ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 85 за 2021 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

А.А. Кислицын, Ю.Н. Орлов

Исследование статистик графов ближайших соседей

Рекомендуемая форма библиографической ссылки: Кислицын А.А., Орлов Ю.Н. Исследование статистик графов ближайших соседей // Препринты ИПМ им. М.В.Келдыша. 2021. № 85. 23 с. https://doi.org/10.20948/prepr-2021-85
https://library.keldysh.ru/preprint.asp?id=2021-85

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.А. Кислицын, Ю.Н. Орлов

Исследование статистик графов ближайших соседей

Москва — 2021

Кислицын А.А., Орлов Ю.Н.

Исследование статистик графов ближайших соседей

В работе рассмотрена структура графов ближайших соседей. Описан статистический подход к анализу графов k-NN, цель которого – поставить в соответствие данному конкретному графу вероятность его реализации в предположении определенного распределения расстояний между вершинами и в зависимости от числа вершин. В качестве основных статистик используются выборочные распределения графа по числу несвязных фрагментов, фрагментов по числу вершин и вершин по степеням входящих ребер. В качестве примера рассмотрена задача идентификации автора текста методом n-грамм.

Ключевые слова: граф ближайших соседей, связность, статистика степеней вершин, выборочный граф, распределение расстояний

Kislitsyn A.A., Orlov Yu.N.

The investigation of the k-NN graph statistics

In this paper the investigation of the structure of k-NN graph is presented. The purpose of this analysis is to match the probability of implementation particular graph under assumption of a certain distribution of distances between vertices and depending on the number of vertices. Sample distributions of the graph by the number of disconnected fragments, fragments by the number of vertices, and vertices by the degrees of incoming edges are used as the main statistics. As an example, the problem of identifying the author of a text by the n-gram method is considered.

Keywords: nearest neighbors graph, connectivity, power vertex distribution, sample graph, distance distribution

Работа выполнена при поддержке гранта РНФ № 19-71-30004

Содержание

1. Постановка задачи о статистике графов 3
2. Практический пример: близость между авторами текстов 6
3. Свойства графов первых ближайших соседей 15
4. Заключение 22
Литература 22

1. Постановка задачи о статистике графов

Анализ графов так называемых k ближайших соседей (k nearest neighbors, k-NN – см., напр., [1]) представляет собой один из широко используемых методов кластеризации данных, для которых тем или иным способом введена функция расстояния между ними. В большинстве случаев целью анализа является установление структурных связей между объектами, которым были поставлены в соответствие наборы чисел и введена мера близости. То есть в задаче кластеризации данных сами числа (в многомерном случае – векторы), как и расстояния между ними, интереса не представляют, важным является результат группировки чисел, позволяющий выявить свойства взаимного расположения собственно объектов. Кластеризация – это эффективный способ сокращения описания, когда вместо большого количества исходных объектов оказывается возможным рассмотреть относительно небольшое число классов и сформировать типичного представителя класса (так называемый паттерн или эталон). Поскольку результат группировки критически зависит от меры близости, то нахождение меры, приводящей к осмысленным выводам применительно к исходным данным, представляет высокий практический интерес и трактуется чуть ли не как «закон природы». В то же время на практике анализируется, как правило, один конкретный набор объектов, относительно которых и были собраны данные. Объектами могут быть, например, учащиеся определенных учебных заведений, группы потребителей того или иного товара, участники социальной сети, клиенты банка и т.п. Тогда практическую важность приобретает вопрос о вариабельности установленных эмпирических закономерностей, поскольку различные множества однотипных объектов (например, учащиеся одного и того же класса, но из разных школ) будут порождать, вообще говоря, разные графы связей. В этом смысле граф связей является случайным, поскольку случаен сам отбор объектов в анализируемую группу.

Теория случайных графов развита очень подробно в направлении «теория вероятностей». Начиная с классической работы [2] свойства графов, вершины которых соединяются между собой по заданному вероятностному закону, активно изучались в контексте решения многих прикладных задач. Развитая теория подробно изложена в книгах [3-5]. Однако в направлении «математическая статистика» конструктивных моделей и результатов в теории графов существенно меньше. Это связано с тем, что «выборка графов» сама по себе не имеет места, ведь граф просто визуализирует некоторое свойство изучаемой группы объектов. Так что понятие выборки адресуется именно группе объектов, которые должны быть однородными по изучаемому свойству в статистическом смысле. Но поскольку изучаемое свойство выявляется как раз в результате анализа структуры графа, заранее сформировать группы объектов с нужным свойством невозможно. Тогда, рассмотрев распределение параметров, которые в той или иной модели системы считаются ключевыми

для конкретной выборки, можно размножить выборки с найденным эмпирическим распределением параметров, после чего изучить статистику графов, отвечающих (по предположению модели) различным состояниям изучаемой системы. Совокупность графов, отвечающих заданному распределению параметров, будем называть статистическим семейством графов. Поскольку же выборка параметров является конечной, то соответствующая выборочная функция распределения флуктуирует от выборки к выборке. Тогда следует изучить вариабельность структуры семейства графов при изменении распределения параметров системы.

Отметим, что большинство результатов исследований по случайным графам применительно к анализу практических ситуаций, таких как модели социальных сетей, транспортных сетей и т.п. (см. работы [6-9]), относится к «асимптотически глобальным» утверждениям. Например, доказывается, что при стремлении числа вершин к бесконечности исследуемый граф почти наверное будет связным или что в нем почти наверное будет существовать полносвязный подграф размерности не ниже определенной. Однако в ряде задач требуется более детальный анализ структуры графа. В частности, применительно к задаче кластеризации определение числа несвязных между собой подграфов является более важным, чем выяснение вопроса о том, есть ли они вообще. Основная проблема состоит в том, что анализ одного конкретного графа не позволяет оценить вероятность его реализации как элемента из определенного множества случайных графов, поскольку нужное множество не формализовано. Для этого необходимо уметь размножать графы с определенным свойством связности его вершин, а каковы эти свойства, на практике как раз и не известно. Редким исключением является граф ближайших соседей, который строится на основе матрицы попарных расстояний между точками. В этом случае функция распределения расстояний известна, что позволяет выяснить, какие графы ей отвечают.

В настоящей работе мы рассмотрим вопрос о вариабельности структуры графа ближайших соседей для множества многомерных векторов, парное расстояние между которыми измеряется в норме L1 как сумма модулей разностей соответствующих координат (хотя выбор нормы в данном случае не принципиален). Также на конкретной задаче – а именно, идентификации автора литературного текста – будет оценена эффективность эвристики ближайшего соседа как инструмента машинного обучения распознавания автора.

Сформулируем основные задачи статистического анализа k-NN графов, которые представляют интерес для классификации и идентификации элементов из некоторого конечного множества.

Рассматриваются расстояния между точками (векторами) в некотором подпространстве из R^d. Пусть в нем выбраны каким-либо способом N различных точек, которые занумерованы (способ нумерации не играет роли). Такая выборка из N точек порождает симметричную матрицу $\left(l_{ij}\right)_{N\times N}$ попарных расстояний. В каждом i-м столбце определяется наименьший элемент (кроме нуля на главной диагонали). Номер j соответствующей строки

отвечает точке, ближайшей к точке i. Вершины i и j соединяются направленным ребром (стрелкой) от i к j. Так образуется граф $\Gamma_1(N,d)$ первых ближайших соседей в наборе из N точек в R^d. Аналогично строится граф $\Gamma_2(N,d)$ вторых соседей, т.е. точек, следующих за ближайшими, и т.д.

Первый вопрос: из скольких несвязных фрагментов состоит типичный граф k-NN? Для этого требуется определить вероятность реализации графа с N вершинами в виде совокупности Q фрагментов. В случае ближайших соседей число Q может меняться от 1 (граф связный) до $[N/2]$ (граф состоит из пар взаимно ближайших вершин).

Второй вопрос: поскольку распределение графа по степеням вершин зависит от числа самих этих вершин, то каково типичное распределение фрагмента по числу V вершин в зависимости от полного числа вершин N графа и числа фрагментов Q?

Третий вопрос: каково распределение вершин фрагмента по степеням при данных значениях N, Q и V?

Четвертый вопрос относится к адекватности модели ближайших соседей применительно к задаче кластеризации: какова вероятность того, что объединение графов k ближайших соседей образует связный граф?

