Научный вестник НГТУ. – 2013. – № 3(52)

УДК 519.217.2

Классификация смоделированных скрытыми марковскими моделями последовательностей в многоклассовом случае*

Т.А. ГУЛЬТЯЕВА, А.А. ПОПОВ

В работе исследуется возможность повышения дискриминирующих свойств скрытых марковских моделей путем использования вторичных признаков, инициированных этими моделями, с применением классификатора, основанного на методе опорных векторов. Рассматривается случай, когда исследователь не обладает точными знаниями о структуре близких между собой по параметрам скрытых марковских моделей, которые смоделировали классифицируемые последовательности.

Ключевые слова: скрытые марковские модели, производные от логарифма функции правдоподобия, метод опорных векторов, многоклассовая классификация

ВВЕДЕНИЕ

Одним из средств моделирования различных процессов являются скрытые марковские модели (СММ) [7, 8, 11]. Особенностью таких моделей является то, что они учитывают внутреннюю структуру исследуемого явления, опираясь на то предположение, что события, происходящие в этом явлении, приводят к появлению характерных особенностей в наблюдаемых последовательностях. Такие модели имеют хорошие описательные способности, но не всегда демонстрируют необходимые дискриминирующие свойства, которые важны для задачи классификации.

В работе в качестве объектов классификации рассматривается множество смоделированных последовательностей, порожденных несколькими близкими по своим параметрам СММ. Классификация последовательностей с использованием СММ при условии того, что конкурирующие модели достаточно хорошо отличимы друг от друга (по вероятности), обычно не вызывает затруднений. Плохо поддаются классификации последовательности, порожденные близкими по своим параметрам СММ. Будем также предполагать, что у исследователя нет априорной информации о структуре СММ. Проведем сравнение в этих условиях возможностей традиционной методики классификации, основывающейся на вычислении вероятности того, что последовательность порождена конкретной СММ, с возможностями классификатора в виде метода опорных векторов *SVM* (Support Vector Machines) [10] в различных пространствах признаков. В качестве признаков будем использовать первые производные от логарифма функции правдоподобия по параметрам СММ.

Результаты, приведенные в статье, являются продолжением серии экспериментов по двухклассовой классификации последовательностей, смоделированных близкими по параметрам СММ в условиях зашумленности генерируемых ими последовательностей [3–5].

* *Статья получена 10 января 2013 г.*

1. КЛАССИФИКАЦИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

СММ описывается матрицей вероятностей переходов, функциями плотности вероятностей распределения наблюдаемых символов и вероятностями начальных состояний: $\lambda = (A,\ B,\ \pi)$, определения которых приведены ниже.

1. Вектор вероятностей начальных состояний $\Pi = \{\pi_i\}$, $i = \overline{1,N}$, где $\pi_i = P\{q_1 = s_i\}$, где q_t – скрытое состояние в момент времени t, множество скрытых состояний $S = \{s_1, s_2, ..., s_N\}$, N – количество скрытых состояний в модели.

2. Матрица вероятностей переходов $A = \{a_{ij}\}$, $i, j = \overline{1,N}$, где $a_{ij} = P\{q_t = S_j \,|\, q_{t-1} = S_i\}$, где q_t – скрытое состояние в момент времени t, $t = \overline{1,T}$, T – длина наблюдаемой последовательности.

3. Функции плотности вероятностей распределения наблюдаемых символов $B = \{b_i(t)\}$, где $b_i(t)$ – это плотности условных вероятностей $P\{o_t \,|\, q_t = s_i\}$, o_t – символ из последовательности наблюдений, наблюдаемый в момент времени $t = \overline{1,T}$. В данной работе рассматривается случай, когда плотности вероятностей распределения наблюдаемых символов описываются смесью нормальных распределений, а сами наблюдаемые символы являются одномерными, т. е. функция плотности вероятности имеет вид

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

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

Обычно для СММ используется классификатор, основанный на отношении логарифмов функций правдоподобия: последовательность O считается порожденной моделью λ_1, если выполняется неравенство

$$\ln L(O \,|\, \lambda_1) \geq \ln L(O \,|\, \lambda_2). \qquad (2)$$

Иначе – считается, что последовательность порождена моделью λ_2.

Было установлено, что если конкурирующие модели близки по параметрам, а наблюдаемые последовательности не являются чисто гауссовскими последовательностями, то традиционная техника классификации с применением (2) далеко не всегда дает приемлемые результаты [3–5].

В качестве альтернативы предлагается использовать классификаторы, использующие признаки, извлекаемые из обученных СММ. Обучение проводится, например, с использованием алгоритма Баум–Велша [10]. Поскольку этот алгоритм чувствителен к выбору начального приближения и может сходиться к локальному максимуму функции правдоподобия, то был использован алгоритм случайного ненаправленного поиска глобального экстремума.

