Применение платформы Microsoft Azure HDInsight для обработки и анализа больших массивов астрономических данных

С.В.Герасимов, А.В.Мещеряков

Аннотация— **Методы машинного обучения позволяют прогнозировать красное смещение галактик и строить объемные карты распределения астрономических объектов во Вселенной, которые в настоящее время широко используются в фундаментальных задачах внегалактической астрономии и наблюдательной космологии. Однако применение построенных прогностических моделей к доступным астрономическим каталогам, содержащим данные широкополосной фотометрии для объектов на всем небе, требует значительных вычислительных ресурсов. В данной статье был обобщен успешный опыт по применению технологии Apache Spark в облачной инфраструктуре Microsoft Azure для решения задачи точного прогнозирования фотометрических красных смещений галактик на большом массиве данных астрономических каталогов из небесного обзора SDSS.**

Ключевые слова— **машинное обучение, большие данные, прогнозирование, каталоги небесных объектов, красные смещения, случайный лес деревьев решений, MapReduce, Apache Spark, Microsoft Azure**

I. Введение

Основным трендом современной наблюдательной астрофизики последней четверти века стал непрерывный экспоненциальный рост объема доступных данных наблюдений, представляющих собой изображения небесных объектов в различных фильтрах (т.н. фотометрические наблюдения в широких полосах частотного спектра), которые получены телескопами в рамках цифровых обзоров неба. Проблематика хранения, обработки и анализа быстро растущих объемов данных астрономических наблюдений (в виде сырых изображений и полученных на их основе астрономических каталогов) из архивов небесных обзоров [1] схожа с подобными задачами, возникающими в настоящее время во всех областях жизнедеятельности человека (например, в Интернет[2], финансах[3], медицине[4], биоинформатике[5]) и обозначаемых общим направлением Большие данные (англ., Big Data). Необходимость эффективного анализа Больших данных, непрерывно накапливаемых человечеством в разных сферах деятельности, спровоцировало появление модели горизонтально-масштабируемых вычислений MapReduce [6], распределенных файловых систем, реализацию на их базе таких технологий как Apache Hadoop[7] и Spark[8], появление облачных инфраструктур Amazon EMR и Microsoft Azure HDInsight, пригодных для пакетной обработки больших данных.

Главными целями современных (например, SDSS[9], PanSTARRS[10]) и будущих (LSST[11], EUCLID[12], Спектр-Рентген-Гамма[13]) цифровых небесных обзоров является проверка космологических теорий возникновения и эволюции Вселенной, исследование природы Темной материи и Темной энергии и исследование других вопросов, лежащих на переднем крае современной фундаментальной физики, на основе объемных карт крупномасштабного распределения вещества (англ., LSS - Large Scale Structure) во Вселенной. Построение подобных карт предполагает точную классификацию и измерение расстояния (т.н. красного смещения) до всех астрономических объектов, наблюдаемых в небесном обзоре. Необходимо отметить, что внешний вид и яркость объекта на отдельном изображении слабо коррелирует с физическим классом и расстоянием до объекта. Поэтому для классификации и измерения красного смещения далекой галактики астрономы стремятся выполнить ее спектральные наблюдения с высоким частотным разрешением, на которых будут различимы детали (спектральные линии) позволяющие точно классифицировать объект и измерить красное смещение. Однако, получение детальных спектральных данных для каждого объекта низкой яркости (а таких галактик - подавляющее большинство в небесном обзоре) требует значительных затрат наблюдательного времени телескопа. Необходимая информация о физических характеристиках, красном смещении и классе для всех астрономических объектов на небе (в общей сложности $>10^{10}$ объектов зарегистрированы на изображениях, полученных современными цифровыми обзорами) не может быть получена напрямую из спектральных наблюдений. Таким образом, возникающие в

Статья получена 3 декабря 2016. Работа поддержана Российским фондом фундаментальных исследований (грант РФФИ №15-29-07085 офи_м и №14-22-03111 офи_м). Авторы благодарят компанию Microsoft за облачные ресурсы, предоставленные коллективу в рамках программы “Microsoft Azure for Research”. А.В. Мещеряков также благодарен финансированию за счет средств субсидии в рамках государственной программы повышения конкурентоспособности Казанского (Приволжского) федерального университета.