Перечисленные вопросы могут быть переформулированы в терминах классических комбинаторных задач, не все из которых решены в настоящее время. В настоящей работе будут доказаны некоторые новые утверждения, касающиеся структур графов ближайших соседей.

Отметим основные свойства графов k-NN.

- С каждой вершиной связано ровно одно выходящее ребро.
- Число вершин графа равно числу ребер.
- Каждый связный граф k-NN содержит ровно один цикл.

Максимальная степень вершины для одномерного расположения точек равна, очевидно, двум.

В двумерном случае максимальное число точек, находящихся на минимальном (допустим, единичном) расстоянии от данной точки, находится на окружности единичного радиуса, причем расстояние между этими точками также равно единице. Следовательно, эти точки образуют вершины равносторонних треугольников с одной общей вершиной, лежащей в центре. Таких треугольников может быть не более шести, но это критический уровень: расстояния между всеми семью точками (включая центральную) равны. Чтобы только одна центральная точка была равноудалена от остальных, точки на окружности надо несколько раздвинуть, для чего одну из них придется удалить. Следовательно, ближайших соседей выделенной точки на плоскости может быть не более пяти.

Для пространства произвольной размерности максимальное число ближайших соседей неизвестно. Здесь мы не будем обсуждать связь этой задачи с диаграммами Вороного и с максимальным числом касаний единичной сферы (т.н. kissing number). Отметим только, что эта задача возникает в

различных приложениях дискретной математики. Возможно, что численные эксперименты в этом направлении могли бы привести к появлению практически полезных эвристик.

Пусть задана функция распределения $F(l)$ попарных расстояний l, из которой генерируются элементы матрицы $\left(l_{ij}\right)_{N\times N}$. Для каждой s-й выборки рассматривается граф k-х соседей. Обозначим его $\Gamma_k^{(s)}(N)$. Цель работы состоит в изучении статистик, характеризующих эти графы. А именно, по совокупности S выборок можно определить, какие графические структуры наиболее вероятны. Статистиками являются параметры, которые в той или иной степени описывают исходные множества.

При заданном полном числе вершин N нас будет интересовать распределение $G_{k,N}(Q)$ графов k-х соседей по количеству Q несвязных фрагментов, распределение $U_{k,N}(Q,V)$ этих фрагментов по числу вершин V и распределение $H_{k,N}(Q,V,P)$ вершин графа данного семейства по степеням P. Вполне естественно, что независимые выборки расстояний из одного и того же распределения будут порождать, вообще говоря, неизоморфные графы. Тогда следует выяснить, какие структуры наиболее вероятны. В результате по виду распределений $G_{k,N}$, $U_{k,N}$ и $H_{k,N}$ можно судить о том, имеем ли мы дело в каждом конкретном случае с типовой ситуацией (в области среднего значения или моды этих распределений), или с редким отклонением, которое, возможно, вызвано специальной причиной, не учитывавшейся при формировании модели сходства объектов. Представляют также интерес предельные распределения, построенные для количества выборок, стремящегося к бесконечности.

2. Практический пример: близость между авторами текстов

Одной из популярных задач, для которой используются методы машинного обучения, является идентификация автора текста в библиотеке, содержащей образцы произведений различных авторов. Идея распознавания основана на той или иной модели близости. В частности, в [10] была предложена модель распознавания на основе анализа распределений буквосочетаний (n-грамм).

Концепция распознавания автора текста, на примере которой мы рассмотрим применение метода ближайшего соседа, формулируется в терминах эмпирических частот буквосочетаний и состоит в том, что предположительно тексты одного и того же автора близки в определенной норме к его эталону и находятся дальше от эталонов других авторов. То есть предполагается, что каждому профессиональному писателю соответствует его персональный эталон, своего рода идеальная генеральная совокупность. Тогда отдельные тексты каждого автора представляют собой выборку из соответствующей совокупности, мало отличаясь от нее и значительно более отличаясь от эталонов других авторов. Заметим, однако, что данное предположение, строго

говоря, не вполне обосновано. Дело в том, что совокупность текстов писателей фактически образует статистически однородное множество по критерию согласия Колмогорова-Смирнова, и потому с помощью этого инструмента тексты не могут быть корректно разделены по авторам. Поэтому исследование свойств других метрик, используемых для распознавания автора текста, является актуальной задачей. В частности, средневзвешенное распределение частот *n*-грамм по произведениям одного и того же автора оказывается весьма хорошим эталоном для идентификации автора по близости текста к эталону в норме L1. Расстоянием между текстами будем называть расстояние между соответствующими распределениями частот.

Пусть $D_a^i(j)$ есть эмпирически определяемое распределение частот символа j в i-м тексте автора a, и N_a^i есть число символов в данном i-м тексте. Пробелы, знаки препинания, числа и символы других алфавитов игнорируются. Прописные и строчные буквы не различаются. Пусть n_a есть число произведений автора a. Тогда эмпирической оценкой эталона $\Psi_a(j)$ автора a является взвешенное распределение частот по совокупности всех текстов, достоверно принадлежащих данному автору:

$$\Psi_a(j) = \frac{1}{N_a}\sum_{i=1}^{n_a} N_a^i D_a^i(j), \quad N_a = \sum_{i=1}^{n_a} N_a^i . \tag{2.1}$$

Тогда расстояние между i-м текстом автора a и эталоном $\Psi_a'(j)$ этого же автора определяется как

$$x_{aa}^i = \sum_{j=1}^{J}\left|D_a^i(j) - \Psi_a'(j)\right|, \tag{2.2}$$

где штрих означает, что распределение $D_a^i(j)$ исключено из эталона (2.1).

Тогда

$$x_{aa}^i = \sum_{j=1}^{J}\left|D_a^i(j) - \frac{N_a\Psi_a(j) - N_a^i D_a^i(j)}{N_a - N_a^i}\right| = \frac{1}{1 - N_a^i / N_a}\sum_{j=1}^{J}\left|D_a^i(j) - \Psi_a(j)\right|. \tag{2.3}$$

Формула (2.3) позволяет протестировать данную модель идентификации автора методом кросс-валидации, последовательно исключая одно из произведений и определяя его автора на основе остальных данных.

Расстояние между i-м текстом автора a и эталоном $\Psi_b(j)$ некоторого другого автора $b \neq a$ есть

$$y_{ab}^i = \sum_{j=1}^{J}\left|D_a^i(j) - \Psi_b(j)\right|. \tag{2.4}$$

Формально для безошибочного распознавания автора текста требуется, чтобы $\forall i,a,b\colon\ x_{aa}^i < y_{ab}^i$. Если построены эталоны (2.1), то авторство b неизвестного текста с распределением $D(j)$ определяется из условия

$$z(a) = \sum_{j=1}^{J}\left|\Psi_a(j) - D(j)\right| = \min \Rightarrow b = \arg\min z(a). \tag{2.5}$$

В качестве примера были взяты 100 отечественных и зарубежных (в переводе) авторов с суммарным объемом около 2000 произведений. Средняя длина одного текста составила 360 тыс. символов. В анализируемом корпусе оказалось (без специального подбора) 40 российских авторов и 60 иностранных. Для данного корпуса текстов были построены авторские эталоны и распределения расстояний между распределениями символов в текстах и эталонах. На Рис. 1 приведены результаты расчетов для биграмм. Показаны распределения расстояний от текста до своего эталона, между текстами одного автора, от текста до чужого эталона и между текстами разных авторов. В указанном порядке моды распределений сдвигаются в сторону увеличивающихся значений, причем расстояния от текста до эталона имеют меньший разброс, чем расстояния между текстами. Следовательно, метод ближайшего эталона представляется наилучшим для идентификации автора.

