гроз и решения практически важных задач грозозащиты технических сооружений и грозовой пожарной опасности лесных массивов на территориях, где отсутствуют инструментальные средства наблюдений за грозами.

Преимуществом использования алгоритма DBSCAN (Density Based Spatiustering of Applications with Noise) для кластеризации данных WWLLN является «естественный» учет особенностей данных о грозовых разрядах, получаемых этой сетью (пространственный характер данных, наличие более плотных «сгустков» объектов и одиночных разрядов («шум»), отстоящих на некотором расстоянии от «сгустков») [1].

При использовании алгоритма DBSCAN необходим выбор входных параметров (ε – минимальное расстояние между разрядами в километрах, ε_{time} – минимальная разница по времени в минутах, $MinPts$ – минимальное количество точек в кластере) таким образом, чтобы результаты кластеризации были сопоставимы с параметрами грозовой активности (например, средняя продолжительность гроз и средняя площадь грозового облака и/или грозовой ячейки в нем). При этом значение ε_{time} должно быть меньше средней продолжительности грозы [2].

С целью согласования результатов кластеризации данных WWLLN с региональной средней продолжительностью гроз был проведен вычислительный эксперимент для данных о грозовых разрядах, зарегистрированных на территории Республики Алтай (регион в градусах) за летний период 2013 года. Были выбраны следующие наборы входных параметров: $20 \le \varepsilon \le 50$, $10 \le \varepsilon_{time} \le 120$, $2 \le MinPts \le 5$ [1], при этом максимальное значение ε_{time} выбрано равным средней продолжительности гроз по выбранному региону за летний период 2013 года. Для каждого набора параметров вычислялась средняя продолжительность по кластерам.

Результаты эксперимента показали, что средняя продолжительность по кластерам согласуется с региональной средней продолжительностью гроз для следующих наборов параметров алгоритма DBSCAN: при значениях $\varepsilon = 45{,}50$, $MinPts = 2$, $\varepsilon_{time} = 105{,}120$ средняя продолжительность по кластерам изменяется в пределах от 110 до 130 минут; при значениях $\varepsilon = 50$, $MinPts = 5$, $\varepsilon_{time} = 40{,}45$ средняя продолжительность по кластерам изменяется в пределах от 100 до 120 минут. Так как в первом случае разница между параметром ε_{time} и средняя продолжительность по кластерам незначительная, то наиболее приемлемыми параметрами используемого алгоритма кластеризации будем считать второй набор параметров.

В дальнейшем планируется проведение дополнительных вычислительных экспериментов и привлечения формальных способов оценки результатов кластеризации.

Библиографический список

1. Беликова М.Ю., Кречетова С.Ю., Перелыгин А.А. Методы и результаты кластеризации данных по грозовым разрядам // Известия Алтайского государственного университета. – Барнаул, 2016. – №1 (89). – С. 97–100.

2. Hutchins, Michael L., Robert H. Holzworth, and James B. Brundell, Diurnal variation of the global electric circuit from clustered thunderstorms, Journal of Geophysical Research: Space Physics 119 (1), 620-629, DOI 10.1002/2013JA019593, Jan 2014 ; [Электронный ресурс]. – URL: http://www.wwlln.net/publications/hutchins.early.view.jgra50799.pdf (дата обращения 20.05.2015).

УДК 519.688

Разработка web-сервиса для диагностики рака молочной железы с помощью Microsoft Azure Machine Learning

А.Ф. Лазарев[1], М.А. Рязанов[2], К.А. Хрулёв[2], А.И. Шаповал[3]

[1]Алтайский краевой онкологический диспансер, г. Барнаул;

[2]АлтГУ, г. Барнаул; [3]РАПРЦ, г. Барнаул

В настоящее время машинное обучение применяется во многих областях науки и производства. Медицина не является исключением. С помощью машинного обучения решается множество таких задач, как классификация больных по видам заболеваний, определение наиболее целесообразного способа лечения, предсказание длительности и исхода заболевания, оценка риска осложнения, нахождение синдромов, наиболее характерных для определённого вида заболевания и т.п.