Герасимов Сергей Валерьевич, инженер Факультета Вычислительной Математики и Кибернетики МГУ им М.В.Ломоносова gerasimov@mlab.cs.msu.su

Мещеряков Александр Валерьевич, к.ф.-м.н., научный сотрудник Института космических исследований РАН и Лаборатории Рентгеновская астрономия Казанского федерального университета mesch@iki.rssi.ru

астрофизике задачи массовой классификации и регрессии необходимо решать на доступных данных из фотометрических каталогов - таблиц, содержащих “измеренные” на снимках значения сотен признаков небесных объектов. Фотометрические каталоги могут быть сопоставлены между собой (например, на основе небесных координат [14]), и такой консолидированный астрономический каталог уже сейчас содержит для каждого небесного объекта $>10^3$ доступных фотометрических признаков. С появлением новых фотометрических каталогов количество признаков небесных объектов будет непрерывно расти. Следует отметить, что фотометрические атрибуты объектов в астрономических каталогах, как правило, сильно коррелированы между собой и содержат большое число пропущенных значений.

Для прогнозирования физических характеристик астрономических объектов на основе фотометрических данных небесных обзоров, в последние годы широкое применение находят алгоритмы машинного обучения с учителем (см. обзор задач в [15]). Наибольшего успеха исследователи добились в прогнозировании фотометрических красных смещений (photo-z) до внегалактических объектов (см. например, [16][17], классификации точечных/протяженных объектов (т.н. классификация звезда/галактика [18]), а также, классификации астрономических объектов по более специализированным типам, например классификации квазаров [19], сверхновых Ia [20], переменных звезд типа RR Лиры [21]. Так в задаче классификации звезда/галактика алгоритмы машинного обучения являются в настоящее время стандартным средством для всех астрономических программ детектирования объектов на изображениях (см. например SExtractor[22]). Для популярной задачи, которая будет подробно рассмотрена в настоящей работе - измерение фотометрических красных смещений галактик - методы машинного обучения показали непревзойденную точность по сравнению с другими подходами [23]. Следует отметить, что при наличии тренировочной (спектральной) выборки достаточного большого размера методы машинного обучения обеспечивают лучшую точность прогнозирования фотометрических красных смещений объектов, чем другие подходы, опирающиеся на физическое моделирование наблюдаемых атрибутов небесных объектов в астрономическом каталоге.

В настоящей работе исследуется проблема применения высокоточных моделей машинного обучения, обеспечивающих аккуратное прогнозирование красных смещений галактик на больших данных астрономических каталогов на основе современных и будущих цифровых обзоров неба. Авторы предлагают решение поставленной задачи прогнозирования photo-z для больших выборок галактик, которое основано на применении фреймворка Apache Spark в облачной инфраструктуре Microsoft Azure.

Статья организована следующим образом. В следующей части рассматривается постановка задачи точного прогноза красных смещений галактик по данным из современных астрономических каталогов (на примере данных из обзора неба SDSS). Далее приводится обзор существующих подходов к применению моделей машинного обучения к большим выборкам объектов и описывается предлагаемое нами решение поставленной задачи. В последней части приведены наши выводы.

II. Постановка задачи: точный прогноз photo-z галактик по данным каталога SDSS

Для исследования проблемы применения высокоточных моделей машинного обучения для прогнозирования фотометрических красных смещений объектов по фотометрическим данным из астрономических каталогов, нами был выбран популярный Слоановский цифровой обзор неба (англ. SDSS - Sloan Digital Sky Survey) [9], содержащий как обширный фотометрический каталог, так и наиболее полные спектроскопические данные, которые могут использоваться для обучения прогностической модели (см. далее). В работе использовались данные из 12 релиза (SDSS DR12).

Для прогнозирования красных смещений галактик нами был применен глубокий (без ограничений по глубине) экстремально случайный лес деревьев решений (англ., ERFDT - Extremely Randomized Forest of Decision Trees) [24]. Алгоритм продемонстрировал высокую точность прогнозирования photo-z, для близких объектов превосходящий точность (см. [25]) достигнутую другими алгоритмами, применяемыми для задачи прогнозирования photo-z галактик, например, с нейронными сетями [26]. Главной особенностью данного метода, которая обеспечивает высокую конечную точность прогнозирования, является использование всей полноты фотометрических атрибутов доступных в астрономическом каталоге при построении регрессионной модели. Применительно к фотометрическому каталогу SDSS, на котором проводились эксперименты в настоящей работе, прогностическая модель строится на 581 атрибуте, часть из которых - оригинальные атрибуты каталога SDSS, другие - экспертно составленные линейные комбинации.