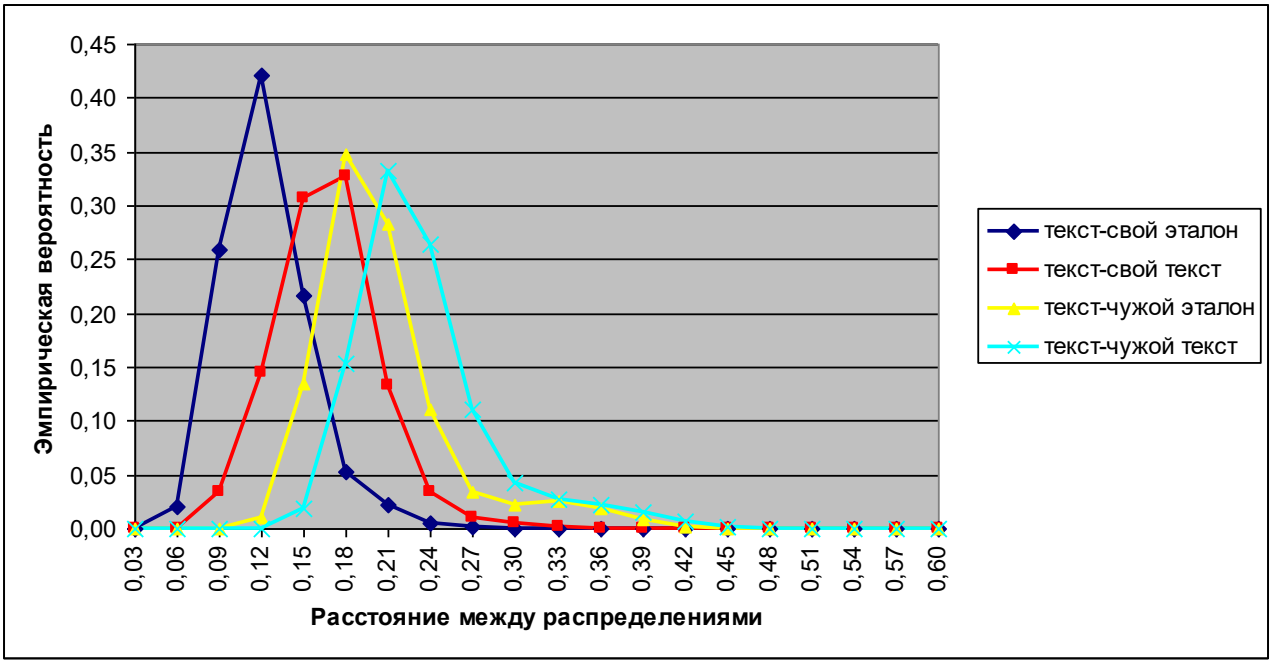

Рис. 1 – Эмпирические плотности распределений расстояний между текстами и авторскими эталонами биграмм

Тем не менее собственно метод идентификации (2.5) дал следующие результаты: для однобуквенных авторских эталонов ошибка составила 39,1 %, для биграмм 12,6 %, для триграмм 8,2 %. Следовательно, выбранную меру близости в терминах биграмм (и тем более триграмм) можно считать достаточно адекватной задаче распознавания авторов. Из анализа совместных распределений расстояний на Рис. 2 также видно, что если расстояние до своего эталона велико, то расстояние до чужого эталона часто оказывается еще больше: почти все распределение лежит справа от диагонали квадрата. Это и позволяет идентифицировать авторов.

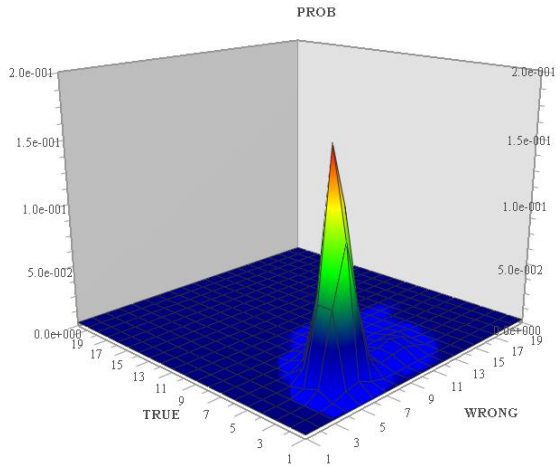

Рис. 2 – Совместная плотность распределения расстояний биграмм между текстами и эталонами «свой-чужой»

Заметим, что метод первого ближайшего соседа лучше работает на однобуквенных распределениях по сравнению с методом ближайшего эталона, а для биграмм и триграмм – наоборот. Так, для однобуквенных распределений ближайший текст дает ошибку идентификации автора на уровне 35 %, для биграмм ошибка составила 16 %, а для триграмм 13 %.

Рассмотрим теперь ближайшего соседа с точки зрения вероятности ошибочного распознавания. Для биграмм методом сравнения с эталоном были ошибочно определены авторы 200 произведений. По логике распознавания автора следовало бы заключить, что ошибочно заявленные авторы близки в плане своих эталонов к истинным авторам. В этой связи интересно изучить ближайших соседей таких ошибочно определенных авторов и сравнить их с фактическими ошибками. Насколько будут совпадать оба списка?

Граф первых ближайших соседей среди эталонов состоит из 16 несвязных фрагментов (см. таблицу 1 и Рис. 3). Каждая вершина графа 1-NN имеет ровно одно выходящее ребро, а статистика по входящим ребрам приведена ниже.

Таблица 1. Статистические свойства графа 1-NN для биграмм корпуса 100 авторов

Число несвязных фрагментов×число вершин во фрагменте	общее число вершин со степенями						
	0	**1**	**2**	**3**	**4**	**5**	**6**
5×2, 2×3, 2×4, 3×9, 2×10, 1×13, 1×17	46	31	9	9	2	1	2

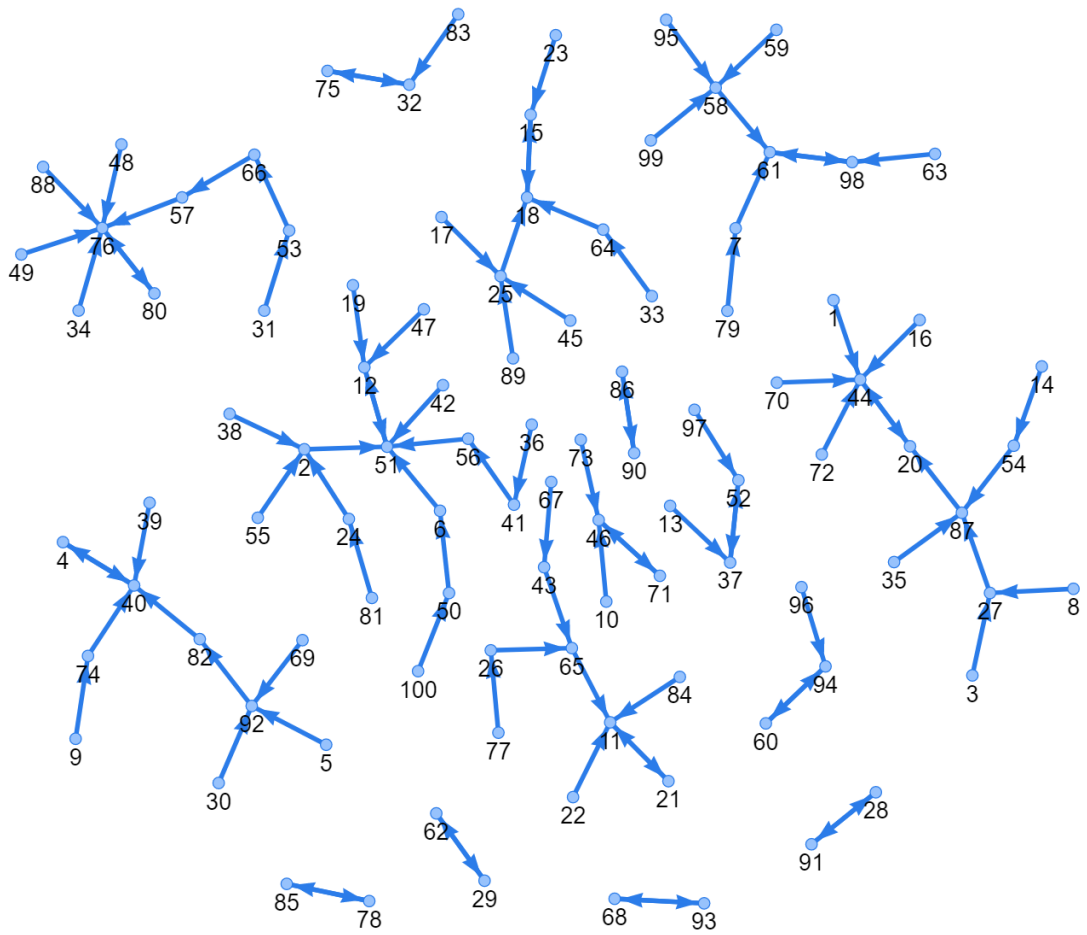

Рис. 3 – Граф 1-NN для эталонов биграмм авторов

Граф вторых соседей состоит из 6 несвязных фрагментов с количеством вершин 2, 3, 8, 12, 27 и 48. В каждом из этих фрагментов есть цикл только длины 2. Граф третьих соседей состоит всего из двух несвязных фрагментов с 18 и 82 вершинами. В большем фрагменте есть цикл длины 2, а в меньшем появился цикл длины 4. Однако не следует полагать, что с увеличением глубины соседства граф будет более связным. Графы четвертых и пятых соседей состоят из четырех несвязных фрагментов. Фрагменты первого графа содержат циклы длин 2, 3, 3 и 4, а второго графа – 2, 2, 4, 4. Распределения вершин графов по степеням приведены в таблице 2.

