УДК 004.62

ТЕХНОЛОГИЯ DATA MINING В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ РАЗВИТИЯ ТРАНСПОРТНОЙ ИНФРАСТРУКТУРЫ

Федосеев А.А., Михеев С.В., Головнин О.К.

ФГБОУ ВПО «Самарский государственный аэрокосмический университет имени академика С.П. Королёва», Самара, Россия (443086, г. Самара, Московское шоссе, 34), e-mail: fedoseevale@gmail.com

Проведён анализ использования технологии Data Mining для выявления скрытых закономерностей в задачах прогнозирования развития транспортной инфраструктуры. В качестве исходных данных (данных об интенсивности транспортных потоков, а также данных о состоянии статических объектов транспортной инфраструктуры) предложено использовать результаты дистанционного зондирования Земли (ДЗЗ), которое проводится как с помощью космических аппаратов (КА), так и методом аэрофотосъёмки. В свою очередь, исходными данными при оценке состояния статических и динамических объектов являются материалы гиперспектральной съёмки объектов транспортной инфраструктуры, результаты которой представляют собой набор пространственно-спектральных данных, позволяющий проводить анализ спектральных характеристик объектов – сигнатур. Показана возможность реализации задачи классификации, обеспечивающей сегментацию изображений – выделения однородных областей. Отражена возможность применения алгоритмов «мягкой» классификации с использованием нечёткой логики для преодоления сложностей, связанных с отсутствием в достаточной степени учёта характеристик съёмки и особенностями обработки.

Ключевые слова: Data Mining, транспортная инфраструктура, дистанционное зондирование Земли (ДЗЗ), гиперспектральная съёмка.

DATA MINING IN PROBLEMS OF TRASNPORT INFRASTRUCTURE DEVELOPMENT FORECASTING

Fedoseev A.A., Mikheev S.V., Golovnin O.K.

Samara State Aerospace University, Samara, Russia (443086, Samara, street Moskovskoe Shosse, 34), e-mail: fedoseevale@gmail.com

The analysis of Data Mining technology utilization has been done for showing up of undetected rules in problems of transport infrastructure detection. The results of Earth remote sensing (by spacecrafts or by aircrafts) was offered as initial data (transport stream intensity data and stream intensity static features condition data. Furthermore the initial data for static and dynamic features estimation are hyperspectral images of transport infrastructure features. Hyperspectral imagery is a composition of spatial and spectral data which gives a possibility to analyze of spectral characteristics (signatures). The possibility of classification, which provide of imagery segmentation, has been shown. The «soft» classification algorithms with «fuzzy» logic utilization implementing has been shown to meet the complexity conducted with specificity of processing and remote sensing process.

Key words: Data mining, transport infrastructure, Earth remote sensing, hyperspectral imaging.

Наблюдающееся в последнее время интенсивное развитие транспортной инфраструктуры мегаполисов, вызванное увеличением количества транспортных средств, возникшими проблемами с заторами, повышением интенсивности транспортных потоков, привело к ухудшению условий движения, транспортным задержкам, социальному дискомфорту, ухудшению экологической обстановки и перерасходу топлива. Традиционные способы решения этих проблем практически утратили свою эффективность, в связи с чем для оптимизации дорожного движения и повышения его безопасности всё чаще используются концепции интеллектуальных транспортных систем (ИТС) [1].

Для прогнозирования показателей развития транспортной инфраструктуры, повышения безопасности дорожного движения необходимо выявить скрытые закономерности, например, между характеристиками улично-дорожной сети, интенсивностью транспортных потоков, дислокацией технических средств организации дорожного движения, дорожно-транспортными происшествиями. В ИТС анализ данных, направленный на выявление скрытых закономерностей, реализуется с помощью методов интеллектуального анализа данных – Data Mining.