Следующий важный этап при построении высокоточного метода машинного обучения состоит в использовании всей доступной тренировочной выборки на этапе обучения модели. Мы использовали тренировочную выборку галактик с хорошим качеством фотометрических данных (про отбор объектов тренировочной выборки подробнее см. [25]) из каталога SDSS DR12, размер используемой тренировочной выборки составил 1,1 млн. объектов. Обучающие данные - это та небольшая (~0,5% из 200 млн.) часть галактик (протяженных объектов на изображениях) в обзоре SDSS, которые содержат целевой атрибут - красное смещение, измеренное напрямую по спектру объекта.

Обучение случайного леса производилось на тренировочной выборке галактик с помощью реализации алгоритма из библиотеки scikit-learn [27]. Процесс построения модели состоял из следующих шагов:

- чтение данных из csv-файла, содержащего

галактики каталога SDSS;

- отфильтровывание нужных данных;
- формирование экспертных атрибутов;
- построение леса деревьев решений.

Для применения результатов прогнозирования в научных исследованиях астрофизикам необходим механизм, позволяющий отбирать только высокодостоверные прогнозы. Для получения оценки достоверности прогноза в [28] было предложено использовать идею квантильного леса регрессии (англ., quantile regression forest) [29] - для каждого объекта оценивать функцию распределения значений прогноза индивидуальных деревьев решений, входящих в ансамбль: скученность значений означает достоверный прогноз, большой разброс - сомнительный.

Целевой выборкой для применения модели прогнозирования красных смещений, в рамках данной работы, являлась фотометрическая подвыборка каталога SDSS, содержащая около 70 млн. объектов. Размеры других выборок, предназначенных для массового прогнозирования красных смещений, могут достигать миллиарда объектов. Временные показатели обработки таких объемов данных могут быть легко определены на основе статистики обработки 70 млн. объектов ввиду линейного роста времени обработки.

Каталог, состоящий из 70 млн. объектов, был сохранен в формате csv и занимал более 300 ГБ. Предварительные эксперименты, проведенные на отдельной рабочей станции (A6 - 4 ядра, 28 ГБ ОЗУ), продемонстрировали, что чтение 1 млн. строк занимает около 4 мин 45 секунд. и еще около 20 мин преобразование данных и прогнозирование. Уже для 70 млн. строк целевой выборки применение модели заняло бы на одном компьютере более суток. Для характерного размера целевой выборки составляющей миллиард астрономических объектов, как требуемый объем хранилища данных, так и время применения модели на одном узле возрастет в 10-15 раз. Также необходимо иметь в виду часто возникающую необходимость пересчета прогноза при применении новых моделей. Все это делает необходимым использование вычислительного кластера.

Для анализа результатов прогнозирования исследователю-астрофизику по каждому объекту выборки могут потребоваться входные атрибуты и прогнозы для выборки деревьев входящих в лес (мы выводили прогнозы 100 деревьев для конструирования оценки достоверности). При работе с такого рода данными исследователям часто используют операции фильтрации по значению атрибутов, группировки и последующей агрегации, также необходимо предоставить возможность соединения набора с самим собой или другими данными с числом объектов в диапазоне 10^4-10^9 (размер доступных астрономических каталогов). Следует также отметить, что в задаче может фигурировать несколько вариантов целевых выборок, несколько моделей прогнозирования с отличающимися гиперпараметрами.

Нам требовалась технология, которая позволит за разумное время (минуты/часы в зависимости от размера выборок) осуществлять высокоточные прогнозы photo-z на больших целевых выборках объектов из астрономических каталогов и предоставит возможность астрофизикам-исследователям интерактивно анализировать результаты построенных прогнозов.

III. Обзор существующих решений

Для решения поставленной задачи нам была необходима программная библиотека, предоставляющая возможность применения готовых случайных лесов деревьев к выборке данных. В первую очередь нас интересовали возможности библиотек по распределенному применению моделей и библиотек, основанных на scikit-learn. Apache Spark MLLib[30] предоставляет распределенный алгоритм построения случайного леса деревьев решений. Ниже мы рассмотрим эти две библиотеки и производные проекты (spark-sklearn и sparkit-learn) более подробно.