Таблица 2. Статистика степеней вершин графов k-NN для корпуса 100 авторов

Порядок соседства	общее число вершин со степенями							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**11**
первый	46	31	9	9	2	1	2	
второй	55	20	12	4	7	1	0	1
третий	53	18	15	7	4	3	0	
четвертый	48	27	12	6	5	1	1	
пятый	48	27	9	9	7	0	0	

Выяснилось, что ближайшие эталоны отнюдь не являются кандидатами на ошибочное распознавание. Только 24 из первых соседей, 10 из вторых и 8 из третьих являются теми, с кем были спутаны истинные авторы. Четвертый и пятый соседи совпадают с четырьмя ошибочными вариантами. Следовательно, примерно 20 % неверно определенных авторов были спутаны со своими ближайшими соседями. В остальных случаях эталон ошибочно определенного автора является скорее дальним соседом эталона правильного автора. Тем не менее все же первый сосед наиболее вероятен как кандидат на ошибочное распознавание. Тогда необходимо выяснить, насколько типичным является граф на рис. 3. Следует ли считать согласно таблице 1, что у любой группы авторов граф первых соседей имеет примерно половину вершин с нулевой степенью входящих ребер и треть вершин – со степенью 1? Иными словами, насколько важна информация, получаемая из анализа конкретного графа, для общего представления об изучаемом явлении как таковом?

Применительно к рассматриваемой проблемной области статистика ответов определяется функцией распределения расстояний между объектами. Тогда следует изучить, насколько устойчива структура графа ближайших соседей при вариации выборок: ведь на самом деле мы рассматриваем некоторый выборочный граф для конкретной группы объектов.

Рассмотрим тогда распределения расстояний между авторскими эталонами биграмм для двух разных наборов по 50 авторов (рис. 4).

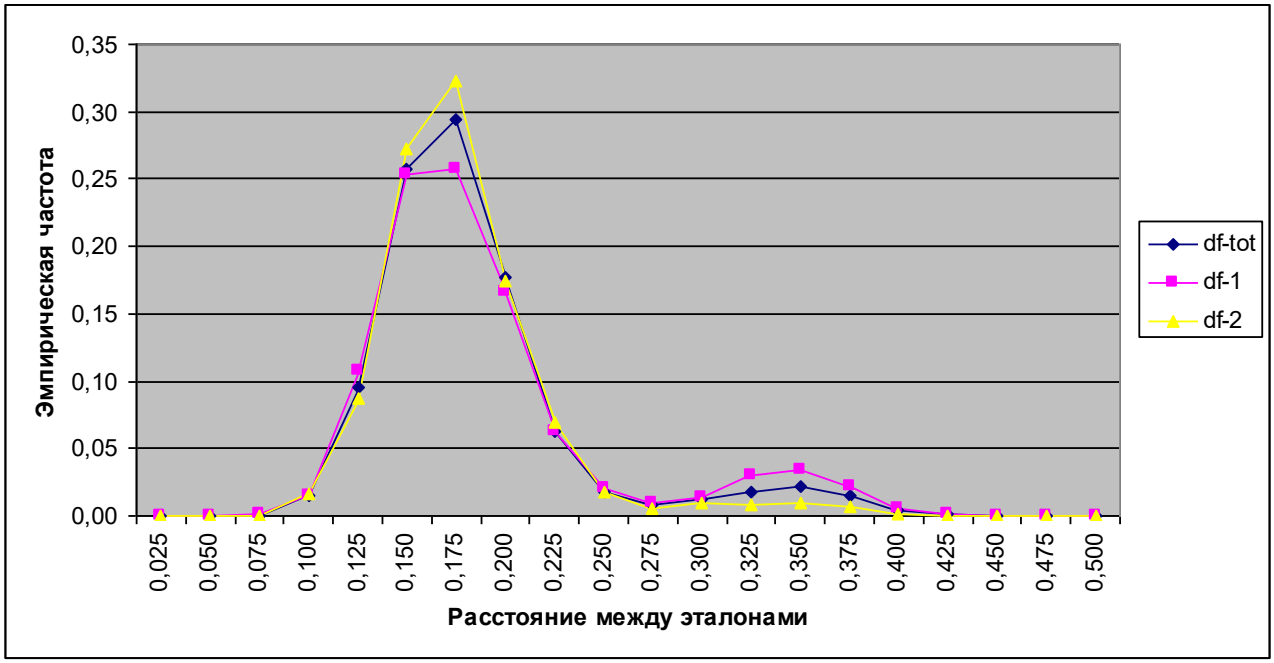

Рис. 4 – Эмпирические плотности распределения между авторскими эталонами биграмм для двух множеств авторов

Эти распределения качественно представляются однотипными. Оценка близости этих двух подвыборок показывает, что гипотеза однородности выполняется на уровне значимости 0,07, что хотя и несколько превосходит стационарную точку (0,06) для выборок таких объемов, все же может считаться приемлемым. Отметим, что такое же расстояние между подвыборками

получается и для однобуквенных эталонов, и для эталонов триграмм. Каждому из двух подмножеств авторов отвечает граф своих первых ближайших соседей. Сравнение этих графов между собой показывает, что они значительно различаются (таблица 3). Первый граф состоит из семи несвязных фрагментов, второй – из восьми. Распределения подграфов по числу вершин и вершин по степеням входящих ребер также различны.

Таблица 3. Статистические свойства графов 1-NN для двух групп авторов по распределению биграмм

Группа авторов	Число вершин в несвязных фрагментах	число вершин со степенями					
		0	**1**	**2**	**3**	**4**	**5**
I	2, 3, 5, 6, 9, 10, 15	25	13	4	5	1	2
II	2, 2, 3, 4, 6, 8, 9, 14	21	17	6	3	3	0

Аналогичные результаты для графов третьих соседей даны в таблице 4. Эти графы различаются более кардинально. Первый состоит из двух фрагментов, а второй – из четырех. Однако отметим, что в первом графе присутствуют циклы длины 2 и 3, а во втором есть два цикла длины 2 и по одному циклу длин 3 и 4.

Таблица 4. Статистические свойства графов 3-NN для двух групп авторов

Группа авторов	Число вершин в несвязных подграфах	число вершин со степенями							
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I	7, 43	24	11	12	1	0	1	0	1
II	5, 8, 13, 24	24	13	7	1	5	0	0	0

Таким образом, для сравнительного анализа полученных результатов необходимо понимать, чем обусловлены выявленные различия и насколько типичной является наблюдаемая структура связей применительно к данной системе объектов. Для ответа на этот вопрос надо сгенерировать выборки случайных величин, имеющих такое же распределение, что и расстояния между изучаемыми объектами, и собрать статистику получающихся графических структур. Насколько известно авторам, размножение выборок как инструмент статистического анализа графов в модели k-NN ранее не применялся. Он в основном использовался для моделирования случайных графов с заданным вероятностным законом связей между вершинами. Здесь же мы рассматриваем в некотором смысле обратную задачу: ищем закон распределения связей при заданной входной информации о системе объектов, а не о графе как таковом.

Матрица расстояний между авторскими эталонами может быть построена для распределений буквосочетаний различных размерностей. Все эти распределения похожи, только сдвинуты вдоль оси абсцисс в соответствии с наиболее вероятными значениями. Тем не менее отношения среднего значения

к среднеквадратичному отклонению для них различны, что и сказывается на точности соответствующих метрик: эти величины равны для однобуквенных эталонов, биграмм и триграмм соответственно 2,7, 3,2, 3,7. Для удобства сравнения преобразуем эти распределения, произведя нормировку расстояния между эталонами на размах. Результат показан на Рис. 5.