В качестве пространств признаков, в которых производится классификация, рассматриваются пространства первых производных от логарифма функции правдоподобия по различным параметрам (более подробно см. [1, 2]). Для каждой обучающей и тестовой последовательностей формируется характеристический вектор, который состоит из двух подвекторов. В первый из них вошли признаки, инициированные первой моделью, а во второй – соответст-

венно второй моделью. Заметим, что для решения задачи многоклассовой классификации была использована схема, называемая «турниром на выбывание» [9].

Обучающие и тестовые последовательности моделировались по методу Монте–Карло. Для проведения экспериментов было сгенерировано по 5 обучающих наборов по 100 последовательностей для каждого класса. К каждому набору этих последовательностей моделировалось по 500 тестовых последовательностей.

2. ПОВЕДЕНИЕ КЛАССИФИКАТОРОВ В УСЛОВИЯХ СТРУКТУРНОЙ НЕОПРЕДЕЛЕННОСТИ

В реальных ситуациях априорная информация о структуре СММ, как правило, отсутствует. Задача структурной идентификации в этом случае может быть поставлена, но на практике чаще всего она не решается в полном объеме.

Исследование проводилось при одинаковом для пяти конкурирующих моделей количестве скрытых состояний $N = 3$, но при разном количестве наблюдаемых состояний: $N^{\lambda_{class}} = (class + 1), class = \overline{1,5}$. При этом матрица переходных вероятностей была для всех СММ одинаковой:

$$A = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.7 \end{pmatrix}.$$

Веса смесей: $\tau_{ij}^{\lambda_{class}} = \left(M^{\lambda_{class}}\right)^{-1}$, $i = \overline{1,N}$, $j = \overline{1, M^{\lambda_{class}}}$, $class = \overline{1,5}$.

Для первой модели параметры математического ожидания смесей нормальных распределений были заданы следующим образом:

$$\mu^{\lambda_1} = \begin{pmatrix} 10 & 50 \\ -100 & -70 \\ -45 & -10 \end{pmatrix}.$$

При этом у каждой последующей модели добавлялось новое наблюдаемое состояние, у которого параметры математического ожидания для нормального распределения имели значения, равные среднему между предыдущими значениями:

$$\mu_{ij}^{\lambda_{class}} = \min\left\{\mu_{ij-2}^{\lambda_{class}}, \mu_{ij-1}^{\lambda_{class}}\right\} + 0.5\left|\mu_{ij-2}^{\lambda_{class}} - \mu_{ij-1}^{\lambda_{class}}\right|,$$

где $i = \overline{1,N}$, $j = \overline{3, M^{\lambda_{class}}}$, $class = \overline{2,5}$. При этом $\mu_{ij}^{\lambda_{class}} = \mu_{ij}^{\lambda_1}$ для $i = \overline{1,N}$, $j = \overline{1,2}$ и $class = \overline{2,5}$.

Параметры дисперсии и вектор начальных состояний были выбраны одинаковыми для всех моделей: $\sigma_{ij}^2 = 1$, $i, j = \overline{1,N}$, $class = \overline{1,5}$, $\pi = (1 \quad 0 \quad 0)$. При таком задании параметров смесей наблюдаемые состояния у конкурирующих моделей становились малоразличимы.

На рис. 1 приведены зависимости среднего процента верно классифицированных последовательностей с использование традиционного метода и предложенного подхода с классификатором SVM от количества скрытых и наблюдаемых состояний, используемых на этапе обучения. С увеличением числа состояний СММ более точно описывают наблюдаемые последовательности, поэтому заметен постоянный рост среднего процента верно классифицированных последовательностей как для традиционного метода, так и для предложенного подхода. Колонки гистограмм отображают результаты: для традиционного метода классификации (белый цвет); в пространстве производных по элементам матрицы переходных вероятностей (светло-

серый); в пространстве производных по параметрам математического ожидания (серый); в пространстве производных по параметрам дисперсии (темно-серый); в пространстве всех этих производных по параметрам математического ожидания (черный).

