

ТЕОРЕТИЧЕСКИЙ ОБЗОР МЕТОДОВ КОДИРОВАНИЯ ГРАФОВ

В.Е. Козловский, Е.Е. Лунева
Научный руководитель: Е.Е. Лунева
Томский политехнический университет
v1teka@tpu.ru

Введение

С помощью теории графов сегодня решаются многие прикладные задачи в различных областях: транспортная логистика; политический анализ, например, анализ настроений граждан относительно политической ситуации для прогнозирования итогов выборов; социологические задачи, в частности, задача выявления возможных друзей в социальных сетях и т.д.

Существует ряд проблем, которые теория графов позволяет решить с помощью трудозатратных алгоритмов, требующих больших вычислительных мощностей и времени. Например, задача коммивояжера или задача идентификации в социальных сетях пользователей экспертов в заданной предметной области. Одним из путей решения такого рода задач является использование машинного обучения. При использовании нейронной сети, оперирующей графами, возникает вопрос, как наилучшим образом представить входные данные.

Цель данной работы – теоретический обзор методов кодирования и представления графов для использования их в качестве входных данных в системах с машинным обучением.

Постановка задачи

Традиционными способами представления графов, являются матрица смежности, список смежности и список ребер. Размер каждой из этих структур напрямую зависит от количества вершин, в то время, как нейронные сети имеют фиксированный размер входного слоя и оперируют нормализованными данными. Следовательно, если входными данными проектируемой нейросети является, например, матрица смежности, она должна будет оперировать лишь графами, количество вершин в которых совпадает.

Если же речь идет о структурах с миллионами участников, описание графа будет иметь несопоставимый размер. Использование нейронных сетей для работы с большими объемами данных подразумевает распараллеливание вычислений. В рамках различных задач рассматриваются различные признаки вершин графов: принадлежит ли вершина той или иной группе, выполняет ли роль связующего звена, и каковы её значения показателей центральности. В силу изоморфизма традиционных структур представления, ни одна из них не позволит выявить эти признаки: при рассмотрении по частям (например, построчная передача матрицы), не отражается информация о положении отдельной вершины в контексте всей структуры.

В качестве примера для рассмотрения алгоритмов будем использовать неориентированный граф “Zachery’s Karate Club” и его матрицу смежности

[1]. Необходимо на основе этих входных данных получить масштабируемое и наиболее полное представление графа, которое может быть передано на вход нейронной сети.

Эмбединг вершин графа

В основу методов работы с графами в нейронных сетях положен принцип эмбединга вершин, который состоит в описании каждой из них с помощью набора численных значений определенного размера [3]. То есть, предлагается сопоставить каждой вершине вектор некоего пространства размерностью, намного меньшей, чем их количество в графе. Такое представление не содержит прямого описания ребер в графе – их наличие выражается только близостью расположения вершин.

Для модификации кодирующего алгоритма, также используется функция декодирования – например, функция восстановления списка ребер и вершин по полученному представлению графа. Последовательно применяя кодирование и декодирование, с помощью минимизации несоответствий между полученными данными и реальной структурой графа, можно оптимизировать алгоритм генерирования представлений. Чаще всего для этого используется метод стохастического градиентного спуска.

Для представления графов в приемлемом для нейросетей виде в предшествующих работах предлагаются следующие методы: DeepWalk (word2vec [2]), Node2Vec [3], 2D CNN [4], Graph Convolutional Networks [5]. Формально их можно разделить на две группы: основанные на принципе случайных блужданий и использующие для кодирования нейронные сети.

Случайные блуждания

Выбрав одну из вершин графа в качестве начальной точки, произвольным образом перемещаемся в одну из ее соседних вершин. Затем повторяем случайное перемещение из вновь выбранной вершины и т.д. Получившаяся последовательность вершин называется реализацией случайного блуждания по графу [6].

DeepWalk

В рамках данного алгоритма для каждой из вершин графа генерируется заданное число случайных блужданий заданной длины. Если затем каждую вершину интерпретировать как слово, то сгенерированные блуждания можно рассматривать в качестве предложений. Таким образом, их можно использовать для обучения механизма обработки естественного языка “word2vec” [2]. Эта технология позволяет составить векторное представление на основе контекстной близости: по аналогии со

словами в текстах, встречающиеся в блуждании рядом вершины будут иметь близкие координаты (рис. 1).