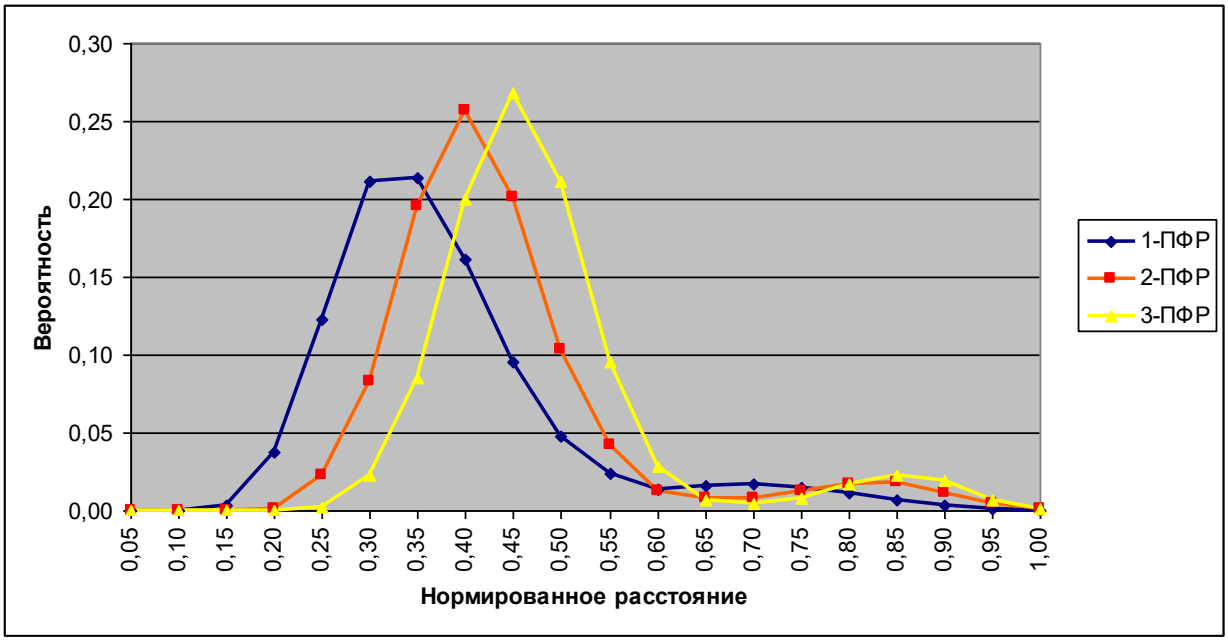

Рис. 5 – Распределения нормированных расстояний между эталонами

Построим графы ближайших соседей для однобуквенных эталонов и триграмм (Рис. 6 и 7). Для биграмм такой граф был построен выше (Рис. 3).

Разумеется, коль скоро эффективность *n*-грамм как инструмента распознавания для разных *n* различна, то не следует ожидать, что графы ближайших соседей на Рис. 3, 6 и 7 будут однотипны. Тем не менее распределения степеней вершин для них достаточно близки, как и количество несвязных фрагментов. С другой стороны, хотя ближайшие соседи относятся к одному и тому же множеству объектов, выяснилось, что только 32 связи из 100 однобуквенных эталонов совпадают со связями биграмм, остальные соседи различны. Для биграмм и триграмм совпадение ближайших соседей более хорошее: 71 вершина из 100. И все же сложно предположить, что эти три графа относятся к одному и тому же множеству объектов.

Следовательно, инструмент ближайшего соседа в данном примере не позволяет понять структуру множества изучаемых объектов на том же уровне точности, что и идентифицирующая метрика: граф k-NN связей между эталонами не позволяет понять взаимное расположение конкретных вершин. Следовательно, k-NN подход не следует применять для детального анализа характеристик множества, он слишком чувствителен к вариации расстояний между конкретными элементами множества.