За годы работы Алтайского краевого онкологического диспансера «Надежда» были накоплены данные по пациентам, проходившим обследования на выявление рака молочной железы. Рак молоч-

ной железы – это заболевание, вызванное перерождением нормальных клеток железистой ткани в раковые. В мире это наиболее частая форма рака среди женщин, поражающая в течение жизни от 1/13 до 1/9 женщин в возрасте от 13 до 90 лет.

В связи с этим диагностика данного вида заболевания и выявление его на ранней стадии является актуальной задачей. Для ее решения были проанализированы данные, полученные от Алтайского краевого онкологического диспансера «Надежда», построена модель машинного обучения с помощью Microsoft Azure Machine Learning и разработан сайт для взаимодействия сотрудников онкологического диспансера с построенной моделью.

Полученные данные представляют собой признаковые описания пациентов, включающие в себя физиологические, психологические и социологические признаки. Всего 74 признака по каждому пациенту, из них 17 бинарных, 20 номинальных и 37 порядковых.

Для предобработки и анализа данных и построения модели бинарной классификации пациентов была использована облачная платформа Microsoft Azure Machine Learning.

Первым этапом решения поставленной задачи являлась предобработка полученных данных, включающая в себя:

– устранение противоречивости информации путем вычисления вероятности появления каждого из противоречивых событий и выбора наиболее вероятного;

– выявление и обработка выбросов и неинформативных объектов с помощью алгоритма STOLP.

Следующим этапом построения модели машинного обучения являлся анализ данных, который включил в себя корреляционный, регрессионный и факторный анализы, что позволило выявить зависимости между признаками и снизить размерность данных [1].

После этого были использованы следующие алгоритмы машинного обучения для построения модели бинарной классификации пациентов:

– Decision Forest;

– Logistic Regression;

– Boosted Decision Tree.

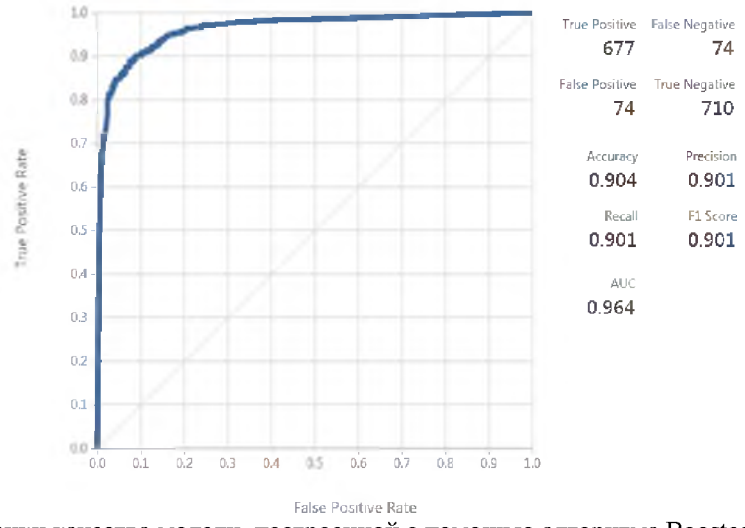

Рисунок – Оценки качества модели, построенной с помощью алгоритма Boosted Decision Tree

Настройка оптимальных параметров была произведена с помощью скользящего контроля, что позволило достичь максимального качества для каждой модели. Сравнив качественные показатели, была выбрана модель, построенная на основе алгоритма Boosted Decision Tree. Она показала максимальные значения чувствительности, специфичности, точности и AUC по отношению к остальным алгоритмам.

После построения модели бинарной классификации пациентов был разработан web-сайт для удобства работы сотрудников онкологического диспансера, который включает в себя следующий функционал:

– просмотр результатов корреляционного, регрессионного и факторного анализов;

– загрузка новых данных для анализа и прогнозирования диагноза.