Различные методы технологии Data Mining, такие как нейронные сети, алгоритмы поиска ассоциативных правил, деревья решений, эволюционные алгоритмы и др., нашли широкое применение в решении задач прогнозирования. Однако не все они обеспечивают высокую точность прогнозов, некоторые из них имеют низкую скорость работы, высокую стоимость, что не позволяет широко и в полной мере использовать их для решения указанных задач. Низкая скорость обработки данных препятствует применению существующих аналитических систем управления транспортными потоками в реальном масштабе времени. В связи с этим актуальным является решение проблемы прогнозирования развития транспортной инфраструктуры мегаполисов с использованием интеллектуальных транспортных систем.

Для решения задач прогнозирования, оптимизации организации дорожного движения и управления транспортными потоками резонно использовать процедуру моделирования. В ИТС используются модели улично-дорожной сети, транспортных потоков, технических средств организации дорожного движения, представляемые большими объемами накопленных данных. Использование методов математической статистики для поиска в данных полезной информации не всегда приводит к успеху. Одна из причин этого – концепция усреднения по выборке, приводящая к операциям над фиктивными величинами (например, среднее количество перевозимых пассажиров). В основу технологии Data Mining положена концепция паттернов, помогающих выявлять однотипные фрагменты многоаспектных взаимоотношений в данных. Эти фрагменты представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск паттернов (шаблонов) производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей. Data Mining – это процесс анализа, выделения, представления детализированных данных неявной конструктивной информации, исследования и моделирования больших объёмов данных для обнаружения неизвестных до этого структур (паттернов), новых значимых корреляций и тенденций, полученных в

результате просеивания большого объёма хранимых данных с использованием методик распознавания образов, статистических и математических методов.

Своей сложностью интеллектуальный анализ данных в значительной мере обязан тем же самым трудностям, связанным с организацией данных, которые характерны для любых методик моделирования. Алгоритмы для интеллектуального анализа данных могут быть сложными, однако их применение, благодаря появлению новых программных средств, значительно упростилось [5].

Процесс обнаружения знаний можно разделить на несколько этапов.

1. Понимание и формулировка задачи анализа.
2. Подготовка данных для автоматического анализа.
3. Применение методов Data Mining и построение моделей.
4. Проверка построенных моделей.
5. Интерпретация моделей.

На первом этапе проводится осмысление поставленной задачи, уточняются цели, которые должны быть достигнуты методами Data Mining, определяются способы оценки результатов исследования. Некорректный выбор целей и методов приведёт к искажению окончательных результатов.

На этапе подготовки данных для автоматического анализа происходит форматирование и, если необходимо, редактирование данных, с целью применения определённых методов интеллектуального анализа данных: поиска ошибок и пропусков, приведения данных к одному формату и т.д.

На третьем этапе осуществляется применение выбранных методов Data Mining. Существует множество вариантов их применения: от использования одного-двух алгоритмов до сложных комбинаций из разнообразных методов, что позволяет выполнить всесторонний анализ.

Четвёртый этап – тестирование полученных моделей. Широко используется следующий приём: весь исследуемый массив данных разделяют на две неравные части. Большая из них (обучающая) является исходным материалом для построения моделей методами Data Mining, а меньшая представляет собой тестовую группу, на которой проверяются полученные модели. Критерием, по которому оценивается модель, является разность в точности между группами.

На пятом этапе производится интерпретация построенных моделей человеком. Этот этап не менее важен, чем предыдущие, поскольку некорректная интерпретация полученных результатов сведёт на нет всю проделанную ранее работу. Окончательная оценка моделей может быть дана только после практического применения полученных результатов.

Для выявления скрытых закономерностей методами Data Mining для прогнозирования показателей развития транспортной инфраструктуры необходимо иметь большой объём исходных данных о характеристиках улично-дорожной сети, интенсивности транспортных потоков, дислокации технических средств организации дорожного движения, а также количественные и качественные показатели, связанные с дорожно-транспортными происшествиями.