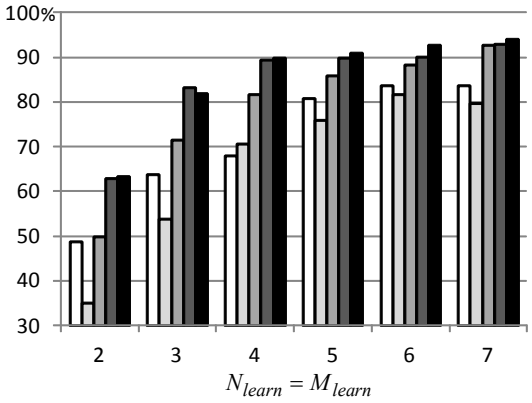


Рис. 1. Зависимость среднего процента верно классифицированных последовательностей от значения параметров N_{learn} и M_{learn}, $T = 100$

При различном числе состояний $N_{learn} = M_{learn}$ наилучшего результата SVM достигает в разных пространствах: либо в пространстве производных по параметрам дисперсии, либо в пространстве, состоящем из производных по всем параметрам. При этом средний выигрыш в сравнении с традиционным методом с использованием первого или второго пространства составляет 14 %, в пространстве производных по параметрам математического ожидания – 7 %, а в пространстве по элементам матрицы переходных вероятностей ухудшение на 7 %. Можно сделать вывод: если исследователь выбрал недостаточное количество состояний, то, используя предложенный подход, он может получить результат при классификации последовательностей гораздо более лучший, чем при использовании традиционного метода.

3. МЕТОДЫ ВЫБОРА ИНФОРМАТИВНЫХ ПРИЗНАКОВ

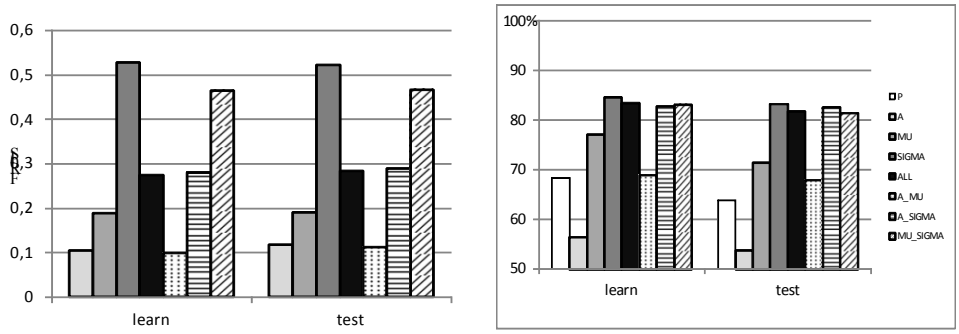
Далее рассмотрено, как можно выбрать пространство признаков, которое обеспечивало бы наилучший процент верно классифицированных последовательностей. В данной работе используются прямой и косвенный подходы для определения информативности признаков, а именно: метод скользящего экзамена [9] и критерий, основанный на функции конкурентного сходства. Функция *FRiS* (Function of Rival Similarity) позволяет вычислять количественную оценку компактности конкурирующих классов [6]. Кроме того, третьим подходом к выбору пространства признаков, обеспечивающего наилучшие результаты классификации, можно отнести решение задачи классификации в каждом из пространств с последующим выбором нужного пространства.

Функция конкурентного сходства для произвольного объекта $z \in X$ со своим классом вычисляется по формуле

$$F(z) = \big(r_2(z, X) - r_1(z, X)\big)\big(r_2(z, X) + r_1(z, X)\big)^{-1},$$

где $r_1(z, X)$ – расстояние до ближайшего объекта своего класса и $r_2(z, X)$ – расстояние до ближайшего объекта конкурирующего класса. В работе [6] было показано, что существует связь между надежностью распознавания со значением *FRiS*-функции: чем больше значение *FRiS*-функции для объекта z, тем больше надежность его верного распознавания (т. е. выше вероятность того, что решение о принадлежности данного объекта к классу правильно).

В данном пункте приводятся результаты для пяти конкурирующих моделей при $N_{learn} = M_{learn} = 3$. На рис. 2, *а* приведены значения функции конкурентного сходства. На рис. 2, *б* слева приведены оценки среднего процента верно классифицированных последовательностей на обучающих наборах, полученные с помощью метода скользящего экзамена; справа – значения среднего процента верно классифицированных последовательностей на тестовых наборах. На гистограммах, приведенных на рис. 2, используются те же оттенки серого цвета, что и на рис. 1. Кроме того, результаты, полученные с использованием пространства производных по элементам матрицы переходных вероятностей и по параметрам математического ожидания, отображаются колонками с точками; пространства производных по элементам матрицы переходных вероятностей и по параметрам дисперсии – с горизонтальными линиями; пространства производных по параметрам математического ожидания и по параметрам дисперсии – с наклонными линиями.