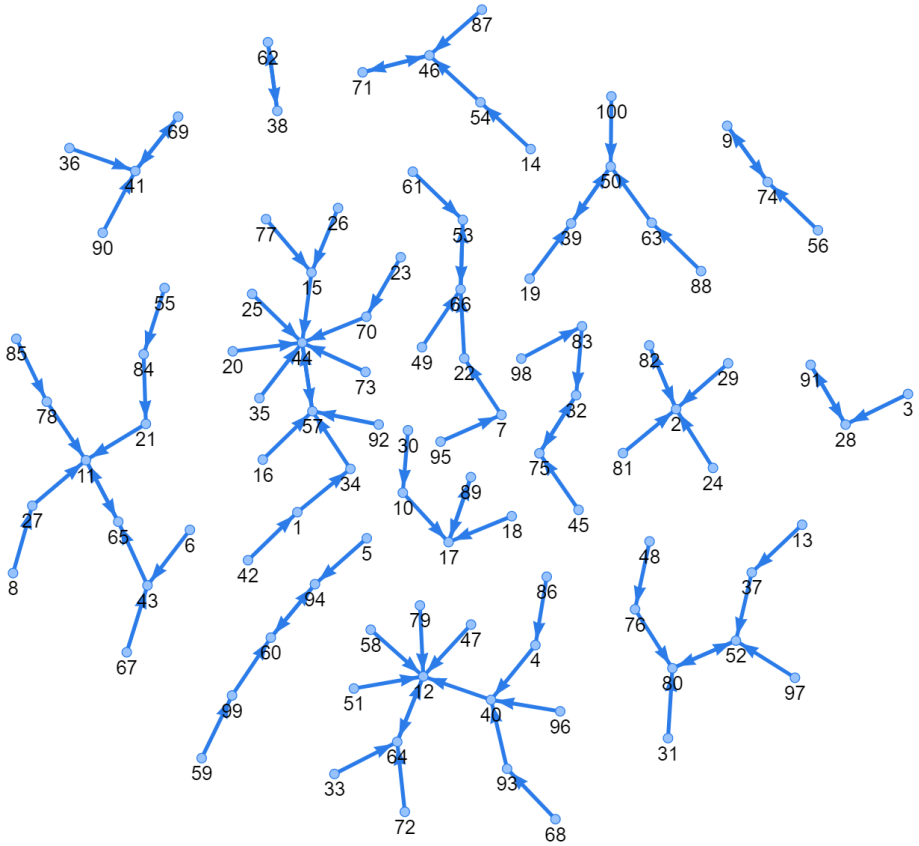

Рис. 6 – Граф 1-NN для однобуквенных эталонов

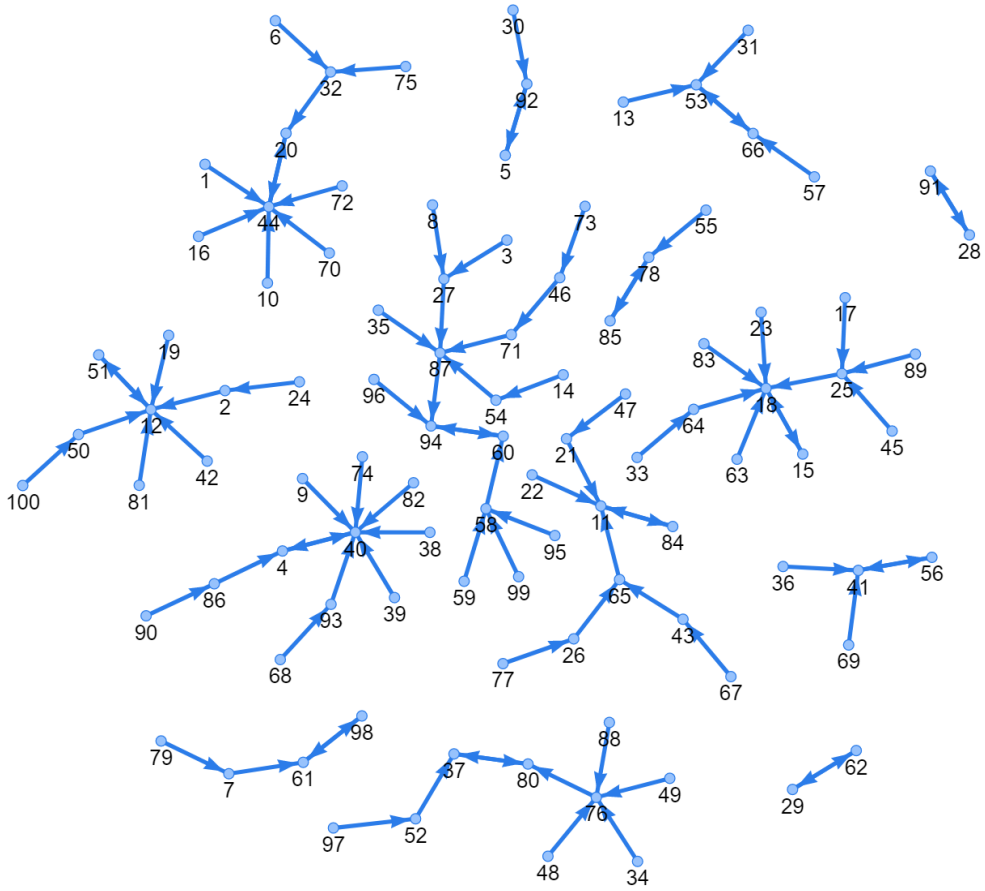

Рис. 7 – Граф 1-NN для эталонов триграмм

3. Свойства графов первых ближайших соседей

Нетривиальность вопроса о ближайших соседях для сгенерированного распределения расстояний между точками состоит в том, что это распределение строится безотносительно размерности пространства, в котором могут находиться точки.

Если количество вершин графа мало, то задача о распределении графа по числу несвязных фрагментов и вершин по числу входящих ребер может быть решена прямым перебором. Сразу оговоримся, что мы будем предполагать единственность ближайшего соседа для каждой точки множества.

Пусть $N = 2$. Тогда при любом распределении расстояний между точками они являются ближайшими соседями одна для другой. Граф связный и является циклом длины 2, так что степень каждой вершины равна единице: $1 \rightleftarrows 2$.

Для $N = 3$ матрица из трех различных расстояний всегда такова, что возможен только цикл длины 2, граф связный, одна вершина имеет степень по входящим ребрам, равную 2, одна вершина – степень 1 и одна – степень 0. Граф ближайших соседей из трех вершин – единственный с точностью до изоморфизма: $1 \to 2 \rightleftarrows 3$. Вариант с циклом длины 3 невозможен. Это – особенность именно графа первых ближайших соседей независимо от числа вершин N.

Предложение 1. Любой связный граф первых ближайших соседей имеет ровно один цикл, и это цикл длины 2. Доказательство. По условию, каждая точка имеет единственную ближайшую. Вследствие этого каждый связный граф должен содержать единственный цикл, причем из каждой вершины выходит ровно одно ребро. Наличие второго цикла означало бы, что у одной из вершин цикла существует две ближайших точки, что невозможно. Предположим теперь, что некоторые три вершины образуют круговой цикл (длины 3). Пусть это цикл $1 \to 2 \to 3 \to 1$. Тогда должны выполняться условия: $l_{23} < l_{12} < l_{13}$, $l_{32} > l_{31}$. Но в силу того, что $l_{ij} = l_{ji}$, второе неравенство противоречит первой группе неравенств, поскольку тогда получается, что $l_{23} > l_{13}$. Следовательно, такой цикл невозможен. Аналогично невозможен вообще любой цикл длины больше двух независимо от числа вершин графа, поскольку должны выполняться противоречащие одно другому условия (для цикла длины n): $l_{12} > l_{23} > ... > l_{n-1,n}$, а также $l_{12} < l_{1n}$, поскольку точка 2 есть ближайшая к точке 1. Но условие замыкания цикла означает, что вершина 1 является ближайшей для вершины n, т.е. $l_{1,n} = l_{n,1} < l_{n,n-1}$. Тогда получается, что $l_{1,n}$ есть самое меньшее расстояние из рассматриваемых, что противоречит исходному предположению о том, что $l_{12} < l_{1,n}$. **Предложение 1 доказано**.

Первая вычислительная нетривиальность появляется для $N = 4$. Здесь мы имеем два варианта графов. Первый вариант: граф связный, т.е. число фрагментов $Q = 1$. Тогда возможны следующие структуры:

1) $1 \to 2 \to 3 \rightleftarrows 4$; степени вершин: 0, 1, 1, 2.

2) $1 \to 2 \rightleftarrows 3 \leftarrow 4$; степени вершин: 0, 0, 2, 2.

3) $\begin{array}{l} 1 \to 2 \rightleftarrows 3 \\ \quad\;\; \uparrow 4 \end{array}$; степени вершин: 0, 0, 1, 3.

Второй вариант: четыре вершины разбиты на пары взаимно ближайших т.е. число фрагментов $Q = 2$. Тогда каждый фрагмент является циклом длины 2, как было указано выше для случая $N = 2$, степень каждой вершины равна 1:

$1 \rightleftarrows 2$, $3 \rightleftarrows 4$.

Таким образом, каждая структура имеет точно известный набор степеней вершин. Заметим, что соответствие между структурой графа и распределением по степеням, вообще говоря, не взаимно однозначно: одно и то же распределение по степеням может соответствовать неизоморфным структурам. Предположим, что известны вероятности реализации каждой из структур. Для $N = 4$ имеем 4 вышеприведенных варианта с вероятностями $q_1 \div q_4$. В данном случае $q_1 = q_2 = q_3 = 4/15;\ q_4 = 1/5$. Тогда, например, вероятность $\varphi(k,0)$ того, что в графе ровно k вершин имеют степень 0, равна:

$\varphi(0,0) = q_4;\ \varphi(1,0) = q_1;\ \varphi(2,0) = q_2 + q_3;\ \varphi(3,0) = 0;\ \varphi(4,0) = 0$.

Аналогично рассматриваются и другие вероятности $\varphi(k,p)$:

$\varphi(0,1) = q_2;\ \varphi(1,1) = q_3;\ \varphi(2,1) = q_1;\ \varphi(3,1) = 0;\ \varphi(4,1) = q_4$;

$\varphi(0,2) = q_4;\ \varphi(1,2) = q_1 + q_3;\ \varphi(2,2) = q_2;\ \varphi(3,2) = 0;\ \varphi(4,2) = 0$;

$\varphi(0,3) = 0;\ \varphi(1,3) = q_3;\ \varphi(2,3) = 0;\ \varphi(3,3) = 0;\ \varphi(4,3) = 0$.

В результате можно определить вероятности степеней вершин для графов с большим числом вершин. Таким образом, следующий вопрос состоит в нахождении вероятности реализации той или иной структуры в зависимости от вида $F(l)$ распределения расстояний между точками.

Предложение 2. Вероятности реализации структур графа ближайших соседей не зависят от распределения расстояний между вершинами. Доказательство. Пусть функция распределения $F(l)$ случайной величины ξ является непрерывной. Тогда, как известно из общего курса теории вероятностей, случайная величина $\eta = F(\xi)$ имеет равномерное распределение на $[0;1]$. На основе этого утверждения строятся алгоритмы генерации выборки $\{x_k\}$ с заданной непрерывной плотностью функции распределения $f(x)$, для которой $F(x)$ есть ее функция распределения. Для этого генерируется равномерно распределенная на $[0;1]$ выборка чисел $\{y_k\}$, после чего из соотношения $y_k = F(x_k)$ находятся значения $\{x_k\}$. Пусть $\{y_k\}\big|_1^n$ есть такая выборка равномерно распределенных расстояний, где $n = N(N-1)/2$. Этой выборке отвечает определенная структура соответствующего графа k-NN. В силу монотонности отображения $y_k = F(x_k)$ обратная функция также непрерывна и монотонна с тем же направлением монотонности, что и прямое отображение.

Следовательно, упорядоченность величин в выборках $\{y_k\}\Big|_1^n$ и $\{x_k\}\Big|_1^n$ совпадает. Поэтому совпадают также и графы соответствующих ближайших соседей. Тогда вероятности реализации структур графа k-NN не зависят от распределения расстояний между точками. **Предложение 2 доказано**.

