СКРЫТАЯ МАРКОВСКАЯ МОДЕЛЬ В OCR И ЕЕ ПРИМЕНЕНИЕ В ДИСТАНЦИОННОМ ОБУЧЕНИИ

Попов А.О.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь, andrei.papou96@gmail.com

Abstract. The article describes what hidden markov model is, how to create it based on markov model and how it can be used in optical character recognition and distance education, what issues people may face during the process of recognition and what are the solution to these issues.

Оптическое распознавание символов может применяться для распознавания текста из любого мультимедиа, такого как изображение, аудио, видео [1].

Распознавание текста при помощи OCR может также применяться при организации дистанционного образования. К примеру, получение и обработка документов, необходимых для передачи или проверки знаний студентов может проводиться при помощи создания специального программного обеспечения, которое будет преобразовать отсканированные документы в набор цифровых данных, который далее будет использоваться в системах по оценке знаний студентов или организации процесса обучения. В статье рассматривается скрытая марковская модель. Данная модель позволяет распознавать текст или символы с очень высоким шансом [2].

Основой для рассматриваемой модели является марковская цепь, которая может быть описана как весовой конечный автомат, который содержит в себе конечный набор N состояний $S = \{s_1, s_2, s_3, s_4...\}$ и набор переходов между этими состояниями. Для каждого состояния существует вероятность π_i, что модель начнет с именно этой точки. Сумма вероятностей для всех состояний равна единице. Также на основе набора вероятностей для каждого состояния создается набор вероятностей переходов $A = \{a_{ij}\}$. Данный набор описывает вероятность перехода из состояния i в состояние j. На рисунке 1 изображена простая марковская цепь с тремя состояниями: s_1, s_2, s_3. Вероятность перехода из состояния s_i в состояние s_j отображена на дуге, которая соединяет эти два состояния.

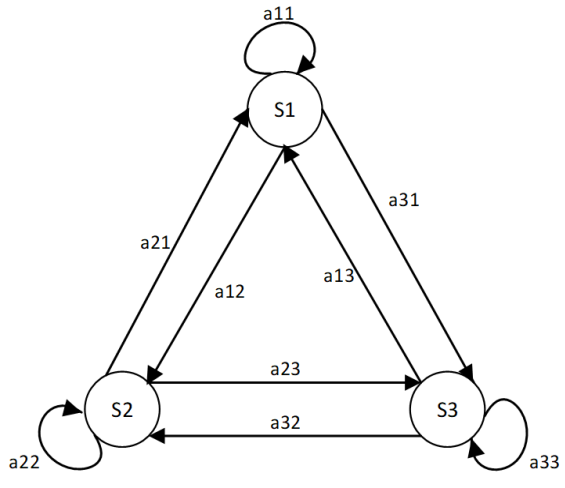

Рисунок 1 – Пример марковской модели с тремя состояниями

Модель предполагает, что был создан марковским источником информации, в котором новые символы зависят только от фиксированного количества предшествующих элементов. Количество элементов зависит от порядка модели. Чаще всего применяются марковские модели первого и второго порядка в связи с тем, что при повышении порядка возрастает и сложность модели, что в свою очередь снижает ее полезность. Для марковской модели первого порядка вероятность определенного состояния зависит только от предыдущего. Определение имеет следующий вид, представленный на формуле (1):

$$P(q_t|q_{t-1} \dots q_1) = P(q_t|q_{t-1}) \quad (1)$$

где q – случайный процесс, участвующий в модели; P – вероятность наступления определенного процесса.

Для марковской модели второго порядка вероятность наступления состояния s_i в момент времени t зависит от состояния s_j, находящегося в моментах t-2 и t-1. Количество m предыдущих состояний, от которых будет зависеть вероятность наступления следующего состояния, является порядком марковской модели.

Процесс является наблюдаемым. Это означает, что все события, которые присутствуют в данной модели являются воспроизводимыми и физическими событиями. При распознавании текста не всегда есть возможность точно предсказать последовательность событий, которая приведет к желаемому результату. В этом помогает скрытая марковская модель. Данная модель является дважды стохастическим вариантом марковской модели, в которой наблюдение за одним стохастическим процессом осуществляется при помощи набора других стохастических процессов, результатом которых является набор наблюдаемых символов. Модель представлена как взаимосвязанный набор состояний, которые соединены между собой набором вероятностей переходов [2] Начало происходит в каком-нибудь из состояний, после чего переходит в новое в зависимости от вероятности перехода. После чего вырабатывается набор символов, какой именно из них будет выбран в этой последовательности будет определено вероятностью результата, которая свойственна состоянию. В результате работы модели производится набор последовательность символов, а так как последовательностей событий, которые могут привести к определенной выходной последовательности символов, множество, последовательность состояний является скрытой.

Преобразуем предыдущую модель в скрытую марковскую модель. Для этого добавим в каждое состояние два символа. 0 и 1 – это наблюдаемые символы каждого состояния. Тогда b_{ij} отвечает за вероятность результата каждого состояния. Пример модели представлен на рисунке 2.

