Секция «Дискретная математика и математическая кибернетика»

Персистентные гомологии марковских цепей и их применение в компьютерном анализе текстов на естественном языке

Научный руководитель – Ирматов Анвар Адхамович

Кушнарева Л.П.[1], *Кузъминых Д.В.*[2]

1 - Московский государственный университет имени М.В.Ломоносова, Механико-математический факультет, Кафедра математической теории интеллектуальных систем, Москва, Россия, *E-mail: K.Lidia@list.ru*; 2 - Московский физико-технический институт, Москва, Россия, *E-mail: kuzminykh@phystech.edu*

Персистентные гомологии - один из основных инструментов топологического анализа данных - молодой дисциплины, которая исследует возможности выделения внутренней структуры в экспериментальных данных различной природы и введения топологических инвариантов на ней. Введение таких инвариантов позволяет примененять методы алгебраической топологии для анализа данных и решения соответствующих прикладных задач. В частности, в статье [1] приведен пример применения топологических методов для анализа изображений. Автор данной статьи сопоставляет каждому фрагменту изображения 3x3 вектор в пространстве размерности 9. Далее он показывает, что векторы, соответствующие фрагментам реальных фотографий, лежат в окрестности подмногообразия существенно меньшей размерности, чем размерность пространства векторов, соответствующих произвольным (случайно сгенерированным) фрагментам. Таким образом, появляется возможность до некоторой степени точности описать с точки зрения топологии, чем осмысленные изображения отличаются от бессмысленных.

Марковские цепи также являются важным прикладным инструментом анализа данных, позволяющим моделировать процессы в большом количестве приложений. Однако, по описанию марковских цепей в явном виде может быть трудно оценить 'структурное' сходство или различие таких процессов. Это приводит к вопросу о том, могут ли марковские цепи быть классифицированы аналогично тому, как это сделано для комплексов и поверхностей методами алгебраической топологии.

В данной работе мы предлагаем метод введения топологических инвариантов для марковских цепей, в качестве основы используя идеи из статьи [1].

А именно, мы вводим на марковских цепях инвариант - аналог персистентных гомологий, описанных в статье [1], который позволяет выделять отличительные признаки марковских цепей различных видов. В нашей работе приводится подробное определение этой новой математической конструкции, а также доказательство того, что она действительно обладает свойством персистентности относительно введенного нами параметра значимости (пороговой вероятности).

В качестве основного примера марковской цепи в данной работе используются марковская цепь, построенная на основе текста на естественном языке. С помощью вычислительного эксперимента продемонстрировано, как можно использовать введенные нами топологические инварианты на практике в комплексе с простыми методами классического машинного обучения, чтобы отличать цепи, построенные по осмысленным текстам, от цепей, построенных по случайно сгенерированным текстам с той же частотой слов. В качестве примеров "осмысленных" текстов были использованы тексты из базы данных The Blog Authorship Corpus, особенности которой подробно описаны в статье [2].

В частности, в эксперименте были вычислены нулевые персистентные гомологии марковских цепей, построенных по указанным текстам, для 8 различных пороговых значений параметра значимости (для каждого текста). Таким образом, мы сопоставили каждому

тексту 8 чисел и использовали их в качестве признаков для обучения логистической регрессии. Логистическая регрессия обучалась с помощью стохастического градиентного спуска с параметром регуляризации C = 1 и максимальным количеством итераций 1000. Эксперимент показал, что нулевые группы гомологий графов, соответствующих осмысленным и бессмысленным текстам, линейно разделимы с точностью 0.96-1, что демонстрирует практическую применимость разработанного в данной работе математического инструмента к анализу реальных данных.

Источники и литература

1) 1. Gunnar Carlsson. Topology and Data. Bulletin of the American Mathematical Society №46, pp. 255-308, April 2009
2) 2. J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of Age and Gender on Blogging. Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.
3) 3. Ширяев А.Н. Вероятность. 3-е изд., перераб. и доп. - М.: Изд-во МЦНМО, 2004.
4) 4. М.С. Цаленко, Е.Г. Шульгейфер. Лекции по теории категорий. Москва, Институт Механики МГУ, 1970.