Применение тематического моделирования в информационном анализе электрокардиосигналов

Шапулин Андрей Валентинович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: a.shapulin@gmail.com

Тематическое моделирование является инструментом статистического анализа текстов и предназначено для выявления латентной тематики коллекций текстовых документов. Тематическая модель представляет каждую тему в виде дискретного распределения на множестве слов, а каждый документ — в виде дискретного распределения на множестве тем. Задача поиска такого представления имеет бесконечно много решений и является некорректно поставленной. Для повышения устойчивости решения к модели предъявляются дополнительные требования. Аддитивная регуляризация тематических моделей (АРТМ) позволяет комбинировать требования в произвольных сочетаниях путём оптимизации логарифма правдоподобия модели с линейной комбинацией регуляризаторов [1].

В последнее время тематическое моделирование всё чаще применяется при анализе сигналов, изображений и видеопоследовательностей — в областях, далёких от обработки естественного языка.

Информационный анализ электрокардиосигналов основан на преобразовании электрокардиосигнала в кодограмму — символьную последовательность, кодирующую знаки приращений интервалов и амплитуд R-зубцов [2, 3]. Кодограмма рассматривается как текстовый документ, короткие подпоследовательности символов — как слова. Для каждой кодограммы имеется список установленных диагнозов обследуемого, и задача заключается в построении алгоритма диагностики. Для решения данной задачи в [3] использовались линейные методы классификации с отбором признаков. В терминах тематического моделирования это эквивалентно предположению, что каждое заболевание хорошо описывается одной темой — вероятностным распределением на множестве слов. Высокие уровни чувствительности и специфичности позволяют утверждать, что генерируемые сердцем сигналы несут существенную информацию для диагностики различных заболеваний внутренних органов, причём не только сердечно-сосудистой системы.

В данной работе ставится задача проверить более общее предпо-

ложение, что каждое заболевание может быть ещё лучше описано вероятностной смесью нескольких тем. Тематические модели успешно применялись в [4] для решения сложных задач классификации с большим числом несбалансированных, пересекающихся, взаимозависимых классов. Однако непосредственное применение таких моделей к кодограммам ЭКГ не дало выигрыша качества классификации по сравнению с линейными моделями SVM, логистической регрессии и даже наивного байесовского классификатора.

Более успешным оказалось применение APTM с комбинацией регуляризаторов. Регуляризаторы разреживания и декоррелирования позволяют находить диагностические эталоны каждого заболевания — попарно различные темы, состоящие из небольшого числа слов. Регуляризатор сглаживания позволяет выделять фоновую тему, не специфичную ни для какого заболевания (в обработке естественного языка аналогом является выделение общей лексики языка). Регуляризатор отбора тем позволяет уменьшать сложность модели и сокращать переобучение. Кроме того, применялась инициализация тематической модели путём построения решающего списка из наивных байесовских классификаторов.

Эксперименты проводились на выборке 11 894 кодограмм с диагнозами по 18 заболеваниям по критерию 10-кратной кросс-валидации AUC и показали преимущество регуляризированных тематических моделей.

Литература

1. Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. — Т. 455., №3. 268–271.
2. Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. М.: Экономика и информатика, 2008. 116 С.
3. Uspenskiy V. M., Vorontsov K. V., Tselykh V. R., Bunakov V. A. Information Function of the Heart: Discrete and Fuzzy Encoding of the ECG-Signal for Multidisease Diagnostic System // in Advances in Mathematical and Computational Tools in Metrology and Testing X (vol.10), Series on Advances in Mathematics for Applied Sciences, vol. 86, World Scientific, Singapore (2015) pp. 375–382.
4. Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, № 1–2.