Использование графов для классификации

Черноскутов Михаил,
ИММ УрО РАН, УрФУ,
mach@imm.uran.ru

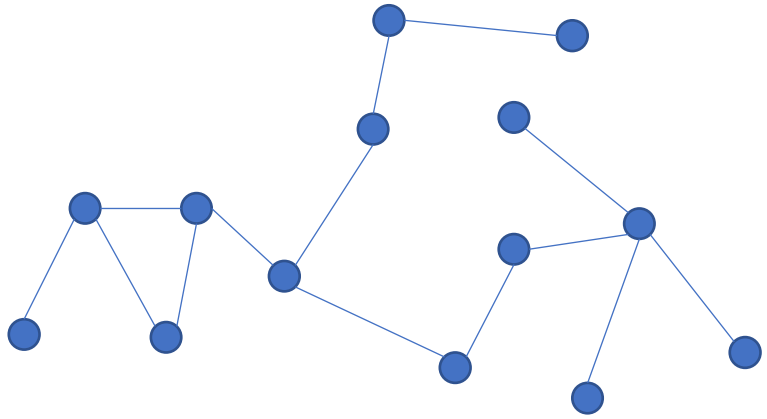
Введение

- Общее описание
  - Алгоритм для классификации, основанный на частоте встречаемости элементов того или иного класса в сообществах, выделенных в графе соседей
- Особенности
  - Алгоритм может использоваться как для бинарной, так и для многоклассовой классификации
  - Могут использоваться разные алгоритмы поиска сообществ для повышения эффективности классификации

Алгоритм

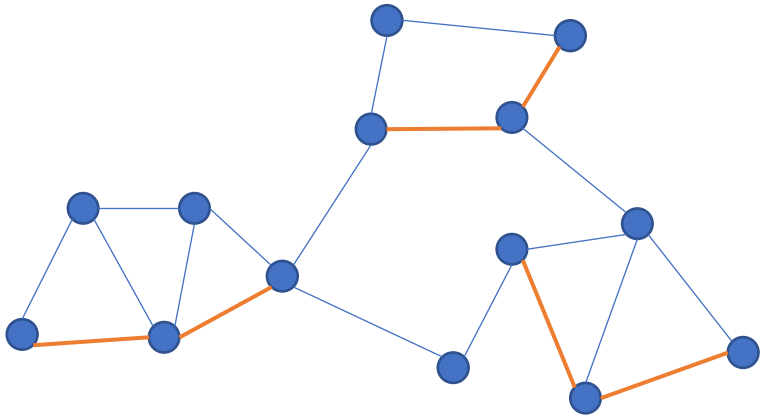
- Шаг 1
  - Построение графа n ближайших соседей
- Особенности
  - Вес рёбер – расстояние между элементами в степени -1 (чем меньше расстояние, тем больше вес)
  - Граф строится на элементах как тренировочной, так и тестовой выборок