Разработанный web-сервис может быть непосредственно использован в качестве системы поддержки принятия решений для сотрудников онкологического диспансера с целью диагностики рака молочной железы.

Библиографический список

1. Юнкеров В.И., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований. – СПб.: ВМедА, 2002. – 266 с.

УДК 581.6

Применение спутниковых данных для картирования растительного покрова Третьяковского района (Алтайский край)

И.С. Минина, Н.В. Овчарова
АлтГУ, г. Барнаул

В настоящее время прослеживается интенсивное развитие и использование спутниковых систем дистанционного зондирования Земли для решения большого разнообразия научных и практических задач. К ним можно отнести исследования характера климата и биосферы, выявление и распознавание на аэроснимках объектов местности, определение их качественных и количественных характеристик и др. (Кравцова, 2005).

Растительный покров, являясь основным компонентом биосферы, а также важным возобновляемым ресурсом, имеет огромное значение, как экономическое, так и экологическое. Системы дистанционного зондирования, обладая возможностью регулярных измерений характеристик покрова земли, дают информацию о пространственном расположении и изменение растительности (Лабутина, 2004).

Исследование направлено на оценку состояния растительного покрова с использованием наземных и дистанционных данных, а также построение среднемасштабной карты растительности на примере Третьяковского района.

Для достижение цели решены следующие задачи: раскрыть понятие дешифрирование и значимость спутниковых данных при изучении растительного покрова; провести анализ методов спутниковых измерений характеристик растительного покрова; расчёт вегетационного индекса растительного покрова; по данным дешифрирования подготовить предварительную классификацию типов фитоценозов на исследуемую территорию; на основе данных спутниковых измерений провести картирование растительного покрова Третьяковского района для оценки и мониторинга растительности на обширных территориях.

Район исследования расположен в южной части Алтайского края. В современных границах занимает площадь 2033, 23 км^2 (около 1,2% площади Алтайского края). Граничит на юге и юго-востоке с Восточно-Казахстанской областью Казахстана, на северо-востоке и севере со Змеиногорским районом, на западе с Локтевским районом. Рельеф района неоднородный, имеет преимущественно равнинный характер, постепенно переходящий в предгорье. Самая длинная река края – Алей длиной 866 км, река Каменка, река Корболиха, а также река Глубокая. Климат резко континентальный с холодной малоснежной зимой и жарким сухим летом. Максимальная температура января +6 град., минимальная –49 град., средняя температура воздуха в зимний период от –12 до –16 град. Почвы обыкновенные черноземы, выщелоченные типичные черноземы, горные серые лесные.

В работе были использованы данные со спутников Landsat 8 ETM+. Территория исследований покрывалась несколькими сценами съемки, что потребовало составления мозаики снимков (2013–2015 гг) (рисунок 1).

Из базы снимков были отобраны безоблачные и малооблачные сцены. Так как снимки в мозаике имеют разные даты и условия съемки, для получения бесшовных мозаик проведено гистограммное выравнивание яркостей для каждого спектрального канала съемки.

При классификации растительности применен эколого-фитоценотический подход, основанный на соотношении основных биоморф и участия эколого-ценотических групп видов в составе и структуре сообщества.

Аналоговая пространственная информация (топокарты, карты гидросети, почвенного и растительного покрова, данные лесной таксации на уровне лесничеств) после сканирования вносилась в ГИС посредством привязки и дальнейшей векторизации в программных продуктах ESRI ArcGIS 10.2 for Desktop Advanced (ArcInfo) Lab Pak.

Для реализации процедуры классификации был рассчитан индекс NDVI по формуле:

$$NDVI = (NIR - RED) / (NIR + RED),$$

где NIR – отражение в ближней инфракрасной области спектра, RED – отражение в красной области спектра.