ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

Векторные представления данных в геномном анализе

Ромашко Дмитрий Александрович, аспирант
Дальневосточный федеральный университет (г. Владивосток)

Группа методов ”word embedding” (часто переводится как векторное представление слов) широко используется в анализе естественных языков. Данные методы основаны на представлении слов или фраз через словарь, сопоставляющий их численным векторам. Общий смысл подхода в том, чтобы выразить через вектор как можно больше семантических, морфологических, контекстных или иерархических данных об объекте. Способ и особенности представления варьируются в зависимости от метода.

Геномные данные легко представить, как разновидность текстовых и использовать соответствующий подход. Данный способ анализа данных широко обсуждается в публикациях последних лет [2, 6], в том числе и по причине относительно недавней разработки новых эффективных методов векторных представлений.

Метод one-hot encoding является базовым, наивным подходом к трансформации слов в вектора и предполагает простой подсчет количества различных слов в документе. В рамках известных задач вместо представления белка через его физические свойства, его можно описать через аминокислотную последовательность. Белок с последовательностью длиной L может быть представлен как матрица размерности L x n, где n это число встречаемых различных аминокислот. Каждая строка матрицы будет состоять из (n − 1) 0 и единственной 1, указывающей на аминокислоту занимающую данную позицию в белке.

Данный метод имеет ряд недостатков. One-hot encoding по своему определению порождает разреженные матрицы. При использовании one-hot encoding нельзя определить меру сходства между последовательностями, они либо идентичны, либо нет. Для one-hot encoding разница между словами “два”, “три” и “кот” одинакова для всех их комбинаций, хотя “два” и “три” интуитивно более схожи, чем “два” и “кот” или “три” и “кот”. Более того, one-hot encoding требует, чтобы все последовательности были выровнены. И это выравнивание необходимо будет обновлять по мере добавления последовательностей к модели. И если обновление выравнивания изменит его длину или даже будут добавлены или удалены аминокислоты, то размерность всех векторов изменится. Задача выравнивания большого числа последовательности является NP-полной и в полной мере не решена. А если выравнивание найдено неудачно, то исходные данные модели уже являются ошибочными.

Более современные способы построения векторных представлений, такие как word2vec и GloVe связаны с представлением слова через частоту его возникновения в определенных контекстах на основании обучающей выборки. Эти методы определяют вектора минимизируя дистанцию между парами слов, встречающихся в схожих контекстах [4, 5]. Предполагается что такие слова имеют некоторую смысловую связь. Данное предположение является верным и для геномных данных, схожие гены имеют схожих соседей по оперону, потому что выполняют одну и ту же функцию.

Подход word2vec был протестирован на данных генома E. Coli взятых из сервиса DOOR [1]. В рамках задачи гены рассматривались как слова, входящие в текст-оперон. В выборку вошло 27983 оперонов и 86563 различных генов. В результате использования алгоритма были получены вектора генов и использованы для создания векторов оперонов через суммирование векторов соответствующих генов. Кластеризация векторов оперонов алгоритмом dbscan правильно определяет классы оперонов в 86% случаев [3].

Использование word2vec позволяет получить эффективное числовое представление геномных данных, без введения дополнительных параметров. Алгоритм позволяет быстро оценивать новые данные при помощи заранее полученных векторных представлений и подходит для работы с большим объемом данных. В отличии от многих алгоритмов машинного обучения word2vec не требует длительной настройки – только длина контекста и искомых векторов. Вектора полученные в результате работы word2vec легко использовать как входные данные для дальнейшего применения стандартных алгоритмов машинного обучения с учителем для решения задач классификации и регрессии [6].

Литература:

1. Cao H., Ma Q., Chen X., Xu Y. DOOR: a prokaryotic operon database for genome analyses and functional inference // Brief Bioinform, 2017. — 8 p.
2. Ehsaneddin Asgari, Mohammad R. K. Mofrad. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics // PLOS One, 2013.
3. Romashko D.A., Medvedev A.U. Применение word2vec в задаче кластеризации оперонов // Программные системы и вычислительные методы, 2018. — 1-6 p.
4. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space // CoRR, abs/1301.3781, 2013.
5. Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, Jeffrey Dean. Distributed representations of words and phrases and their compositionality. // Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems, 2013. — 3111-3119 p.
6. Xu Min, Wanwen Zeng, Ning Chen, Ting Chen, Rui Jiang. Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding // Bioinformatics, Vol. 33, Issue 14, 2017. — 92-101 p.