	y_1	…	y_j
x_1	x_{11}		x_{1j}
…		…	
x_i	x_{i1}		x_{ij}



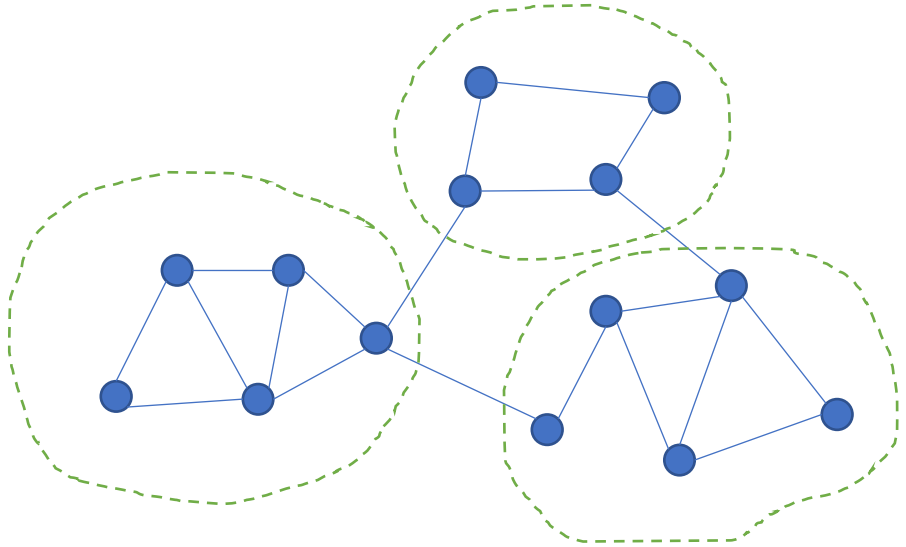
Алгоритм

- Шаг 2
  - Связывание узлов, находящихся на расстоянии двух рёбер
  - $\alpha(w_{ij} + w_{jk}) > \frac{1}{r_{ik}}$
    - α – коэффициент ($0 < \alpha < 1$), w – веса рёбер, r – расстояние между элементами данных
- Особенности
  - Повышение локальной плотности графа
  - Коэффициент α для регулирования количества новых рёбер



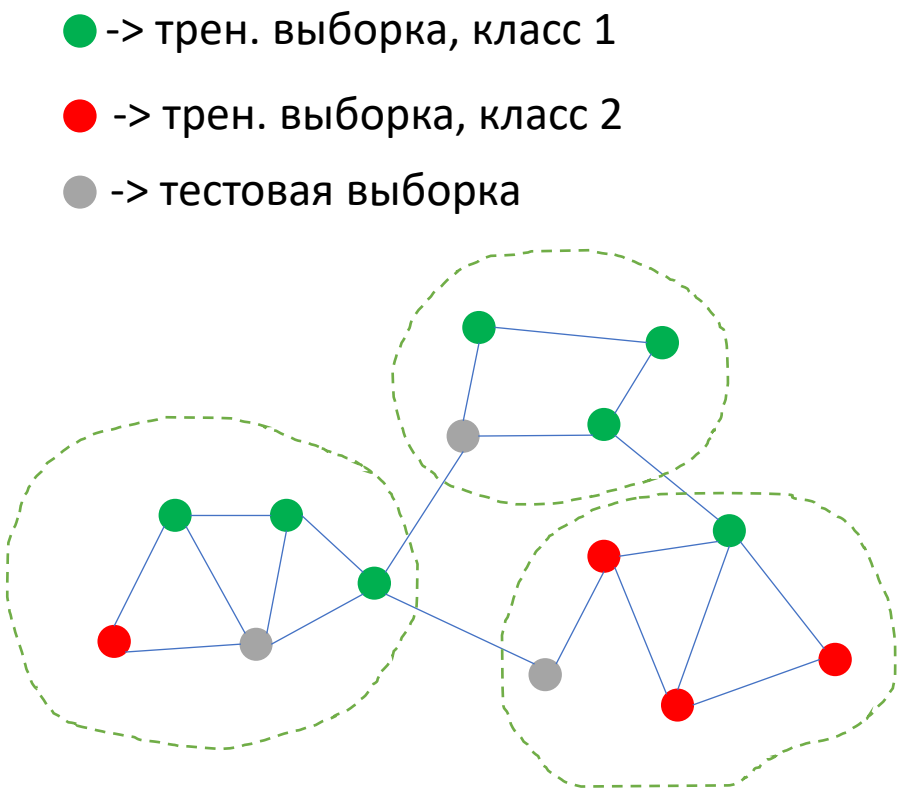
Алгоритм

- Шаг 3
  - Поиск сообществ в полученном графе
- Особенности
  - Граф разобьётся на сообщества, количество которых может не совпадать с количеством размеченных классов в тренировочной выборке