Рис. 1. Процесс обнаружения знаний с помощью технологии Data Mining.

Получение исходных данных происходит различными способами. Для получения данных об интенсивности транспортных потоков, а также данных о состоянии статических объектов транспортной инфраструктуры используется дистанционное зондирование Земли (ДЗЗ), которое проводится как с помощью космических аппаратов, так и методом аэрофотосъёмки. Каждый метод ДЗЗ имеет свои достоинства и недостатки и применяется в зависимости от типа решаемой задачи. Исходными данными при оценке состояния статических и динамических объектов являются материалы гиперспектральной съёмки объектов транспортной инфраструктуры, результаты которой представляют собой набор

пространственно-спектральных данных, позволяющий проводить анализ спектральных характеристик объектов – сигнатур. Для эффективного извлечения из массива («гиперкуба») гиперспектральной информации скрытых данных (рис. 2), используемых для решения задач определения состояния объектов транспортной инфраструктуры, используется технология Data Mining. В данном случае предполагается реализация задачи классификации, которая обеспечивает сегментацию изображений – выделение однородных областей. Основная идея классификации объекта на гиперспектральном изображении сводится к идентификации его гиперспектральной характеристики (ГСХ). Как правило, ГСХ объекта сравнивается с эталонным ГСХ (паттернами) из спектральной библиотеки по определённому алгоритму. Эталонные значения ГСХ представляют собой набор значений коэффициентов спектральной яркости (КСЯ) в зависимости от длины волны λ. КСЯ представляет собой фотометрическую функцию, характеризующую структуру отражённого поверхностью излучения, как по длинам волн λ, так и по условиям наблюдения и освещения [6]. Если учесть, что все значения ГСХ в спектральной библиотеке и исследуемая ГСХ нормированы (значения КСЯ от 0 до 1, спектральный диапазон библиотеки и исследуемой характеристики совпадают), то простейшим случаем классификации является нахождение меры сходства евклидова расстояния. Евклидово расстояние между двумя точками *i* и *j* на плоскости, когда известны координаты *x* и *y*, определяется по формуле:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (1)$$

В этом случае фактически решается задача одномерной классификации, когда координата *x* жёстко задана и всегда одинакова для исследуемой и эталонных ГСХ. Таким образом, реализована следующая формула нахождения меры сходства [3]:

$$D_i = \sqrt{\sum_{i=1}^{n} (y_i - y_{i\_эталона})^2}, \quad (2)$$

где *i* – номер канала гиперспектрометра, y_i = значение КСЯ исследуемой характеристики для *i*-го канала гиперспектрометра, $y_{i\_эталона}$ – значение КСЯ эталонной характеристики для *i*-го канала гиперспектрометра.

Данная мера сходства обеспечивает хороший результат при идеальных условиях съёмки.

Однако при обработке гиперспектральных изображений ряд характеристик (угол съёмки, влияние атмосферы и т.д.) будет учтён в недостаточной степени. В результате обработки полученная ГСХ объекта может значительно отличаться от реальной, и сравнение с эталоном с помощью алгоритмов «жёсткой» классификации, к которым относятся алгоритмы классификации на основе меры сходства евклидова расстояния, не даст положительных результатов.

Кроме того, исследуемый объект транспортной инфраструктуры может одновременно попадать под разные категории, и алгоритм «жёсткой» классификации однозначно отнесёт

его к заданной категории, хотя это и не совсем верно. Применение алгоритмов «мягкой» классификации с использованием нечёткой логики позволяет преодолеть указанные сложности [4].

Как правило, методы нечёткой классификации данных ДЗЗ широко используются в тех ситуациях, когда исследуемый объект транспортной инфраструктуры изначально является не вполне определённым – не подходит ни под один из известных паттернов.

