Секция «Вычислительная математика и кибернетика»

Предсказание связности графа

Кириллов Александр Николаевич
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: arhipisk@gmail.com

Задача предсказания связности графа (Link Prediction Problem) — относительно новая проблема для анализа данных. Она быстро набирает популярность в связи с активным развитием и исследованием социальных сетей. На данный момент уже разработан ряд методов, которые активно используются при её решении.

В общем виде задача ставится как предсказание появления новых ребер в динамически изменяющемся графе, более строго — дан граф G = в дискретные моменты времени $t = t_0, t_1, \ldots, t_n$, требуется предсказать, каким будет граф в момент времени t_{n+1}. Такой динамический граф легко интерпретируется как модель социальной сети, где множество вершин это пользователи, а множество ребер это дружеские связи между ними. Основной проблемой задачи является большой размер данных, что требует разработки эффективных с точки зрения производительности методов.

Целью работы являлся поиск новых методов, которые позволяют улучшить качество решения. Исследования проводились на реальных данных социальной сети flickr, предоставленных для проведения международного конкурса «IJCNN Social Network Challenge» компании KAGGLE [3]. В исследуемом графе более 1-го миллиона вершин и около 7-ми миллионов ребер. Целью конкурса было упорядочить набор из пар вершин по вероятности возникновения ребра между ними в ближайшем будущем. Функционалом качества решения являлся AUC (площадь под ROC-кривой).

Для построения признакового пространства использовались классические признаки, такие как коэффициент Жаккара, коэффициент Adamic/Adar, Katz, PageRank и другие методы, описанные в [2]. Результирующий вектор – линейная комбинация этих признаков, настроенная с помощью алгоритма «LENKOR» [1]. Два различных подхода позволяют расширить признаковое пространство и существенно улучшить результат. Первый подход является обобщением идеи использования количества общих соседей: вершины x и y соединены, если между множеством вершин смежных с x и множеством вершин смежных с y большое количество ребер. Второй подход эксплуатирует гипотезу о том, что вершины соединены ребром, если уже соединены похожие на них. Для данного подхода была использована техника коллаборативной фильтрации.

Нам удалось выделить ряд новых признаков, которые в комбинации с классическими методами значительно улучшают качество решения. Топология динамического графа позволяет с большой точностью предсказывать появление новых связей. Итоговый результат – AUC = 0.94.

Литература

1. D'yakonov A. Two Recommendation Algorithms Based on Deformed Linear Combinations // ECML-PKDD 2011 Discovery Challenge Workshop. 2011. Pp. 21-27.

2. Liben-Nowell D., Kleinberg J. The Link Prediction Problem for Social Networks // Society for Information Science and Technology. 2007. Vol. 58. No. 7. Pp. 1019-1031.

3. Kaggle: http://www.kaggle.com/c/socialNetwork

Слова благодарности

Работа выполнена при финансовой поддержке РФФИ, проект 12-07-00187-а.