Библиотека scikit-learn включает в себя многонитиевую реализацию случайного леса деревьев решений на языке Си. Для обучения алгоритма необходимо, чтобы тренировочная выборка помещалась в памяти (хотя настройка, позволяющая добавлять новые деревья в существующий лес, позволяет обучать алгоритм на фрагментах выборки, но этот режим не эквивалентен обучению на всей выборке за раз). В многонитевом стиле реализованы как обучение, так и применение модели. Применение модели в scikit-learn на больших выборках можно осуществлять путем загрузки выборок в память и прогнозирования по частям.

Apache Spark MLLib[30] предоставляет распределенный алгоритм построения случайного леса деревьев решений. Алгоритм основывается на идее распределенного построения отдельного дерева решений. Значения атрибутов разбиваются на “бины”, число “бинов” является параметром алгоритма. Для формирования “бинов” используются процентили, таким образом, “бины” заданного атрибута имеют одинаковую частоту встречаемости в данных. В драйвере поддерживается очередь вершин дерева, требующих разбиения. Изначально в очередь помещается корневая вершина. Драйвер в цикле изымает из очереди разом несколько (количество определяется настройками памяти) узлов для разбиения. Рабочие (англ., workers) вычисляют в “бинах” статистику по целевой переменной, необходимую для вычисления значения функции неоднородности (англ., impurity function). Каждый рабочий рассчитывает статистику в рамках своих локальных данных. Статистика отсылается с помощью “тасовки” на один из рабочих, который выбирает наилучший атрибут и точку разбиения, после чего отправляет статистику о выбранном разбиении в драйвер. Драйвер обновляет структуру дерева и добавляет в очередь новые вершины. Данный алгоритм использует загрубление данных с помощью “бинирования” и эффективен для построения неглубоких деревьев. При применении модели лес копируется на каждый из воркеров и применяется для локальных данных. Мы не стали переходить на реализацию в Apache Spark MLLib поскольку, во-первых, спектральные части каталогов небесных

обзоров - это несколько миллионов объектов, данные которых умещаются в оперативную память современных компьютеров, во-вторых, производительность распределенной реализации леса деревьев решений в MLlib сильно зависит от степени загрубления данных с помощью дискретизации, а для нашей задачи критичной является точность прогнозирования.

Библиотека spark-sklearn[31] решает задачу распределенного подбора на Spark гиперпараметров алгоритмов в ходе кросс-валидации. При этом используются оригинальные нераспределенные реализации алгоритмов из scikit-learn.

Библиотека sparkit-learn[32] обеспечивает бо̀льшую интеграцию между scikit-learn и PySpark. Библиотека предоставляет возможность осуществлять на базе Spark распределенную предобработку данных, распределенное обучение некоторых алгоритмов scikit-learn (тех из них, для которых подходит идея обучения нескольких моделей на частях данных и объединения их прогнозов при применении, в т.ч. случайного леса), распределенное применение и подбор гиперпараметров.

Следует отметить, что библиотеки spark-sklearn и sparkit-learn находятся на ранних стадиях разработки, поэтому в дальнейшем они были исключены из рассмотрения.

При выборе собственного решения мы учитывали специфику поставленной задачи: необходимость построения максимально точных прогнозов, небольшой объем доступных в астрономии тренировочных спектральных выборок (которые могут помещаться в память отдельного узла кластера). Опираясь на представленный выше обзор библиотек, мы приняли решение обеспечить распределенное применение случайных лесов деревьев решений за счет “простого” копирования уже обученных моделей scikit-learn на отдельные вычислительные узлы. Реализация данного подхода на Apache Spark подробно рассматривается ниже.

IV. Предложенное решение

Для решения поставленной задачи, прогнозирования photo-z галактик на больших данных астрономических каталогов мы решили применить готовый инструментарий аналитики больших данных (фреймворк Apache Spark) в облачной инфраструктуре Microsoft Azure, поскольку это, на сегодняшний день, один из самых легко доступных, удобных и недорогих способов развернуть вычислительный кластер с современным стеком технологий аналитики больших данных. Высокоскоростной канал между облаком Microsoft Azure и астрономическим архивом SDSS, позволяет быстро перенести данные фотометрического каталога SDSS (~3ТБ) в развернутый кластер. Из скачанного каталога SDSS была отфильтрована целевая выборка 70млн. галактик для применения моделей прогнозирования photo-z.