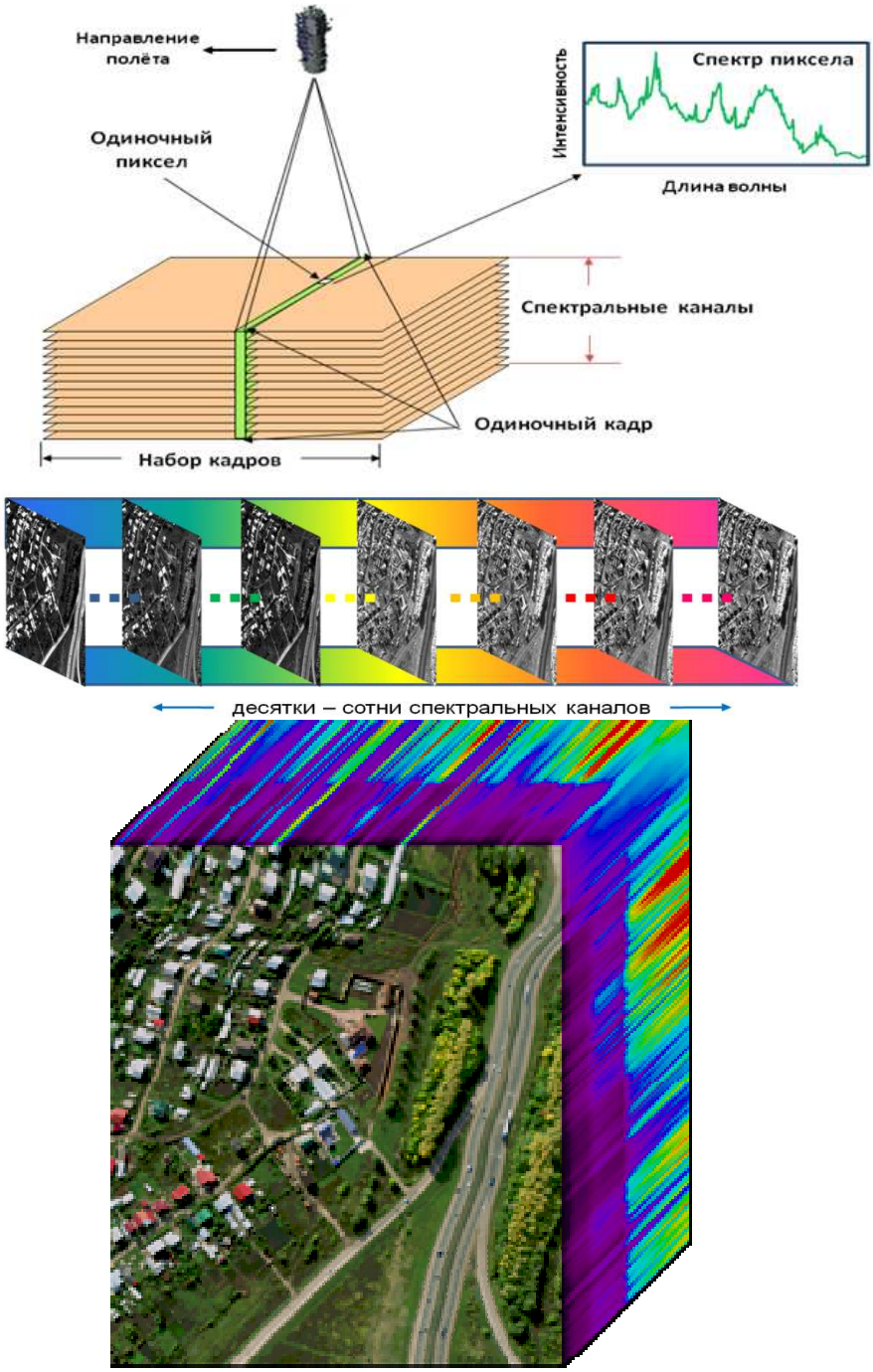

Рис. 2. Процесс формирования «гиперкуба».