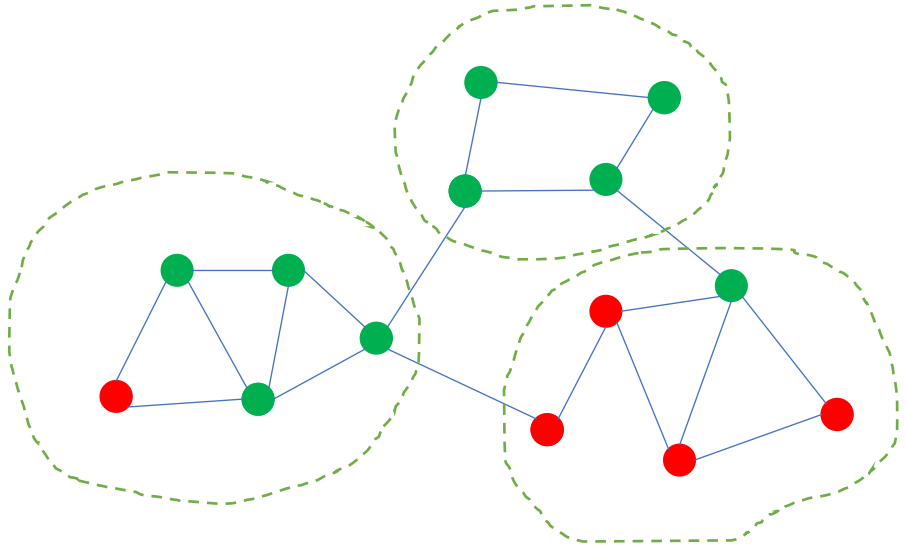
Алгоритм

- Шаг 4
  - Разметка узлов по принадлежности к тренировочной или тестовой выборке
- Особенности
  - В одном сообществе могут присутствовать элементы из разных классов



Алгоритм

- Шаг 5
  - Классификация элементов на основе частоты встречаемости узлов того или иного класса в каждом из сообществ



Тестирование

- Наборы данных для классификации
  - Breast cancer
    - Бинарная классификация
    - 569 элементов в выборке
    - 30 признаков у каждого элемента
  - Handwritten digits
    - 10 классов
    - 1797 элементов в выборке
    - 64 признака у каждого элемента
- Метрика
  - Доля верно классифицированных элементов тестового множества
- Соотношение размеров тренировочной и тестовой выборок
  - 80% / 20%
- Алгоритм для поиска сообществ
  - Asynchronous Label Propagation
  - U.N. Raghavan, R. Albert, and S. Kumara, \Near linear time algorithm to detect community structures in large-scale networks," Phys. Rev. E, vol. 76, p. 036106, Sep 2007. [Online]. doi:10.1103/PhysRevE.76.036106

Результаты (предложенный метод)

- Breast cancer
- Обозначения
  - n – количество ближайших соседей
  - α – коэффициент связываемости
  - N_e - количество рёбер в графе
  - N_c - количество сообществ в графе
  - $score$ – усреднённое по пяти запускам значение метрики эффективности (доля верно классифицированных элементов)

	n	α	N_e	N_c	$score$
1	2	1.0	1729	84	**0.924**
2	2	0.8	1729	86	0.901
3	2	0.6	1726	85	0.899
4	2	0.4	1644	86	0.915
5	2	0.2	808	140	0.880
6	3	1.0	2841	52	0.917
7	3	0.8	2841	58	**0.947**
8	3	0.6	2840	55	0.936
9	3	0.4	2746	54	0.931
10	3	0.2	1201	102	0.925

Результаты (предложенный метод)

- Handwritten digits
- Обозначения
  - n – количество ближайших соседей
  - α – коэффициент связываемости
  - N_e - количество рёбер в графе
  - N_c - количество сообществ в графе
  - $score$ – усреднённое по пяти запускам значение метрики эффективности (доля верно классифицированных элементов)

	n	α	N_e	N_c	$score$
1	2	1.0	7454	158	**0.988**
2	2	0.8	7454	165	0.977
3	2	0.6	7450	157	0.985
4	2	0.4	5984	186	0.980
5	2	0.2	2671	385	0.965
6	3	1.0	13164	79	0.984
7	3	0.8	13164	85	0.981
8	3	0.6	13157	76	**0.989**
9	3	0.4	10573	109	0.986
10	3	0.2	3871	294	0.981

Результаты (другие методы)

- Другие методы
  - Алгоритм случайного леса (Random Forest)
  - Градиентный бустинг (XGBoost)
- *score* – усреднённое по пяти запускам значение метрики эффективности (доля верно классифицированных элементов)

	score (breast cancer)	*score* (handwritten digits)
random forest	0.961	0.967
XGBoost	0.954	0.958

Обсуждение

- Проблемы
  - Долгое построение графа ближайших соседей
    - Эвристики, пространственные хэш-функции
  - Пересекающиеся/непересекающиеся сообщества?
    - Отделение узлов, участвующих в нескольких сообществах
- Преимущества
  - Интерпретируемость результатов
  - Малое количество параметров
  - (Потенциальное) отсутствие требований к размеру обучающей выборки

Вопросы?