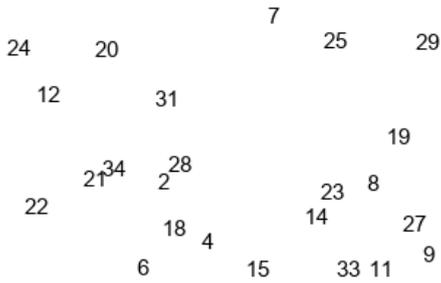


Рис. 1. Представление графа, полученное с помощью метода DeepWalk

Node2vec

Основное отличие данного метода от DeepWalk заключается в том, что на каждой итерации блуждания вычисляются вероятности перехода в ту или иную вершину. В зависимости от значений двух параметров p (параметр возврата) и q (параметр углубления), блуждания более склонны к приближению или отдалению относительно начальной вершины [3]. Корректирование этих параметров позволяет сформировать блуждания, наиболее полно отражающие те или иные признаки структуры каждого отдельного графа.

2D CNN

Метод предлагает преобразование графа в изображение, которое затем может быть пропущено через сверточную нейронную сеть [4]. Весь процесс может быть описан тремя шагами:

- получение эмбединга графа с помощью одного из представленных способов;
- сжатие векторного пространства;
- фрагментирование сжатого пространства и создание 2D-гистограммы каждого фрагмента;
- решение задачи теории графов путем выявления зависимостей по каждой отдельной гистограмме.

Graph Convolutional Networks

В рамках этого метода производится программное моделирование сверточной нейронной сети, которая принимает на вход матрицу смежности графа и векторные представления вершин (на начальном этапе подобранные случайно). На каждом скрытом слое представления изменяются с использованием выбранной функции активации и передаются следующему слою [5]. Корректировка весов нейронов позволяет получать более полные векторные представления графа:

1. случайным образом генерируется матрица X векторных представлений для каждой из вершин;
2. для каждой вершины создаются ребра – петли:

$$A = A + I$$

, где I – единичная матрица;

3. матрицы A и X перемножаются, значения нормализуются с помощью степеней вершин:

$$H = D^{-1} * A * X$$

4. генерируются несколько слоев нейронной сети со случайными весами;
5. на выходе имеем откорректированную матрицу векторных представлений вершин.

Алгоритм был реализован, представление, полученное для исходного графа визуализировано (рис. 2).

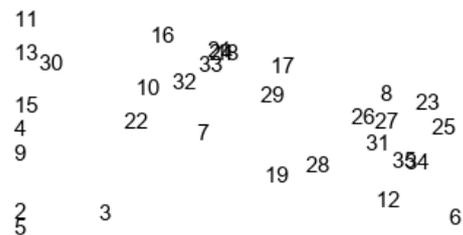


Рис. 2. Представление графа, полученное с помощью метода сверточной сети

Заключение

В рамках данной работы рассмотрены методы кодирования графов, основанные на принципе случайных блужданий и с использованием нейронных сетей. Реализованы алгоритмы обоих типов, результаты визуализированы.

Оптимизация алгоритмов с помощью обучающих данных позволит получить представления графов, отражающие их структуру наиболее полно.

Работа выполнена при финансовой поддержке РФФИ (проект №17-07-00034 А).

Список используемых источников

1. Zachary's karate club [Электронный ресурс]/Wikipedia. – URL: https://en.wikipedia.org/wiki/Zachary%27s_karate_club (дата обращения 15.11.2018).
2. Word2vec [Электронный ресурс]/Wikipedia. – URL: <https://ru.wikipedia.org/wiki/Word2vec> (дата обращения 15.11.2018).
3. Aditya Grover, Jure Leskovec. node2vec, Scalable feature learning for networks. Phoenix, Arizona: International Conference on Knowledge Discovery and Data Mining, 2016. – 4 p.
4. Tixier, Antoine Jean-Pierre, Giannis Nikolentzos, Polykarpos Meladianos and Michalis Vazirgiannis. Graph Classification with 2D Convolutional Neural Networks. Melbourne, Australia: ACL, 2017. – 4 p.
5. Thomas N. Kipf, Max Welling. Semi-supervised classification with graph convolutional networks. Toulon, Palais des Congrès Neptune: International Conference on Learning Representations, 2017. – P. 3-6.
6. А. Б. Теслюк. О случайном блуждании по графу веб-документов. – М.: Московский физико-технический институт, 2004. – С. 4-5.