В таких условиях, когда входные данные представляют собой нечёткие множества, для решения задачи идентификации объекта, можно использовать алгоритм кластеризации k-средних (fuzzy C Means FCM), предполагающий выполнение следующих шагов [2].

1. Инициализация начального нечёткого разбиения.
2. Вычисление координат центров кластеров.
3. Вычисление новых значений функций принадлежности.
4. Шаги 2 и 3 повторяются до тех пор, пока абсолютная разница значений целевой функции на двух последовательных итерациях не станет менее заранее выбранной погрешности.

С учётом того что количество кластеров (паттернов ГСХ спектральной библиотеки) известно, задача кластеризации сводится к задаче классификации. Искомой величиной, которая рассматривается в качестве меры близости, в данном алгоритме выступает вектор степеней принадлежности исследуемой ГСХ *U* для *J* эталонных ГСХ:

$$U = (u_1, u_2, \ldots, u_j) \tag{3}$$

Степень принадлежности в этом случае рассчитывается по формуле:

$$u_i = \frac{1}{\sum_{i=1,j} \left(\frac{D_i}{D_j}\right)^{\frac{2}{m-1}}}, \tag{4}$$

где D_i – мера сходства для исследуемой ГСХ по отношению к *i*-й эталонной ГСХ, рассчитанная как евклидово расстояние по формуле (2); m – экспоненциальный вес (от единицы до бесконечности).

Экспоненциальный вес характеризует степень «размытости» вектора степеней принадлежности. При $m \to \infty$ получается, что исследуемая ГСХ принадлежит эталонным ГСХ с одной и той же степенью. Практически значение экспоненциального веса устанавливается в пределах от 2 до 5.

Таким образом, использование задач и методов технологии Data Mining открывает новые возможности обработки гиперспектральных данных, используемых в качестве исходной информации для оценки состояния объектов транспортной инфраструктуры, необходимой при решении задачи прогнозирования развития транспортной инфраструктуры.

Список литературы

1. Михеева Т.И., Рудаков И.А. Мониторинг транспортной инфраструктуры // Математика и ее приложения : труды II Междунар. науч. конф. «Математика. Образование. Культура». – Тольятти : ТГУ, 2005. – С. 70-74.

2. Нестеров Н.И., Тишкин Р.В., Юдаков А.А. Кластеризация гиперспектральной информации на основе инструментария нечётких множеств // Труды XVII Всерос. науч.-

техн. конф. студентов, молодых учёных и специалистов «НИТ – 2012». – Рязань : РГРТУ, 2012. – С. 190-191.
3. Пылькин А.Н., Тишкин Р.В. Методы и алгоритмы сегментации изображений. – М. : Горячая линия – Телеком, 2010. – 92 с.
4. Труханов С.В. Классификация объектов на гиперспектральных изображениях в условиях нечётких множеств // Труды XVII Всерос. науч.-техн. конф. студентов, молодых учёных и специалистов «НИТ – 2012». – Рязань : РГРТУ, 2012. – С. 190-191.
5. Чубукова И.А. Data Mining. Основы информационных технологий. Специальные курсы. – М. : Бином. Лаборатория знаний, 2006. – 384 с.
6. Шовенгердт Р.А. Дистанционное зондирование: методы и модели обработки изображений. – М. : Техносфера, 2010. – 560 с.

Рецензенты:

Титов Борис Александрович, доктор технических наук, профессор, заведующий кафедрой организации и управления перевозками на транспорте, ФГБОУ ВПО «Самарский государственный аэрокосмический университет имени академика С.П. Королева (национальный исследовательский университет)», г. Самара.

Хайтбаев Валерий Абдурахманович, доктор экономических наук, профессор кафедры организации и управления перевозками на транспорте, ФГБОУ ВПО «Самарский государственный аэрокосмический университет имени академика С.П. Королева (национальный исследовательский университет)», г. Самара.