4. Alex Belotserkovskiy – Microsoft представляет новый сервис машинного обучения Azure Machine Learning [Электронный ресурс]. – Режим доступа: http://blogs.msdn.com/b/albe/archive/2014/07/15/microsoft-azure-machine-learning.aspx, свободный.

5. Облачные службы – концентраторы событий [Электронный ресурс]. – Режим доступа: http://azure.microsoft.com/ru-ru/services/event-hubs/, свободный.

6. Azure Newsletter [Электронный ресурс]. – Режим доступа: http://ibmpw.blogspot.ru/2015/03/azure-newsletter-2015.html, свободный.

УДК 004

КЛАСТЕРИЗАЦИЯ СПУТНИКОВЫХ СНИМКОВ С ИСПОЛЬЗОВАНИЕМ ОБЛАЧНЫХ ТЕХНОЛОГИЙ MACHINE LEARNING AZURE

Е.Е. Васильева
Научный руководитель: Н.Г. Марков, д.т.н., профессор, зав.каф. ВТ ИК ТПУ
Национальный исследовательский Томский политехнический университет
E-mail: ekaterina.vasilyeva9@gmail.com

This article deals with a modern cloud computing technology – Machine Learning Azure. The technology gives a variety of tools to design the systems which implement popular machine learning algorithms and approaches such as anomaly detection, classification, clustering and regression. It can be used for data analysis in different fields but article shows how it can be used for satellite image processing and recognition.

Keywords: cloud computing, clustering, satellite image recognition, Machine Learning Azure.
Ключевые слова: облачные технологии, кластеризация, распознавание спутниковых снимков, Machine Learning Azure.

Введение

В связи с развитием методов машинного обучения и компьютеризацией, все большее количество задач становится возможным решать при их помощи. В связи с этим Microsoft представили в 2014 г. новый облачный сервис – Machine Learning Azure [1], позволяющий пользователям по мере необходимости быстро анализировать данные. Это графическая среда, в которой некоторые инструменты и алгоритмы машинного обучения представлены в виде блоков. Для реализации собственных алгоритмов используются блоки, компилирующие код на языке R или Python. Создавая определенную схему из блоков с корректно настроенными параметрами, можно получить модели классификации, кластеризации, линейной регрессии и статистического анализа. Далее рассмотрена кластеризация спутникового снимка при помощи Machine Learning Azure.

Обработка данных

В ходе исследования использовался набор данных Urban Land Cover Dataset [2], который состоит из текстур, описываемых множеством признаков (яркость, индекс формы, NDVI (вегетационный индекс), средние значения в красном и зеленом каналах и др.), полученных из космоснимка высокого разрешения. Эти данные нуждаются в предварительной обработке. К мерам обработки относятся: удаление дубликатов, восстановление отсутствующих значений, нормализация.

Нормализация данных необходима, когда параметры измеряются в разных шкалах. Рассматриваемый набор данных был нормализирован по Z-score:

$$z = \frac{x - mean(x)}{stdev(x)},$$

где x – фактическое значение параметра, mean(x) – среднее значение параметра на всем наборе данных, stdev(x) – стандартное отклонение.

Модель кластеризации

Для сравнительного анализа в модели были реализованы два алгоритма кластеризации: K-Means и иерархическая кластеризация [3] (рис. 1).

Данная модель получает исходный набор данных training.csv, все числовые параметры (Project Columns отфильтровывает параметры) которого нормализуются в блоке Normalize Data. В блоке K-Means Clustering задаются параметры модели кластеризации (количество кластеров, мера расстояния, принцип начальной инициализации центров кластеров, ограничение числа итераций). Нормализованные данные разбиваются на кластеры, согласно выбранному алгоритму кластеризации в блоке Train Clustering Model. Блок Execute R Script может содержать произвольный код на языке R для обработки и табличного и/или графического вывода данных. В данном блоке реализован алгоритм иерархической кластеризации, так как Machine Learning Azure содержит готовый блок только для K-Means алгоритма кластеризации. Полученные кластеры сравниваются с реальными значениями классов.

