, по всей видимости, не нужны особо точные оценки, в то время как для традиционного классификатора этот момент является критическим.

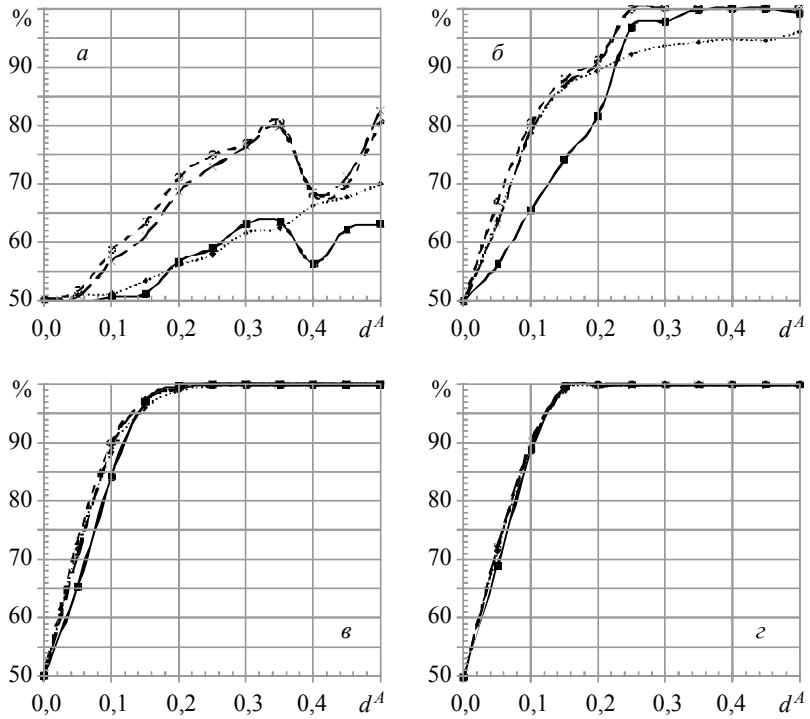

Рис. 2. Зависимость разности среднего процента верно классифицированных последовательностей от параметра близостей моделей d^A при $M = 4$ и при $N_l = M_l = 2$ (а); $N_l = M_l = 3$ (б); $N_l = M_l = 4$ (в); $N_l = M_l = 6$ (г)

При увеличении параметра d^A заметна общая тенденция к уменьшению выигрыша от использования классификаторов в пространстве производных. Это связано с тем, что традиционный классификатор при достаточном различии моделей уже сам показывает хорошие результаты, близкие к 100 %. В то же время анализ рис. 1 – 3 говорит о том, что чувствительность классификаторов *kNN* и *SVM* к ошибкам спецификации структуры СММ несколько ниже, чем у традиционного классификатора. Выигрыш от их использования составил: на рис. 1, *а* – до 14 %, на рис. 2, *а* – до 18 %, на рис. 3, *в* – до 40 %. Общая тенденция такова: чем меньше

взято количество N_l и M_l, тем больший выигрыш в точности классификации можно получить, используя *kNN* или *SVM*.

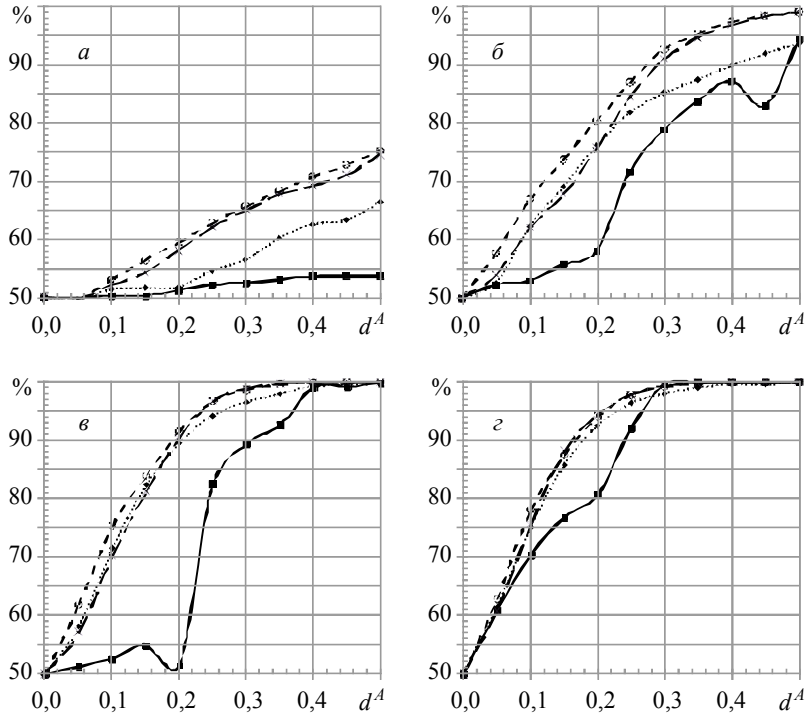

Рис. 3. Зависимость среднего процента верно классифицированных последовательностей от параметра близостей моделей d^A при $M=6$ и при $N_l=M_l=2$ (*а*); $N_l=M_l=3$ (*б*); $N_l=M_l=4$ (*в*); $N_l=M_l=6$ (*г*)

Традиционный классификатор показывает хорошие результаты, когда выбираются количества скрытых состояний и смесей N_l и M_l большие, чем истинные значения N и M.