а *б*

Рис. 2. Значения функции конкурентного сходства (*а*) и среднего процента верно классифицированных последовательностей (*б*) на обучающих и тестовых наборах

Критерий, основанный на функции конкурентного сходства так же, как и метод скользящего экзамена на обучающей выборке, показали, что необходимо выбрать пространство, состоящее из производных по параметрам дисперсии. В этом пространстве наблюдается наилучший процент верно распознанных последовательностей. Заметим, что самым затратным по временным ресурсам является метод, основанный на перекрестном экзамене, так как он предполагает многократное построение решающего правила классификации. Наименее затратный в этом смысле метод, использующий значения *FRiS*-функции.

ЗАКЛЮЧЕНИЕ

Исследования показали, что в условиях структурной неопределенности использование 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 многоклассовой задаче в сравнении с традиционным подходом в ряде случаев может достигать 25 %.

Таким образом, когда нет возможности провести структурную идентификацию СММ, можно рекомендовать использовать рассматриваемый в данной работе подход к классификации последовательностей в силу его меньшей чувствительности к такой ошибке спецификации структуры как недобор числа скрытых состояний и компонент гауссовых смесей.

СПИСОК ЛИТЕРАТУРЫ

[1] **Гультяева Т.А.** Вычисление первых производных от логарифма функции правдоподобия для скрытых марковских моделей / Т.А. Гультяева // Сб. науч. тр. НГТУ. – 2010. – № 2(60). – С. 39–46.

[2] **Гультяева Т.А.** Особенности вычисление первых производных от логарифма функции правдоподобия для скрытых марковских моделей при длинных сигналах / Т.А. Гультяева // Сб. науч. тр. НГТУ. – 2010. – № 2(60). – С. 47–52.

[3] **Гультяева Т.А.** Классификация зашумленных последовательностей, порожденных близкими скрытыми марковскими моделями / Т.А. Гультяева, А.А. Попов // Научный вестник НГТУ. – 2011. – № 3(44). – С. 3–16.

[4] **Гультяева Т.А.** Классификация последовательностей, смоделированных скрытыми марковскими моделями при наличии аддитивного шума / Т.А. Гультяева, А.А. Попов // Научный вестник НГТУ. – 2012. – № 3(48). – С. 17–24.

[5] **Гультяева Т.А.** Исследование возможностей применения алгоритма k ближайших соседей и метода опорных векторов для классификации последовательностей, порожденных скрытыми марковскими моделями / Т.А. Гультяева, Д.Ю. Коротенко // Сб. науч. тр. НГТУ. 2011. – № 3(65). – С. 45–55.

[6] **Загоруйко Н.Г.** Когнитивный анализ данных / Н.Г. Загоруйко. – Новосибирск: Академическое изд-во «ГЕО», 2012. – 186 с.

[7] **Моттль В.В.** Скрытые марковские модели в структурном анализе сигналов / В.В. Моттль, И.Б. Мучник. – М.: Физматлит, 1999. – 351 с.

[8] **Cappe O.** Ten years of HMM [Электронный ресурс] / O. Cappe; CNRS, LTCI & ENST, Dpt. TSI. – Режим доступа: http://perso.telecom-paristech.fr/~cappe/docs/hmmbib.html.

[9] **Friedman H.** Another approach to polychotomous classification. Technical report, Stanford Department of Statistics, 1996.

[10] **Platt J.C.** Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines [Электронный ресурс]: Technical Report MSR-TR-98-14 / J.C. Platt; Microsoft Research. – Режим доступа: http://luthuli.cs.uiuc.edu/~daf/courses/Optimization/Papers/smoTR.pdf.

[11] **Rabiner L.R.** A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition / L.R. Rabiner // Proceedings of the IEEE. – 1989. – 77 (2). – С. 257–285.

Попов Александр Александрович, доктор технических наук, профессор, заведующий кафедрой программных систем и баз данных Новосибирского государственного технического университета. Основные направления научных исследований – статистические методы анализа данных, оптимальное планирование эксперимента. Имеет более 100 публикаций, в том числе 1 монографию. E-mail: alex@fpm.ami.nstu.ru.

Гультяева Татьяна Александровна ассистент кафедры программных систем и баз данных Новосибирского государственного технического университета. Основные направления научных исследований – скрытые марковские модели, статистические и структурные методы распознавания. Имеет 33 публикации. E-mail: gtany@mail.ru.

T.A. Gultyaeva, A.A. Popov

The Classification in a Multiclass Case of the Sequences Generated by Hidden Markov Models

In work possibility of increase of discriminating properties Hidden Markov Models (HMM) by use of the secondary features initiated by these models with use of different classifiers in case of multiclass classification is investigated. The case when the researcher doesn't possess exact knowledge of structure of relatives among themselves on parameters of the HMM models which simulated classified sequences is considered.

Key words: Hidden Markov Models, Derivative of Likelihood Function, Support Vector Machines, Multiclass Classification.