

Шурыгина А.А.¹, Самсонов Т.Е.²

ИСПОЛЬЗОВАНИЕ ЭЛЕМЕНТОВ ТЕОРИИ ГРАФОВ И МАШИННОГО ОБУЧЕНИЯ В РАСПОЗНАВАНИИ СЕТЕЙ ПРИРОДНЫХ И АНТРОПОГЕННЫХ ЛИНЕЙНЫХ ОБЪЕКТОВ

АННОТАЦИЯ

Представляемая работа является частью исследования, которое направлено на выявление подходов к распознаванию пространственных объектов по векторным моделям данных. На предыдущем этапе классификация объектов основывалась на наборе их морфометрических признаков, – был сделан вывод о недостаточности подобных характеристик для автоматизированного распознавания и необходимости дополнительного изучения пространственных взаимосвязей объектов, т. е. перехода от объектного уровня распознавания к классовому, от анализа формы объектов к исследованию сочетаний форм – рисунков. Важно рассмотреть пространственные взаимосвязи как объектов одного класса, так и разных. В рамках текущей работы решалась задача классификации групп пространственных объектов различных классов.

В статье описаны направления использования результатов исследования, методика решения поставленной задачи, материалы, задействованные в работе. Проведено сравнение результатов классификации объектов по морфометрическим и графовым признакам.

Представлены подходы к автоматизированному распознаванию природных и антропогенных объектов линейной локализации на векторной модели данных. Применена теория графов и элементы машинного обучения для их классификации. Результатом работы стал инструмент, разработанный на языке программирования Python, который конвертирует сеть линейных объектов в граф и рассчитывает определенный перечень его параметров. Также в программном продукте **Rapid Miner Studio** подготовлена модель для машинного распознавания сетей автомобильных и железных дорог, рек, каналов. Тестирование разработанной технологической схемы проводилось на цифровых данных трех уровней детализации и шести типов ландшафтов, чтобы оценить устойчивость алгоритмов к генерализации данных и изменчивости рисунков сетей в различных природных условиях.

КЛЮЧЕВЫЕ СЛОВА: машинное обучение, распознавание графических образов, теория графов.

¹ Московский государственный университет имени М.В. Ломоносова, Ленинские горы, д. 1, 119991, Москва, Россия, *e-mail*: shur.a17@yandex.ru

² Московский государственный университет имени М.В. Ломоносова, Ленинские горы, д. 1, 119991, Москва, Россия, *e-mail*: tsamsonov@geogr.msu.ru

Anastasia A. Shurygina¹, Timofey E. Samsonov²

RECOGNITION OF NATURAL AND ANTROPOGENIC NETS BASED ON THE GRAPH THEORY AND METHODS OF MACHINE LEARNING

ABSTRACT

Current study is a part of the research which is aimed at identifying approaches to the recognition of spatial data objects presented in the vector data models. In the previous stage the classification of spatial objects was based on the set of their morphometric features – shape characteristics. A conclusion was drawn that such attributes were insufficient for automated recognition and there was a need in additional study of the spatial relationships of objects. That means the transition from the object recognition level to the feature classes recognition level, from shape analysis to the exploration of objects combinations – patterns. It is important to analyse spatial relationships between objects of the same classes as well as different ones. That study covers the problem of classifying various feature classes.

Application options of the study results are described in the paper on a par with the methodology of solving the task and materials involved in the study. Moreover, the results of objects recognition based on their morphometric features and graph characteristics are compared.

The paper presents the ways of automated recognition of natural and antropogenic vector objects with the linear localization. The study applies Graph theory and Machine learning to classify them. The Python script for calculating graph parameters of linear objects was created. Furthermore, the model in Rapid Miner Studio application program was prepared for spatial objects' nets recognition.

KEYWORDS: graph theory, machine learning, pattern recognition.

ВВЕДЕНИЕ

Теория распознавания образов, занимающаяся классификацией групп графических объектов на основе заданных признаков [Лурье, Косиков, 2003], находит сегодня многоаспектное практическое применение. Распознавание объектов на растровых моделях представления данных входит в перечень задач цифровой фотограмметрии и дистанционного зондирования. Развитие технологий распознавания в картографии является одним из факторов создания системы полностью автоматизированного составления карт, что важно в первую очередь для их оперативного обновления и укладывается в концепцию Цифровой Земли. Исследования в этом направлении проводятся, например, компаниями Facebook и DigitalGlobe, о чем упоминается в статье Е. Лисовского «Искусственный интеллект и нейросети в картографии»³.