Для решения вопроса прогнозирования на целевой выборке большого объема на PySpark был реализован алгоритм распределенного прогнозирования с помощью обученной модели scikit-learn. Идея алгоритма - клонирование модели на каждого исполнителя (англ., executor) и применение модели - загрузка данных в оперативную память, формирование экспертных атрибутов, прогнозирование - по разделам (англ., partition), составляющих исходный RDD (англ., Resilient Distributed Dataset). Клонирование модели реализовано с помощью широковещательных переменных (broadcast variables), размер леса из 100 деревьев составил около 3.1 ГБ. Для применения описанных действий над RDD по разделам использовалась функция RDD.mapPartitions.

Более 70 миллионов строк входных файлов были разбиты на около 7,4 тыс. разделов каждый, примерно по 9,5 тыс. строк. Граф вычислений (рис.1) содержал

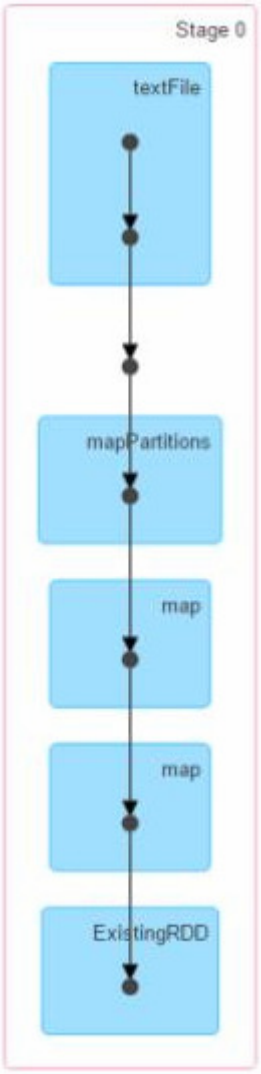

Рис. 1 Граф вычислений

всего одну стадию (англ., stage), состоящую из загрузки файла из хранилища, применения логики обработки для каждого раздела (функция, передаваемая в mapPartitions), нескольких вспомогательных вызовов map и сохранения результата (листинг 1). 7,4 тыс. задач обработаны 12ю исполнителями за 2 часа и 1 мин. Ускорение оказалось даже более 12 раз по сравнению со временем обработки на одном компьютере. Это может быть объяснено более высокой производительностью процессоров Xeon, использованных в кластере в узлах серии D v2, по сравнению процессором с серии A, использованном в отдельном компьютере). Для

вычислений использовалась следующая конфигурация кластера Azure HDInsight:

Операционная система: Linux

Версия Apache Spark: 2.0.0

12 рабочих узлов D12 V2 (4 ядра, 28 ГБ ОЗУ, внешняя память 200 ГБ SSD)

Настройки исполнителей:

- число исполнителей (конфигурационное свойство spark.executor.instances) - 12,
- оперативная память исполнителя (свойство spark.executor.memory) - 11264 МБ,
- число ядер исполнителя (свойство spark.executor.cores) - 3.

Код обработки каждого раздела RDD - функция f - имеет следующую структуру:

- чтение данных из фрагмента csv-файла, соответствующего разделу RDD (функция read_csv библиотеки pandas), типы атрибутов фиксированы, автоопределение типов не используется;
- формирование экспертных атрибутов (функция DataFrame.eval билиотеки pandas);
- получение прогнозов каждого дерева, входящего в лес (функция ExtraTreesRegressor.predict библиотеки scikit-learn).

Листинг 1

```
  from io import StringIO
  import numpy as np
  import pandas as pd

  # b_estimator – широковещательная переменная – объект лес деревьев решений
  names=... # имена столбцов csv-файла
  dtype=... # типы столбцов csv-файла
  columns=... # атрибуты и экспертные комбинации – арифметические выражения, участвующие в прогнозировании

  def f(rows):
      content='\n'.join(rows)
      file=StringIO(content)
      df=pd.read_csv(file,             header=None, names=names, dtype=dtype);

      for c in columns:
          if c not in df.columns:
              df[c]=df.eval(c)

      e_preds=[]
      for e in b_estimator.value.estimators_:
          e_pred=e.predict(df[columns])
          e_preds.append(e_pred)
      test_pred=np.vstack(e_preds).T
      df2=pd.DataFrame(test_pred)

      return [r.tolist() for r in df[orig_columns].join(df2).to_records(index=False)]
```