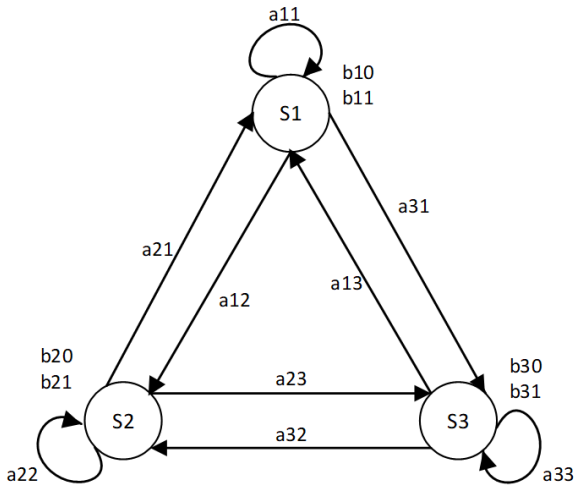

Рисунок 2 – Пример скрытой марковской модели с тремя состояниями

Каждое состояние содержит набор параметров, где $S = \{ s_1, s_2, s_3, s_4 \ldots\}$ – конечный набор N состояний, $V = \{v_1, v_2, v_3, v_4 \ldots m\}$ – набор размерностью M возможных символов в словаре, вероятность π_i, что модель начнет с именно этой точки, набор вероятностей переходов перехода из состояния i в состояние j $A = \{a_{ij}\}$ и $B = \{b_i(v_k)\}$ – набор вероятностей результата, где $b_i(v_k)$ – вероятность генерации символа v_k в состоянии i.

У скрытой марковская модели тоже есть свои недостатки [4]. Первая проблема – это оценка. Предположим, что дана последовательность наблюдений $O = \{o_1, o_2, o_3, o_4 \ldots o_t\}$ и модель λ. Как высчитать вероятность того, что последовательность наблюдений была выработана этой моделью? Данную проблему решает алгоритм прямого-обратного хода.

Вторая проблема – это декодирование. Различные последовательности состояний могут выдать одну и ту же наблюдаемую последовательность. Для каждой наблюдаемой последовательности нужно выбрать те состояния, которые будут иметь наивысшую вероятность выработки наблюдений для нахождения оптимальной последовательности символа. Для решения такой задачи часто используется алгоритм Витерби.

Еще одна проблема в оптическом распознавании символов – обучение. Задача состоит в том, чтобы правильно подобрать параметры модели для того, чтобы как можно больше увеличить возможность генерации наблюдаемой последовательности. Наиболее популярные методы по обучению модели являются метод максимального правдоподобия и алгоритм Баума-Велша, который в свою очередь использует алгоритм прямого-обратного хода. Основные ошибки, которые могут встречаться при обучении модели по распознаванию символов – это пропуск символа, вставка лишних символов, замена одного символа другим, замена двух символов на один и наоборот и замена двух символов на два других символа [3].

Также источником проблем при оптическом распознавании символов является качество документов, предоставляемых на обработку, различных шрифты, которые. Курсивный стиль и шрифтовые шрифты символов могут перекрывать друг друга, что затрудняет выполнение некоторых основных процессов распознавания, таких как сегментация. Символы различных шрифтов имеют большие вариации внутри класса и образуют множество подпространств шаблона, что затрудняет точное распознавание, когда число классов символов велико.

Поскольку работа на различных расстояниях предназначена для многочисленных цифровых камер, важным фактором является фокусировка цифровой камеры. Для лучшей точности распознавания символов и сегментации символов требуется четкость символов. При больших значениях диафрагмы и коротких расстояниях неравномерная фокусировка может наблюдаться при изменении маленькой точки зрения. По большей части, связанных с фотографией, существует два вида затемнения: нечеткость неясности и неясность движения. В точке захвата движущегося объекта, когда скорость затенения камеры недостаточно высока, датчик отображается постоянно меняющейся сцене. Соответственно, размытие будет наблюдаться в деталях в движении.

Обучение и тренировка модели являются двумя важными шагами в классификации. Изначально данные приводятся к виду, подходящему для обучения, далее происходит выборка, цель которой заключается в уменьшении количества данных, путем взятия только нужной информации. При эстимации модели из конечного набора векторов производится оценка модели для каждого класса в наборе данных.

Литература

1. Hiral Modi, M. C. Parikh “A Review on Optical Character Recognition Techniques” International Journal of Computer Applications (0975 – 8887). – Volume 160 – No 6. – February, 2017.

2. J. Hu, S.G. Lim and M.K. Brown “Writer independent on-line handwriting recognition using an HMM approach,” J. PATTERN Recognit. Soc. – vol. 33. – p. 133– 147. – 2000.

3. J. Esakov, D.P. Lopresti and J.S. Sandberg, “Classification and distribution of optical character recognition errors,” in Proceedings of the IS&T/SPIE International Symposium on Electronic Imaging. – San Jose, CA. – February 1994.

4. S.M. Thede and M.P. Harper “A second-order Hidden Markov Model for part-of-speech tagging,” in Proceedings of the 37th Annual Meeting of the ACL, pp. 175–182. – 1999.