Задача распознавания пространственных объектов на векторных моделях также актуальна. Накопление знаний об особенностях геометрических и структурных признаков (формы и рисунка) объектов дополнит методы классификации растровых изображений, сделает возможным создание системы контроля автоматизированного и краудсорсингового наполнения картографических баз данных (как геометрической составляющей, так и семантической), позволит генерировать картографические образы геосистем заданных свойств.

Цель работы – определить способы формализованного описания сетей линейных пространственных объектов для их автоматизированного распознавания методом машинного обучения с целью решения геоинформационных задач, связанных с распознаванием пространственных образов.

¹ Lomonosov Moscow State University, Leninskie gory, 1, 119991, Moscow, Russia, *e-mail*: shur.a17@yandex.ru

² Lomonosov Moscow State University, Leninskie gory, 1, 119991, Moscow, Russia, *e-mail*: tsamsonov@geogr.msu.ru

³ URL:<http://www.forbes.ru/tehnologii/351121-iskusstvennyy-intellekt-i-neyroseti-v-kartografii-2-kogda-narodnye-karty-kruche> (дата обращения 06.12.17).

Достижение цели осуществлялось в несколько этапов:

- Определение набора исследуемых характеристик классов объектов как графов и разработка на языке программирования Python специализированных инструментов для вычисления их значений.
- Создание тестового набора линейных пространственных объектов разных классов и применение к ним разработанного инструмента анализа.
- Подготовка модели в специализированном программном продукте для машинного распознавания выбранных классов на основе обучающей выборки.
- Формулирование выводов об эффективности использованного подхода.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Природные и антропогенные объекты, которые на общегеографических картах изображаются линейным знаком, можно воспринимать как неориентированный граф. Например, дорожная сеть состоит из перекрестков – вершин и собственно дорог – ребер.

Во многих прикладных задачах, в которых графы используются как модели определенных явлений, требуется анализ положения некоторых вершин относительно остальных: определение центра сети, кратчайшего расстояния между двумя вершинами, возможности проложить путь между заданными вершинами и др. Вследствие этого математика предлагает широкий аппарат для анализа взаиморасположения элементов графа, характеристик его составных частей, нежели графовой структуры целиком.

В рамках решаемой в работе задачи необходим именно целостный подход к описанию пространственных сетей, поскольку перекрестки, места слияния рек, железнодорожные узлы и фрагменты дорог и водотоков между ними по отдельности не несут информации об особенностях явлений, не характеризуют рисунок явлений, потому что являются его фрагментами. Кроме того, ранее была исследована возможность использования такого признака, как среднее арифметическое индексов узлов графов (сетей линейных объектов общегеографических карт): будучи признаком, обобщающим характеристики отдельных элементов графа, он не позволил качественно различать сети – необходимы более сложные характеристики, опирающиеся на свойства графа как абстрактной модели.

Из общих свойств графа в работе использованы: *средний коэффициент кластеризации* (отражает полноту связей в сети); *транзитивность сети* – отношение числа всех возможных треугольников, которые могут быть образованы тремя узлами и тремя ребрами, к числу присутствующих в сети триад – структур, состоящих из трех узлов и двух ребер [Schank, Wagner, 2004]; *плотность графа*,

$$D = \frac{2 \cdot e}{v \cdot (v - 1)}$$

где e – количество ребер в графе, v – число вершин; *коэффициент ассортативности*, характеризующий наличие явления тяготения узлов с большими степенями друг к другу [Newman, 2003]; число *компонент связности* графа – частей графа, в пределах которых возможно проложить путь из любой вершины в любую другую [Reingold, 2008].

Природные и антропогенные сети в пространственных базах данных хранятся в виде упорядоченного набора объектов, например, рек. Все реки обязательно имеют две точки – исток и устье, причем значительное число рек обладает притоками, устьевые точки которых расположены на принимающих водотоках. Для того чтобы представить речную сеть в виде графа, необходимо все точки истоков и устьев принять за его узлы, а участки водотоков между ними – за ребра графа. Следовательно, возникает задача предварительной подготовки данных, а именно: выделение всех пар указанных точек и разрезание в них объектов более высокого порядка (крупных рек в местах впадения притоков).