Таким образом, когда нет возможности провести структурную идентификацию СММ можно рекомендовать использовать классификатор *kNN* или *SVM* в пространстве первых производных от логарифма функции правдоподобия в силу их меньшей чувствительности к такой ошибки спецификации структуры как недобор числа скрытых состояний и компонент гауссовых смесей.

При выборе пространства производных по тому параметру, по которому генерирующие последовательности модели отличаются, выигрыш в сравнении с использованием объединенного пространства, получается максимально 4 % (рис. 1, *а*), 3% (рис. 2, *в*), 6 % (рис. 3, *б*).

Классификатор *SVM* всегда показывает лучшие результаты, чем *kNN* в пространстве первых производных по элементам матрицы переходных вероятностей: до 10 % (рис. 1, *а*), 16 % (рис. 2, *б*), 8 % (рис. 3, *а*). Кроме того, по рис. 2, *а*, *б*, 3, *в* видно, что *kNN* при некоторых значениях параметра d^A проигрывает традиционному классификатору.

Заключение

Исследования показали, что в условиях структурной неопределенности использование классификатора *k* ближайших соседей и классификатора, основанного на методе опорных векторов, приводит к повышению качества классификации в сравнении с традиционным подходом, основанным на отношении логарифмов функций правдоподобия. При этом прирост процентов верной классификации в рассмотренной двухклассовой задаче в сравнении с традиционным подходом в ряде случаев может достигать 40 % как для *kNN*, так и для *SVM*. Последний показывает более точные результаты в сравнении с *kNN*. Максимальное улучшение достигает около 18 %.

ЛИТЕРАТУРА

1. *Rabiner L.R.* A tutorial on hidden markov models and selected applications in speech recognition // Proc. IEEE. 1989. V. 77(2). P. 257–285.
2. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Института математики, 1999. 270 с.
3. *Platt J.C.* Sequential minimal optimization: a fast algorithm for training support Vector Machines [Электронный ресурс]: Technical Report MSR-TR-98-14; Microsoft Research. URL: http://luthuli.cs.uiuc.edu/~daf/courses/Optimization/Papers/smoTR.pdf.
4. *Гультяева Т.А.* Вычисление первых производных от логарифма функции правдоподобия для скрытых марковских моделей // Сб. научных трудов НГТУ. Новосибирск: Изд-во НГТУ, 2010. № 2(60). С. 39–46.
5. *Гультяева Т.А.* Особенности вычисление первых производных от логарифма функции правдоподобия для скрытых марковских моделей при длинных сигналах // Сб. научных трудов НГТУ. Новосибирск: Изд-во НГТУ, 2010. № 2(60). С. 47–52.
6. *Гультяева Т.А., Попов А.А.* Классификация зашумленных последовательностей, порожденных близкими скрытыми марковскими моделями // Научный вестник НГТУ. Новосибирск: Изд-во НГТУ, 2011. № 3(44). С. 3–16.

Гультяева Татьяна Александровна
Попов Александр Александрович
Новосибирский государственный технический университет
E-mail: gult_work@mail.ru, alex@fpm.ami.nstu.ru

Поступила в редакцию 12 апреля 2012 г.

Gultyaeva Tatyana A., Popov Alexander A. (Novosibirsk State Technical University). **Classification of sequences with use of hidden Markov models under conditions of the inexact task of their structure.**

Keywords: hidden Markov models, derivative of log likelihood function, classifier of k nearest neighbors, support vector machines.

The problem of sequences classification with use of methodology of hidden Markov models (HMM) is considered. Classification is spent both with use of the standard approach, and with classifier of *k* nearest neighbors (*kNN*), and a support vector machines in space of the signs initiated by HMM. The behavior of qualifiers was investigated while errors in the structure specification of HMM are presented.

Researches have shown that under conditions of structural uncertainty use of qualifiers of *k* nearest neighbors and the classifier based on a SVM, leads to improvement of classification quality in comparison with the traditional approach based on the ratio of logarithms of likelihood function. Thus the gain in correct classification in the considered two-class problem in comparison with the traditional approach in some cases can reach 40 % both for *kNN*, and for *SVM*. The last shows more exact results in comparison with *kNN*. The maximum improvement reaches about 18 %.