Предположим теперь, что для произвольного N мы перечислили все доступные структуры реализации каждого графа ближайших соседей с точностью до изоморфизма. Структурой является конкретная совокупность фрагментов с заданным числом вершин. Пусть этих структур $M(N)$, вероятность реализации каждой из них известна (допустим, из вычислительного эксперимента) и равна $A_m(N)$, $m = 1,2,...,M$. Число $M(N)$ есть число способов, которыми можно разбить связное множество из N точек на Q множеств, $Q = 1, 2,..., [N/2]$, с числом вершин $V_1, V_2,..., V_Q$ и с условием $\sum_{k=1}^{Q} V_k = N$. В идеале это и есть решение задачи о реализуемости конкретного графа. Распределение степеней вершин для каждой структуры известно по построению, обозначим его $h_m(N,P)$, $0 \le P \le N-1$. Это распределение рассматривается именно в рамках данной структуры, так что

$$\forall m, N: \sum_{P=0}^{N-1} h_m(N,P) = 1. \tag{3.1}$$

Тогда распределение случайного графа по степеням вершин есть

$$H(N,P) = \sum_{m=1}^{M} A_m(N) h_m(N,P). \tag{3.2}$$

Проблема состоит в том, что задача определения $A_m(N)$ связана с полным перебором вариантов и имеет экспоненциальную сложность, что для больших N становится непреодолимым препятствием. Число возможных структур графа связано с числом $R(N)$ способов разбиения данного натурального числа N в сумму натуральных чисел. Асимптотическая оценка числа разбиения дается формулой Харди – Рамануджана [12]:

$$R(N) \approx \frac{1}{4\sqrt{3}N} \exp\left(2\pi\sqrt{(N - 1/24)/6}\right). \tag{3.3}$$

Возможные разбиения представляются обычно в виде диаграмм Юнга [13], предельной формой которых является полукруговое распределение Вигнера, которым также определяется предельная плотность распределения собственных значений стохастической матрицы при увеличении ее размерности. Эти замечания о связи данной задачи с классическими результатами комбинаторики и теории вероятностей служат мотивацией проводимого исследования. В частности, естественно возникает вопрос о числе различных фрагментов, которые можно наблюдать для графа ближайших соседей.

Предложение 3. Количество $M(N)$ различных фрагментов, в виде которых может быть реализован граф первых ближайших соседей из N точек,

асимптотически равно главной части производной числа разбиений $R(N)$ по числу вершин:

$$M(N) \approx \frac{\pi R(N)}{\sqrt{6N}}. \qquad (3.4)$$

Доказательство. Число структур $M(N)$ – это число $R(N)$ представлений числа N в виде суммы натуральных чисел за вычетом количества невозможных структур графов применительно к задаче k-NN. Таким образом, среди возможных фрагментов нет изолированных точек, которых может быть одна, две, …, N. Обозначим это количество недопустимых структур через $M^1(N)$. Тогда $M(N) = R(N) - M^1(N)$. Заметим теперь, что число разных структур учитывается с точностью до изоморфизма. Это означает, что существует $M(N-1)$ вариантов, когда имеется одна изолированная вершина, $M(N-2)$ вариантов для двух изолированных вершин и т.д. Следовательно, число $M^1(N)$ представляется в виде суммы $M^1(N) = \sum_{k=2}^{N-1} M(k)$. Тогда из вышеприведенного равенства $M(N) = R(N) - M^1(N)$ следует, что $R(N) = \sum_{k=2}^{N} M(k)$. Таким образом, $M(N) = R(N) - R(N-1)$, откуда и следует (3.4). **Предложение 3 доказано**.

Экспоненциальный рост числа способов разбиения связного графа на фрагменты приводит к тому, что надо отказаться от детального рассмотрения структуры графа и перейти к более укрупненным его характеристикам. Укрупнение приводит к изучению вероятности реализаций тех или иных распределений, т.е. к так называемым метараспределениям. Пусть $G_N(Q)$ есть вероятность реализации графа ближайших соседей в виде Q фрагментов, а $U_N(Q,V)$ есть распределение этих фрагментов по числу вершин, т.е. это есть условная вероятность того, что в графе из N вершин, состоящем из Q фрагментов, существует фрагмент с данным количеством V вершин.

Для описания этого семейства графов ключевым объектом является количество неизоморфных связных графов с данным числом вершин. Обозначим эту величину через $Y(V)$. Если, например, граф состоит из двух фрагментов с числом вершин в них V_1 и V_2, то число неизоморфных графов в таком семействе равно $Y(V_1)Y(V_2)$. В силу независимости структур графа k-NN от вида распределения расстояний между вершинами вероятность каждого графа в пределах данного семейства одинакова и равна $1/\prod_k Y(V_k),\ k = 1,2,...,Q;\ \sum_k V_k = N$.

Предложение 4. Количество неизоморфных представлений связного графа первых ближайших соседей из V вершин равно

$$Y(V)=\sum_{k=0}^{[V/2]-1} C(k)C(V-k-2), \qquad (3.5)$$

где $C(n)=\frac{1}{n+1}C_{2n}^{n}$ – числа Каталана. Доказательство. Цикл длины 2 разбивает связный граф из V вершин на два подграфа, каждый из которых присоединяется только к одной из двух вершин цикла. Эти графы являются деревьями с вершинами $k+1$ и $V-k-1$ соответственно: это количество вершин, расположенных слева (справа) от цикла, плюс одна из вершин цикла, которая является частью дерева – левого или правого. Неизоморфных комбинаций таких графов будет $[V/2]-1$: цикл длины 2 перемещается от крайнего (допустим, левого) положения, когда слева от него 0 ветвей, а справа – дерево с $V-2$ ветвями, до середины. Как известно из теории графов, неизоморфных деревьев общего вида с n вершинами существует $C(n-1)$, где $C(n)$ – числа Каталана. Тогда число возможных сочетаний дерева с числом вершин $k+1$ и независимо от него дерева с числом вершин $V-k-1$ равно $C(k)C(V-k-2)$. **Предложение 4 доказано**.

В частности, $C(0)=1,\ C(1)=1,\ C(2)=2,\ C(3)=5,\ C(4)=14,...$. Тогда из (3.5) получаем следующие числа для определения количества неизоморфных связных графов 1-NN в зависимости от количества N вершин: $Y(2)=1$, $Y(3)=1$, $Y(4)=3$, $Y(5)=6$, $Y(6)=23$.

Асимптотика чисел Каталана имеет вид:

$$C(n)\approx\frac{4^N}{N^{3/2}\sqrt{\pi}}, \qquad (3.6)$$

что следует из формулы Стирлинга. Отсюда, в свою очередь, следует, что величины $Y(N)$ из (3.5) имеют ту же асимптотику, что и $C(n)$, только с меньшим коэффициентом. Главная часть по N суммы (3.5) положительных членов совпадает с главной частью соответствующего интеграла, так что с учетом (3.6) получаем

$$Y(N)\approx\int_1^{N/2-1} C(x)C(N-2-x)dx=\frac{4^{N-2}}{\pi}\int_1^{N/2-1}\frac{dx}{\left((N-2)x-x^2\right)^{3/2}}=$$

$$=\frac{4^{N-1}}{\pi(N-2)^2}\int_0^{1-2/(N-2)}\frac{dx}{\left(1-x^2\right)^{3/2}}= \qquad (3.7)$$

$$=\frac{4^{N-1}}{\pi(N-2)^2}\frac{1-2/(N-2)}{\sqrt{1-\left(1-2/(N-2)\right)^2}}=\frac{4^N}{8\pi N^{3/2}}\left(1+o(1)\right).$$

Таким образом, асимптотика числа неизоморфных графов 1-NN примерно такая же, как и число неизоморфных деревьев.

Подчеркнем, что распределение вершин по степеням зависит от конфигурации, которая хотя и равновероятна в пределах данного разбиения на

фрагменты с заданным числом вершин, но для разных разбиений, естественно, различна. Поэтому распределение вершин по степеням, отвечающее некоторой конфигурации из Q фрагментов безотносительно к конкретной реализации этих фрагментов в виде определенного распределения вершин, не определено однозначно. Оно само имеет некоторое распределение $B_N(Q,H,P)$: это есть вероятность того, что в данной конфигурации степень P имеет частоту реализации H. На практике вероятность $B_N(Q,H,P)$ оценивается по результатам S_Q экспериментов, в ходе которых получилась реализация Q фрагментов (доля этих экспериментов и равна эмпирической вероятности $G_N(Q)$). В каждом таком j-м эксперименте определяется количество $n_j(P)$ вершин, которые имеют степень P, так что соответствующая эмпирическая частота равна $H_j(P)=n_j(P)/N$. Совокупность $\{n_j(P)\}$, $j=1,2,...,S_Q$ образует эмпирическое распределение $B_N(Q,n,P)$ числа n вершин, имеющих в данной конфигурации Q степень P:

$$\sum_{n=0}^{N-1} B_N(Q,n,P)=1. \tag{3.8}$$

Если мы теперь рассматриваем произвольную реализацию графа, то его распределение по степеням вершин имеет вид

$$\Phi_N(n,P)=\sum_Q G_N(Q)B_N(Q,n,P), \quad \sum_{n=0}^{N-1}\Phi_N(n,P)=1. \tag{3.9}$$

Таким образом, эмпирическое метараспределение $\Phi_N(n,P)$ представляет собой стохастическую матрицу. Каждая реализация графа с N вершинами есть выборка пар $(n_0,0),(n_1,1),...$ с условием $\sum_k n_k = N$.