Для ее решения был создан программный код на языке программирования Python, использующий функциональность библиотеки *arcpy* и позволяющий сформировать таблицу, в которой каждому фрагменту между двумя точками пересечения линий (ребру будущего графа) поставлены в соответствие идентификационные номера его концевых точек (узлов) и длина.

Таблица становится исходным материалом для работы программного кода, реализованного на языке Python: на ее основе строится граф. Далее к нему применяются готовые методы библиотеки *NetworkX* и авторские функции и рассчитанные характеристики графа записываются в таблицу MS Excel.

В программном пакете Rapid Miner Studio, который работает с табличными данными, подготовлен шаблон модели для классификации объектов по ограниченному набору признаков. Ядром модели является инструмент голосования нейронной сети и дерева решений.

Решающие правила создавались на основе обучающей выборки, в качестве которой использовались сети автомобильных и железных дорог, речная сеть и сеть каналов территории Московской области, Республики Коми, Алтайского края в трех масштабах (1 : 10 000, 1 : 100 000 и 1 : 1 000 000), фрагментов территории Республики Карелия, Республики Дагестан, Камчатского края и Самарской области в двух масштабах (1 : 10 000 и 1 : 1 000 000) – всего 58 наборов. Так как многие из используемых коэффициентов интуитивно кажутся близкими по смыслу и есть вероятность линейной связи их значений для тестовых наборов данных, был проведен попарный корреляционный анализ полученных рядов данных – наибольшая связь обнаружена между значениями коэффициентов транзитивности и кластеризации – 0,5.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

На основе обучающей выборки была выявлена значимость каждого атрибута при классификации (в порядке убывания): транзитивность, кластерность, количество компонентов, ассортативность и плотность.

Построено дерево решений (рис.): в узлах указан признак разбиения, на его ребрах – границы классов, листья – итоговые классы. Ширина полосы под названием класса говорит об относительном числе объектов, пришедших при распознавании именно к этому листу, а цвета – о том, правильно ли они были распознаны, много ли ошибочных решений в данном листе. Видно, что главным признаком железнодорожных сетей является относительно высокая степень кластеризации – наличие разрозненных уплотнений сети, узлов, вокруг которых возрастает интенсивность железнодорожного сообщения. Однако на территориях, где его концентрация уменьшается, железные дороги иногда принимаются за автомобильные. Низкой кластерностью и транзитивностью отличаются реки, потому что для речной сети не характерно явление сосредоточения вокруг некоторого узла, только близ нескольких ребер, составляющих водоток высшего порядка. Также им не свойственна транзитивность – наличие треугольников. Тем не менее речные сети равнинных обводненных территорий, с множеством протоков, могут по признаку транзитивности быть распознанными как каналы, отличающиеся низкой кластерностью и высокой транзитивностью. Основным свойством автомобильных дорог выступает их транзитивность – наличие альтернативных маршрутов между пунктами. Из этого следует повышенная плотность и сниженная относительно железнодорожной сети кластерность таких сетей.

Учитывая, что выборка была составлена из сетей разного уровня генерализации и типов ландшафтов и на ее основе было получено компактное дерево решений, можно рассматривать графовые свойства природных и антропогенных сетей как значимые дополнительные признаки при их распознавании.



Дерево решений для классификации линейных пространственных объектов по графовым признакам

The decision tree for the classification of linear spatial objects by graph characteristics

Следующий возникающий вопрос – корректно ли признаки линейных объектов как графов анализировать совместно с их морфометрическими характеристиками? С одной стороны, значения графовых свойств рассчитываются для совокупности аналогичных типов пространственных объектов, а не для каждого объекта в отдельности (как это выполнялось для морфометрических расчётов), поэтому в случае их распространения на элементы сети обучающая выборка станет нерепрезентативной, потому что такие атрибуты будут однозначно идентифицировать класс объекта и не будут давать диапазон значений, в пределах которого может находиться признак, что затруднит классификацию произвольных объектов. С другой стороны, задача используемых алгоритмов распознавания заключается не в соотношении атрибутов неклассифицированных объектов с вариациями значений соответствующих признаков обучающей выборки, а в нахождении границ классов в признаковом пространстве. Поэтому наличие констант в векторах признаков объектов не должно влиять на качество работы алгоритмов.