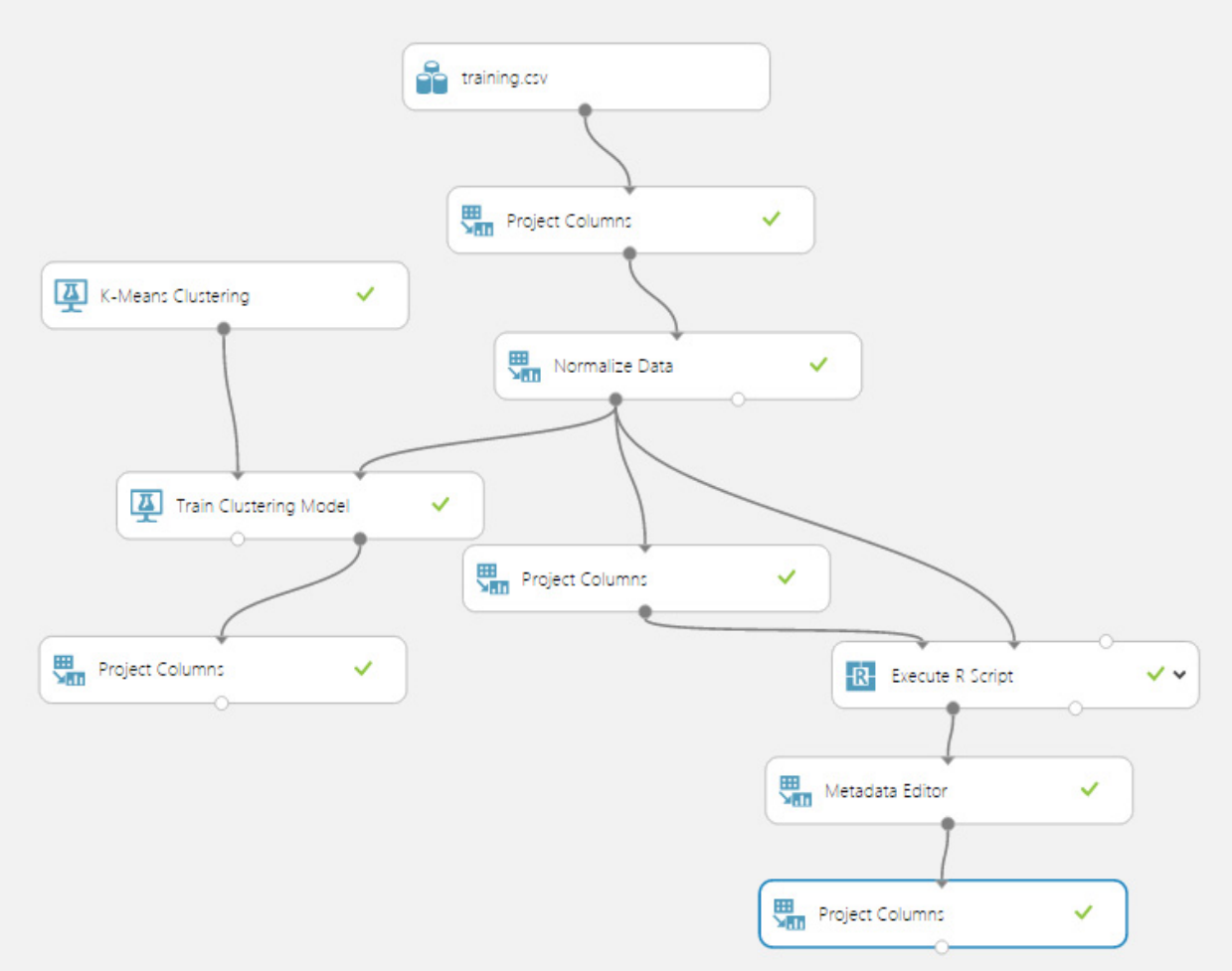

Рис. 1. Модель кластеризации в Machine Learning Azure

Результаты кластеризации

Кластеризация снимков проводилась при помощи двух алгоритмов – иерархического и K-Means. В ходе эксперимента были использованы различные настройки параметров данных алгоритмов, и были получены следующие лучшие результаты кластеризации (рис. 2). В случае идеального распознавания, объекты были бы отнесены только к одному классу – одно значение на строку. При кластеризации K-Means (косинусное расстояние, инициализация центров K-Means++ Fast, максимум 500 итераций) были выделены 3 четко различимых класса, два класса: трава и деревья – были отнесены к одному. Иерархическая кластеризация (метод Варда, Евклидово расстояние) показала следующий результат: 3 четко различимых класса, смешение классов трава и деревья, асфальт и тени.