Для того, чтобы предоставить астрофизикам-исследователям необходимые возможности дальнейшего анализа результатов, последние были сохранены в поколоночном формате Parquet[33] в хранилище Azure Blob Storage. Доступ к данным был предоставлен с помощью интерфейса Spark SQL в Jupyter Notebook. Предварительно были проведены тесты производительности на следующих типах запросов:

Запрос	Время выполнения
select * from res where run < 4617	20 сек
select avg(z_pred) from res	1 мин 8 сек
select run, avg(z_pred) from res group by run	1 мин 55 сек
select a.z_pred from res a inner join res b where a.objID=b.objID	3 мин

Для обработки аналитических запросов использовалась та же конфигурация кластера. Ввиду линейной горизонтальной масштабируемости реализованной процедуры применения случайного леса и обработки SQL запросов над данными, хранящимися в поколоночном формате, для достижения близких временных показателей при обработке больших объемов данных, например, 1 млрд. объектов, достаточно увеличить число узлов в кластере пропорционально росту объема данных (для 1 млрд. - в 10-15 раз).

V. Выводы

Методы машинного обучения позволяют прогнозировать красное смещение галактик и строить объемные карты распределения астрономических объектов во Вселенной, которые в настоящее время широко используются в фундаментальных задачах внегалактической астрономии и наблюдательной космологии. Применение построенных прогностических моделей к доступным астрономическим каталогам, содержащим данные широкополосной фотометрии для объектов на всем небе, требует значительных вычислительных ресурсов. В данной статье авторами был обобщен успешный опыт по применению технологии Apache Spark в облачной инфраструктуре Microsoft Azure для решения актуальной задачи точного прогнозирования фотометрических красных смещений (photo-z) галактик на большом массиве данных астрономических каталогов из небесного обзора SDSS.

Фреймворк Apache Spark позволяет программисту быстро реализовать необходимую логику пакетной обработки и одновременно предоставляет возможность исследователям-астрофизикам эффективно анализировать полученный прогноз красных смещений для всех объектов астрономического каталога используя доступные в данном фреймворке возможности SQL и средств визуализации.

Современные тренировочные данные в астрофизике, доступные для обучения модели photo-z галактик относительно невелики (несколько миллионов тренировочных объектов может быть собрано в общей сложности путем комбинации данных из различных спектральных обзоров) и, в данной работе, авторы предпочли выполнять обучение высокоточной регрессионной модели ERFDT на таких выборках нераспределенно. Тем не менее, в ближайшем будущем объем тренировочных данных в астрофизике должен существенно вырасти, за счет введения в строй таких проектов как DESI[34] (предполагается получение спектров десятков миллионов объектов). Ожидаемое увеличение на порядок доступных тренировочных выборок потребует использования распределенных

вычислений не только на этапе применения модели, но и на этапе ее обучения. Авторы планируют продолжить исследования в этом направлении.

Библиография

[1] Zhang, Y., Zhao Y. “Astronomy in the Big Data Era”, 2015, Data Science Journal, 14, p.11

[2] C.Snijders, U.Matzat, U. Reips “Big Data”: Big Gaps of Knowledge in the Field of Internet Science International Journal of Internet Science 2012, 7 (1), 1–5 ISSN 1662-5544

[3] T. Seth and V. Chaudhary, “Big Data in Finance”, in Big Data: Algorithms, Analytics, and Applications, Chapman and Hall/CRC Big Data Series, CRC Press, 2014

[4] A. Belle, R. Thiagarajan, S. Soroushmehr, F.Navidi, D. Beard, K. Najarian “Big Data Analytics in Healthcare” BioMed Research International Volume 2015 (2015)

[5] A. Greene, K. Giffin, C. Greene, J. Moore “Adapting bioinformatics curricula for big data” Briefings in Bioinformatics, 2015, 1–8

[6] J.Dean, S.Ghemawat “MapReduce: Simplified Data Processing on Large Clusters” OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004

[7] “Apache Hadoop” http://hadoop.apache.org дата запроса 3.12.2016

[8] “Apache Spark” http://spark.apache.org дата запроса 3.12.2016

[9] “The Sloan Digital Sky Survey: Mapping the Universe” http://www.sdss.org дата запроса 3.12.2016

[10] “ Pan-STARRS” http://pan-starrs.ifa.hawaii.edu/public/ дата запроса 3.12.2016

[11] “LSST Information for Scientists ” https://www.lsst.org/scientists дата запроса 3.12.2016