Для исследования влияния наличия графовых признаков на результат классификации использовались выборки из линейных классов объектов цифровых топографических карт масштабов 1 : 10 000, 1 : 100 000 и 1 : 1 000 000 на территорию Московской области. В атрибутивные таблицы исходных классов пространственных объектов были добавлены поля, отвечающие за графовые характеристики, и рассчитаны морфометрические (извилистость отдельных линий, ортогональность, схематичность [Шурыгина, Самсонов, 2017]). Данные региона были условно разделены на две зоны – северо-западную и юго-восточную. Для каждой половины графовые показатели рассчитывались отдельно. Из данных первого фрагмента формировались обучающие выборки, по данным второго – наборы для тестирования моделей обучения.

Основная тенденция после добавления новых, графовых, признаков для распознавания заключается в смещении значимости характеристик в сторону графовых параметров: ассортативность, кластерность и число компонент теперь входят в пятерку важнейших признаков классификации по оценке моделей машинного обучения. Достоверность распознавания тестовых объектов имеет две особенности. При первичной классификации экспериментальных данных правильность распознавания практически всех групп выборок имела более низкое (в пределах нескольких десятков процентов) значение, чем при работе моделей только по морфометрическим признакам. В то же время путем подбора усеченной группы параметров практически для всех выборок удалось найти такое сочетание характеристик, которое позволяет классифицировать независимые данные с абсолютной достоверностью. Однако такие наборы индивидуальны для каждой группы классов объектов, их сложно унифицировать даже в пределах одного масштаба. Как правило, перечень необходимых и достаточных признаков содержит некоторые графовые характеристики (в первую очередь ассортативность) и 1–2 морфометрические.

Объекты масштаба 1 : 10 000 успешно распознавались всего лишь по двум параметрам – ассортативности и количеству компонент. Единственный набор данного масштаба,

качество классификации которого не удалось довести до 100 %, состоит из классов рек, каналов, автомобильных и железных дорог (каналы относились к железным дорогам). Тем не менее полученный результат улучшился по сравнению в предыдущим этапом исследования: достоверность повысилась с 59 до 77 %.

В более мелком масштабе абсолютной достоверности также не удалось достичь только в самой крупной по числу классов группе. Наиболее важными признаками в данной масштабной совокупности стали ассортативность и кластерность. Относительно часто использовались длина линии и ее фрактальная размерность. Ранее ключевыми признаками объектов двух крупнейших масштабов выступали схематичность, замкнутость линии, ее длина, фрактальная размерность и извилистость.

Наконец, в масштабе 1 : 1 000 000 повышена распознаваемость группы из всех доступных линейных объектов; не изменилась достоверность классификации набора «реки, авто- и железные дороги» (83 %). Остальные совокупности успешно распознавались по признакам ассортативности и кластерности, часто задействовались показатели извилистости и длины линий, реже – ортогональности, фрактальной размерности и плотности графа. Ранее объекты данной масштабной группы классифицировались алгоритмом по схожему набору морфометрических характеристик: схематичность, ортогональность, извилистость, длина линии и замкнутость.

Проведенный эксперимент показал, что рассмотрение линейных пространственных объектов как графов может значительно повысить качество их распознавания и что допустимо составлять векторы признаков объектов как из индивидуальных морфометрических показателей, так и из общеклассовых графовых. Однако обучение и верификация моделей проводились на однотипных по виду рисунка территориях, графовые признаки которых действительно близки, и успешным распознавание объектов было только по неполному набору признаков, подобранному для каждой выборки, причем такие наборы состоят из 2–3 графовых характеристик и нескольких морфометрических параметров-помощников. Данное замечание говорит об отсутствии гибкости моделей: высокое доверие моделей к графовым характеристикам может привести к ложной классификации объектов похожего рисунка сети, но незначительно отличающихся значений графовых признаков, и ошибки распознавания по этой группе параметров с малой вероятностью будут исправлены морфометрическими показателями, потому что обновленный набор необходимых и достаточных признаков содержит недостаточное для классификации число индивидуальных показателей формы.

