Обучение вероятностных конечных автоматов

Автор: Степанов Всеволод
Руководитель: Кураленок Игорь Евгеньевич

6 семестр

Мотивация

Сейчас основной подход для работы с последовательностями — рекуррентные нейронные сети.

Цель

Придумать и реализовать другой подход для работы с последовательностями, догнать нейронные сети по точности

Задачи

1. ~~Обучение детерминированного автомата~~
2. Обучение вероятностного автомата
3. Бенчмарки

LSTM

LSTM — популярная сейчас архитектура RNN

Важная особенность: решена проблема затухания градиента

План

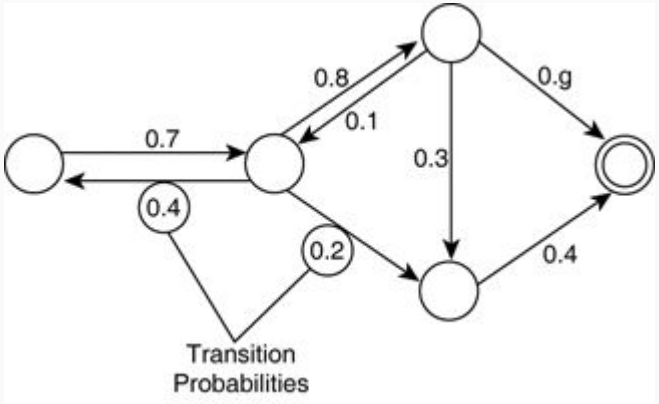
1. Увеличение размера алфавита
2. Gradient boosting: регрессия вместо классификации, комбинирование слабых моделей
3. Обучение слабой модели — автомата

Вероятностный автомат

У каждого перехода есть вероятность

P_c — матрица перехода для символа c

v_i — значение в вершине i



Слабая модель

$s = \{s_k\}$ — последовательность, которая дается на вход

$t = (1, 0 \ldots, 0)$ — начальное распределение по состояниям

$h(s) = t \cdot (\prod P_{s_k})v$ — гипотеза

$$\sum_{s,y} (h(s) - y)^2 \to min$$

Обучение

- Стохастический градиентный спуск
- Затухание градиента, если автомат далек от детерминированного

Улучшения

- Хотим автомат, близкий к детерминированному
- подбор начальных весов — матрица с диагональным преобладанием
- регуляризация:

$$(h(s) - y)^2 - \lambda \sum_{i,j,k} P_{ijk}^2 \to min$$

Бенчмарк

	DFA	PFA	LSTM
Splice dataset	94%	88%	90%
IMDB dataset	—	86%	87%

Ссылка на репозиторий:
https://github.com/spbsu-ml-community/jmll/tree/seva

Splice dataset
https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Splice-junction+Gene+Sequences)

IMDB dataset
http://ai.stanford.edu/~amaas/data/sentiment/