[12] “Euclid” http://sci.esa.int/euclid/ дата запроса 3.12.2016

[13] “Спектр-Рентген-Гамма” http://srg.iki.rssi.ru дата запроса 3.12.2016

[14] Pineau F. X. et al. “Probabilistic multi-catalogue positional cross-match” arXiv preprint arXiv:1609.00818. – 2016

[15] Ivezic et al “Statistics, Data Mining, and Machine Learning for Astronomy” Princeton University Press, 2014

[16] Bilicki et al. “WISE × SuperCOSMOS Photometric Redshift Catalog: 20 Million Galaxies over 3/pi Steradians” The Astrophysical Journal Supplement Series, Volume 225, Issue 1, article id. 5, 24 pp. (2016)

[17] Beck et al. “Photometric redshifts for the SDSS Data Release 12” Monthly Notices of the Royal Astronomical Society, Volume 460, Issue 2, p.1371-1381

[18] Soumagnac et al. “Star/galaxy separation at faint magnitudes: application to a simulated Dark Energy Survey” Monthly Notices of the Royal Astronomical Society, Volume 450, Issue 1, p.666-680 (2015)

[19] Brescia et al. “Automated physical classification in the SDSS DR10. A catalogue of candidate quasars” Monthly Notices of the Royal Astronomical Society, Volume 450, Issue 4, p.3893-3903 (2015)

[20] Möller et al. “Photometric classification of type Ia supernovae in the SuperNova Legacy Survey with supervised learning” arXiv:1608.05423 (2016)

[21] Elorrieta et al. “A machine learned classifier for RR Lyrae in the VVV survey” Astronomy & Astrophysics, Volume 595, id.A82, 11 pp. (2016)

[22] “SExtractor” http://www.astromatic.net/software/sextractor дата запроса 3.12.2016

[23] Abdalla et al. “A comparison of six photometric redshift methods applied to 1.5 million luminous red galaxies” Monthly Notices of the Royal Astronomical Society, Volume 417, Issue 3, pp. 1891-1903 (2011)

[24] Geurts, P., Ernst, D. & Wehenkel, L. “Extremely randomized trees” Mach Learn (2006) 63: 3. doi:10.1007/s10994-006-6226-1

[25] Meshcheryakov A. et al. “High-accuracy redshift measurements for galaxy clusters at z < 0.45 based on SDSS-III photometry” Astronomy Letters, Volume 41, Issue 7, pp.307-316 (2015)

[26] Brescia et al. “A catalogue of photometric redshifts for the SDSS-DR9 galaxies” Astronomy & Astrophysics, Volume 568, id.A126, 7 pp. (2014)

[27] “scikit-learn” http://scikit-learn.org дата запроса 3.12.2016

[28] Carrasco Kind M. & Brunner R. “TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests” Monthly Notices of the Royal Astronomical Society, Volume 432, Issue 2, p.1483-1501 (2013)

[29] N. Meinshausen “Quantile Regression Forests” Journal of Machine Learning Research 7 (2006) 983–999

[30] “Apache Spark MLlib” http://spark.apache.org/mllib/ дата запроса 3.12.2016

[31] “Scikit-learn integration package for Spark” http://github.com/databricks/spark-sklearn дата запроса 3.12.2016

[32] “PySpark+Scikit-learn=Sparkit-learn” http://github.com/lensacom/sparkit-learn дата запроса 3.12.2016

[33] “Apache Parquet” http://parquet.apache.org/ дата запроса 3.12.2016

[34] “Dark Energy Spectroscopic Instrument ” http://desi.lbl.gov дата запроса 3.12.2016

Processing and analysis of large amount of astronomical data on Microsoft Azure HDInsight

S.V.Gerasimov, A.V.Mesheryakov

Annotation — **Machine learning provides effective techniques to accurately measure photometric redshifts (photo-z) of extragalactic astronomical objects, which allows researchers to build maps of Large Scale Structure of the Universe. These maps are widely used in various fundamental research fields of extragalactic astrophysics and observational cosmology. Though making predictions by these models for a huge number of objects in astronomical catalogs, containing a broad-band photometry over all the sky, is a challenging task and requires a significant computational resources. In the article we tested the Apache Spark horizontally-scalable framework, deployed in the cloud Microsoft Azure, for the task of photo-z measurements for galaxies from the big photometric dataset of Sloan Digital Sky Survey.**

Keywords — **machine learning, big data, forecasting, sky catalogues, red shift estimation, random forest, MapReduce, Apache Spark, Microsoft Azure**