Дополнительно была проверена модель для распознавания рек, каналов, автомобильных и железных дорог масштаба 1 : 100 000 на данных фрагмента Ленинградской области. Общая достоверность классификации составила 62 %. Безошибочно была распознана речная и дорожная сеть, однако все каналы и железные дороги вошли в класс автомобильных дорог. В этом заключается еще один аспект влияния общеклассовых графовых характеристик: если модель ошибается, то ошибается сразу на всех объектах одного класса, потому что они имеют постоянные значения графовых признаков и, как уже говорилось выше, морфометрические показатели не могут «спасти» классификационный процесс.

Вероятно, использование графовых характеристик при распознавании линейных пространственных объектов совместно с морфометрическими не всегда эффективно и требует отдельного поэтапного использования.

ВЫВОДЫ

Взаиморасположение пространственных объектов, в частности линейных, определяет рисунок местности, ее образ. С изменениями, связанными с процессом генерализации, он отображается в моделях пространства – картах и снимках. Анализируя сочетания

объектов геоизображений, человек получает информацию, пропорциональную его знаниям и опыту. Автоматизация этого процесса становится актуальной задачей, поскольку ее решение позволит повысить эффективность дешифрирования снимков, открыть новые возможности анализа, генерализации и использования карт, выявить неочевидные свойства пространственных объектов, автоматизировать процесс создания карт.

В работе представлены способы применения теории графов для решения задач, связанных с распознаванием классов линейных пространственных объектов. Параметры, характеризующие графовые структуры как цельные объекты, позволили в совокупности с методами машинного обучения распознавать группы объектов и отдельные объекты с удовлетворительной точностью (70–80 %).

Однако использование графовых характеристик при распознавании линейных пространственных объектов совместно с морфометрическими не всегда эффективно и требует отдельного поэтапного использования.

Дальнейшее направление исследования темы связано с разработкой методов распознавания типов рисунков объектов одного класса, например, гидрографических сетей: автоматизированное выделение перистого, центробежного, решётчатого и других типов.

СПИСОК ЛИТЕРАТУРЫ

1. *Лурье И.К., Косиков А.Г.* Теория и практика обработки изображений. Дистанционное зондирование и географические информационные системы / Под ред. А.М. Берлянта. М.: Научный мир, 2003.
2. *Шурыгина А.А., Самсонов Т.Е.* Исследование и систематизация морфометрических характеристик основных классов объектов общегеографических карт // Научные исследования молодых ученых-картографов, выполненные под руководством сотрудников кафедры картографии и геоинформатики географического факультета МГУ имени М.В. Ломоносова / Ред. М.В. Зимин, П.Е. Каргашин, М.В. Кусильман и др. КДУ. М., 2017. С. 110–121.
3. *Newman M.E.J.* Mixing patterns in networks // *Physical Review E*. 2003. Т. 67, No 2. S. 026126.
4. *Reingold O.* Undirected connectivity in log-space // *Journal of the ACM (JACM)*. 2008. Т. 55, No 4. P. 17.
5. *Schank T., Wagner D.* Approximating clustering-coefficient and transitivity. Universität Karlsruhe, Fakultät für Informatik, 2004.

REFERENCES

1. *Lurie I.K., Kosikov A.G.* Theory and practice of image processing. Remote sensing and geographic information systems / Edited by A.M. Berlant. M.: Nauchniy mir, 2003 (in Russian).
2. *Newman M.E.J.* Mixing patterns in networks. *Physical Review E*. 2003. Т. 67, No 2.
3. *Reingold O.* Undirected connectivity in log-space. *Journal of the ACM (JACM)*. 2008. Т. 55, No 4. P. 17.
4. *Schank T., Wagner D.* Approximating clustering-coefficient and transitivity. Universität Karlsruhe, Fakultät für Informatik, 2004.
5. *Shurygina A.A., Samsonov T.E.* Analysis and systematization of morphometric characteristics of major classes of geographic maps' objects. Scientific research of young scientists-cartographers performed under the guidance of the staff of the Department of Cartography and Geoinformatics of the Faculty of Geography of Lomonosov Moscow State University / Ed. by M.V. Zimin, P.E. Kargashin, M.V. Kusilman et al. KDU, Moscow, 2017. P. 110–121 (in Russian).