УДК 519.173.5

Исследование эффективности графовых метрик при решении задач классификации на стохастических сетях

В.С. Ивашкин

Московский физико-технический институт (государственный университет)

Случайные кластерные сети широко применяются в машинном обучении. Одна из центральных задач машинного обучения для кластерных сетей – классификация. Классификация позволяет отнести каждую вершину к определенному классу, если уже размечена некоторая выборка (обучение с учителем). Классификаторы используют графовые метрики для определения наиболее подходящего класса для каждой вершины. Целью данной работы является сравнение эффективности наиболее известных графовых метрик применительно к задаче классификации.

Для исследования использовались генератор больших кластерных сетей[1], метрики, а также kNN классификатор с параметром α, задающим разницу в весе между самым близким и более дальними ближайшими соседями. Сравнивая результат классификации и заранее известные данные о кластерах графа, можно определить эффективность графовой метрики.

Метрики образуют параметрические семейства. В работе использовались следующие семейства метрик:

1. Комбинации метрик Shortest Path и Commute Time – тривиальное семейство, служащее базой для сравнения;
2. “Plain” Walk distances [2];
3. Walk distances [3];
4. [“Plain”] Forest distances [4];
5. Logarithmic Forest distances [3];
6. Communicability distances [3];
7. Logarithmic Communicability distances [3];
8. Helmholtz free energy distances [5].

Было проведено два эксперимента.

В первом эксперименте были построены графики зависимости качества классификации от параметра семейства метрики при различных параметрах генерируемого графа. Метрики Logarithmic Communicability и Helmholtz free energy побеждают при всех исследуемых параметрах графов.

Идея второго эксперимента такова: вначале генерируем два набора графов с одинаковыми параметрами: обучающий набор графов и набор для соревнований. Для каждого семейства

метрик находим оптимальные для первого набора графов параметры классификатора k и α, а также параметр семейства метрики. После этого проводятся «соревнования» метрик с найденными оптимальными параметрами на втором наборе графов. Для каждого графа метрика получает столько баллов, сколько метрик она «победила». Таким образом, можно не только определить лучшую метрику, но и увидеть, насколько хорошо можно предсказать качество классификации каждой метрикой.

Результаты показывают, что монотонной зависимости между предсказываемым качеством и количеством полученных баллов в соревнованиях не прослеживается. Особенно сильная разница проявляется для метрики Communicability – будучи одной из лучших в первом эксперименте, эта метрика занимает одно из последних мест в проведенных соревнованиях. Это связано с тем, что на результат соревнований влияет также «стабильность» результата в зависимости от графа. Результаты, полученные в этой части исследования, подтверждают, что метрики Logarithmic Communicability, Walk, “Plain” Walk, Free Energy лучше всего подходят для классификации на исследуемых графах.

Следует отметить, что показавшая себя хорошо метрика Logarithmic Communicability ранее практически не исследовалась, поэтому представляется перспективным продолжить ее исследование как теоретически, так и на экспериментах.

Литература

1. *Kluge R.* An Efficient Generator for Large Clustered Dynamic Random Networks: Bachelor Thesis. – Karlsruhe Institute of Technology. – 2011.
2. *Чеботарев П.Ю., Шамис Е.В.* О мерах близости вершин графов. – Автоматика и Телемеханика, 1998. – № 10. – С. 113–133.
3. *Chebotarev P.* Studying new classes of graph metrics. – Geometric Science of Information. – Springer, 2013. – Pp. 207–214.
4. *Chebotarev P., Shamis E.* The forest metrics for graph vertices. – Electronic Notes in Discrete Mathematics. – 2002. – Vol. 11. — Pp. 98–107.
5. *Kivimäki I., Shimbo M., Saerens M.* Developments in the theory of randomized shortest paths with a comparison of graph node distances. – Physica A: Statistical Mechanics and its Applications. – 2014. – Vol. 393. – Pp. 600–616.