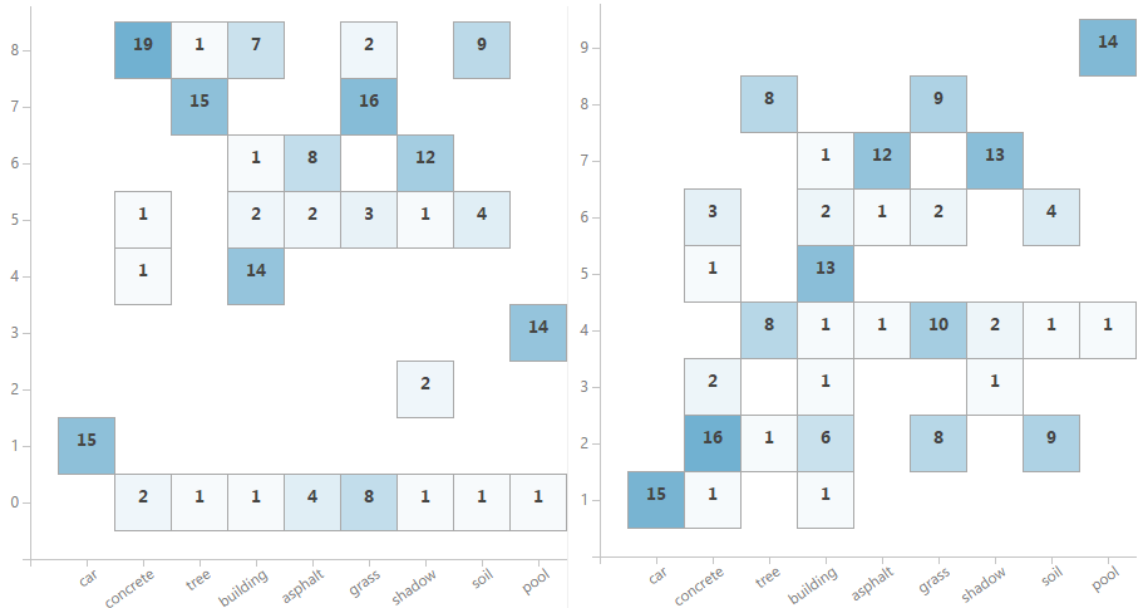

Рис. 2. Результаты кластеризации
(слева – K-Means, справа – иерархическая)

Заключение

Machine Learning Azure – гибкий инструмент для обработки и интерпретации данных, который позволяет упростить процесс освоения среды для начинающих, реализует наиболее распространенные методы и алгоритмы машинного обучения и имеет возможность расширения. Полученные в результате анализа спутникового снимка данные говорят о том, что используемые алгоритмы кластеризации плохо справляются с задачей распознавания, что, в свою очередь, говорит о необходимости развития алгоритмов кластеризации.

Список литературы

1. Microsoft Azure Machine Learning [Электронный ресурс] / Официальный сайт Microsoft Azure Machine Learning. – Электрон. дан. – 2015. – Режим доступа: https://studio.azureml.net/, свободный. – Загл. с экрана. – Яз. англ. (Дата обращения 10.03.2015).

2. Urban Land Cover Data Set [Электронный ресурс] / UC Irvine Machine Learning Repository. – Электрон. дан. – 2015. – Режим доступа: https://archive.ics.uci.edu/ml/datasets/Urban+Land+Cover, свободный. – Загл. с экрана. – Яз. англ. (Дата обращения 03.03.2015).

3. Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. – 412 с.