УДК: 004.942
Название: Разработка алгоритма восстановления значений атрибутов в комплексных сетях

Авторы:
Калинин А.М, Университет ИТМО, г. Санкт-Петербург;
Боченина К.О., Университет ИТМО, г. Санкт-Петербург

Научный руководитель: Боченина К.О., Университет ИТМО, г. Санкт-Петербург

Тезис доклада:

Для исследования систем связанных объектов реального мира (транспортные, компьютерные, социальные и т.п.) применяется формализм комплексных сетей. Вершины комплексной сети могут дополнительно характеризоваться набором атрибутов (например, пол, возраст, вектор интересов для пользователей социальной сети). В сетях реального мира часть атрибутов вершин сети может быть недоступна вследствие специфики метода сбора данных или настроек приватности информации. В этом случае возникает задача восстановления значений атрибутов узлов по связям между вершинами.

Целью исследования является разработка метода восстановления пропущенных атрибутов вершин комплексных сетей, идентифицируемых по разнородным данным.

Методы восстановления пропущенных значений в графе можно разделить на три класса: локальные модели, статистические реляционные модели и эмбеддинги.

Локальные модели рассматривают только локальные свойства вершины графа без учёта его связей. *Статистические реляционные модели* дополнительно учитывают отношения вершин друг с другом. Как и локальные, статистические реляционнные модели могут быть реализованы на основе машинного обучения (во втором случае включаются признаки с информацией о связях между вершинами). Примерами являются: Weighted-Vote Relational Neighbor Classifier (wvRN), Class-Distribution Relational Neighbor Classifier (cdRN), Network-Only Link-Based Classification (nLB).

Статистическое реляционное обучение предполагает итеративное применение локальных и реляционных моделей, при котором для вершин с неизвестными метками на каждом шаге обновляются их вероятности принадлежности к классам на основе состояний смежных вершин. Способы обхода графа и то, как вершины используют оценки смежных соседей, определяется т.н. методами коллективного вывода, среди которых наиболее известны: Iterative Classification, Gibbs Sampling, Relaxation Labeling.

Эмбеддинги проецируют объекты пространства высокой размерности в пространство низкой размерности. Широкое применение эмбеддинги получили в обработке текстов, наиболее известна реализация word2vec. Графовые эмбеддинги основаны на том же подходе, что и текстовые, но формируются не с помощью контекстов слов, а путём случайных блужданий от каждой вершины графа. Разработаны различные варианты графовых эмбеддингов, такие как: node2vec, LINE, DeepWalk, ARCTE.

Предложенный в данной работе подход комбинирует графовые эмбеддинги и реляционные модели. Вероятности, полученные из обученного на эмбеддингах классификатора, передаются на вход реляционной модели, которая использует их в качестве априорных.

Для тестирования проанализированных методов извлечён набор данных, содержащий информацию о ~435 тыс. пользователей социальной сети «ВКонтакте» и ~4 млн. связей между ними. Для восстановления выбраны следующие социо-демографические атрибуты: пол, род деятельности и место учёбы. Максимальное значение доли правильных ответов достигается при большом (от 70%) количестве известных данных и равняется ~80%. Комбинированный метод превосходит реляционные модели до 25% при малом (менее 50%) количестве известных вершин, что позволяет сделать вывод о целесообразности его применения.

Результатом работы является комбинированный алгоритм, объединяющий статистические реляционные модели и графовые эмбеддинги. Разработанный метод может быть применён на любых данных, организованных в виде комплексной сети.

Автор _____ /_____Калинин А.М._____/

Научный руководитель _____ /_____Боченина К.О._____/