Наиболее вероятное распределение вершин по степеням определяется максимумом совместного распределения вероятностей $\Psi(P_0,...,P_{N-1})$ для степеней $0,...,N-1$.

При числе экспериментов S, стремящемся к бесконечности, эмпирические распределения $G_N(Q)$, $B_N(Q,n,P)$ и, как следствие, $\Phi_N(n,P)$ сходятся почти наверное к соответствующим генеральным распределениям (то есть предел при $S\to\infty$ вероятности того, что разность между эмпирической и генеральной функциями распределения меньше ε для любого $\varepsilon>0$, равен единице.) Нахождение вида генеральных распределений при больших N есть одна из основных задач анализа графов k-NN.

В дальнейшем планируется провести детальный анализ структур распределений $\Phi_N(n,P)$ в зависимости от N и n. Здесь мы приведем иллюстрирующие ситуацию рисунки по результатам 1 млн генераций выборок длины $N(N-1)/2$ из равномерного распределения как вариантов допустимых исходных множеств парных расстояний, т.е. элементов матрицы $(l_{ij})_{N\times N}$.

Поскольку, как было показано выше, статистика выборочных k-NN графов не зависит от вида распределения расстояний между точками, достаточно ограничиться равномерным распределением. Мы проанализируем, насколько ситуация, описанная выше для графов ближайших соседей эталонов писателей, является типичной. На Рис. 8-9 показаны распределения $\Phi_N(n,P)$ для $N=50$ и $N=100$.

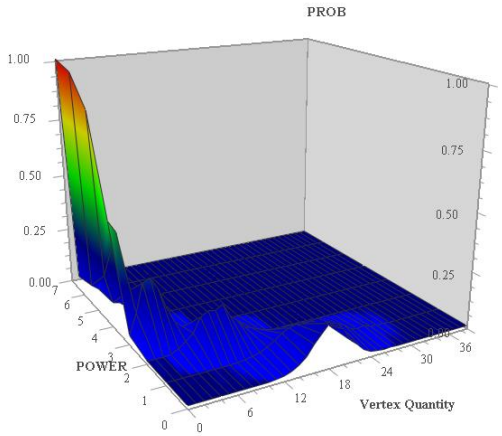

Рис. 8 – Распределение вершин по степеням для N=50

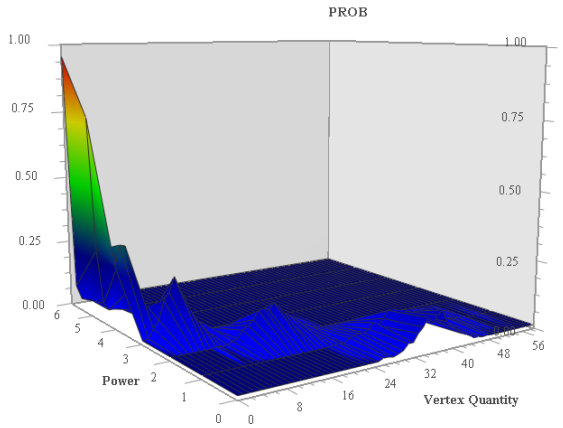

Рис. 9 – Распределение вершин по степеням для N=100

Сравнивая, например, данные на Рис. 9 с аналогичными показателями из табл. 2 и Рис. 3, делаем вывод о том, что в выборке авторов текстов граф связей весьма нехарактерен по сравнению с типовым, то есть наиболее вероятным случаем. На Рис. 9 наиболее вероятное значение числа вершин с нулевой степенью равно 40, со степенью 1 – 37, со степенью 2 – 15, со степенью 3 – 4. Вероятность такого распределения степеней, оцененная из статистического эксперимента, равна $3\cdot10^{-4}$. Надо, конечно, учитывать, что вероятность реализации каждой конкретной ситуации достаточно мала, поэтому интерес представляют относительные величины. Так, вариант на Рис. 3 отвечает

вероятности его реализации на уровне $3 \cdot 10^{-8}$, то есть он в десять тысяч раз менее вероятен, чем типичный граф ближайших соседей. Это свидетельствует о том, что данная конкретная выборка образована во всяком случае объектами с высокой степенью коррелированности. Применительно к писателям этот вывод представляется логичным: по-видимому, они имеют некоторый общий уровень литературного образования, читали одни и те же книги, а возможно, и тексты друг друга. Следовательно, результаты кластеризации в данном случае могут трактоваться как существенные применительно к конкретной группе объектов. Если бы граф был типичным, причин именно таких связей, а не каких-то других, искать не имело бы смысла. Здесь же можно содержательно обсуждать причины возникновения конкретных кластеров.

4. Заключение

В работе ставится задача о создании математической статистики применительно к графам. В частности, рассматриваются графы ближайших соседей, для которых могут быть получены не только численные, но и некоторые аналитические результаты.

В результате проведенного анализа выяснилось, что для практического решения задачи кластеризации данных методом k-NN необходимо знать априорные оценки вероятностей графических реализаций, и даже в этом относительно простом случае требуются значительные машинные вычисления. Основные распределения, характеризующие семейства графов k-NN, были введены в разделе 3. Существенный вывод состоит в том, что число возможных структур графа (т.е. его несвязных фрагментов) экспоненциально растет с ростом числа вершин, причем число неизоморфных конфигураций, отвечающих каждой такой структуре, растет по степенному закону. Поэтому практически важная информация, содержащаяся в распределении степеней вершин графа для решения задачи об определении вероятности реализации наблюдаемой структуры, может быть получена численно только для относительно небольших размерностей. Большой интерес представляют асимптотические оценки этих вероятностей, что является темой отдельного исследования.

Численно определяемые распределения вершин графа по степеням можно использовать как новый метод тестирования множества на наличие скрытых взаимосвязей: если реализовалась наиболее вероятная структура, то, по построению, она отвечает случайному расположению объектов.

Раздел 1 написан Ю.Н. Орловым, раздел 3 написан А.А. Кислицыным, раздел 2 написан совместно.

Литература

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – Springer, 2001. – 764 p.

2. Erdos P., Renyi A. On random graphs I // Publ. Math. Debrecen, 1959. V. 6. P. 290-297.
3. Колчин В.Ф. Случайные графы. – М.: Физматлит, 2004. – 256 с.
4. Bollobas B. Random Graphs. – Cambridge University Press, 2001. – 520 p.
5. Janson S., Luczak T., Rucinski A. Random graphs. – New York: Wiley, 2000. – 333 p.
6. Райгородский А.М. Модели случайных графов и их применения. // Труды МФТИ, 2010. Т. 2. № 4. С. 130-140.
7. Райгородский А.М. Модели Интернета. – Долгопрудный, Издательский Дом «Интеллект», 2013. – 64 с.
8. Barabasi L.-A., Albert R. Emergence of scaling in random networks // Science, 1999. V. 286. P. 509-512.
9. Leskovec J., Chakrabarti D., Kleinberg J., Faloutsos C., Gharamani Z. Kronecker graphs: an approach to modeling networks // J. Machine Learning Research, 2010. V. 11. P. 985-1042.
10. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 312 с.
11. Полянин А.Д., Манжиров А.В. Справочник по интегральным уравнениям. – М.: Физматлит, 2003. – 608 с.
12. Hardy G.H., Ramanujan S. Asymptotic Formulae in Combinatory Analysis. Proc. London Math. Soc., 1918, vol. 17, pp. 75-115.
13. Вершик А.М., Керов С.В. Асимптотика меры Планшереля симметрической группы и предельная форма таблиц Юнга // Доклады Академии Наук СССР. 1977. Т. 223 № 6. С. 1024-1027.