Impact Factor:	ISRA (India) = 1.344	SIS (USA) = 0.912	ICV (Poland) = 6.630
	ISI (Dubai, UAE) = 0.829	РИНЦ (Russia) = 0.207	PIF (India) = 1.940
	GIF (Australia) = 0.564	ESJI (KZ) = 4.102	IBI (India) = 4.260
	JIF = 1.500	SJIF (Morocco) = 2.031	

SOI: 1.1/TAS DOI: 10.15863/TAS

International Scientific Journal

Theoretical & Applied Science

p-ISSN: 2308-4944 (print) **e-ISSN:** 2409-0085 (online)

Year: 2018 **Issue:** 03 **Volume:** 59

Published: 30.03.2018 **http://T-Science.org**

Aleksandra Dmitrievna Soboleva
Bachelor of Peter the Great St. Petersburg Polytechnic University
sania.soboleva@mail.ru

Oleg Yurievich Sabinin
Candidate of technical sciences, Docent, Department of Intellectual Sciences and Technology of Peter the Great St. Petersburg Polytechnic University
olegsabinin@mail.ru

SECTION 4. Computer science, computer engineering and automation.

ENSEMBLE LEARNING METHOD DEVELOPMENT FOR SOLVING THE PREDICTION PROBLEM ON THE EXAMPLE OF ORACLE DATA MINING TECHNOLOGY

Abstract: *The article reviews the existing machine learning methods, which solve the prediction problem, and related issues. The modification of the bagging method, which aggregates two fundamentally different basic machine learning algorithms, is proposed and justified. The research of the method, based on the examples of risk estimation of the cardiovascular disease and forecasting the dynamics of General Electronic company stock, confirms the effectiveness of the method.*

Key words: *prediction problem, classification, regression, bagging, data mining, machine learning, Oracle Data Mining*

Language: *Russian*

Citation: Soboleva AD, Sabinin OY (2018) ENSEMBLE LEARNING METHOD DEVELOPMENT FOR SOLVING THE PREDICTION PROBLEM ON THE EXAMPLE OF ORACLE DATA MINING TECHNOLOGY. ISJ Theoretical & Applied Science, 03 (59): 147-154.

Soi: http://s-o-i.org/1.1/TAS-03-59-24 ***Doi***: crossref https://dx.doi.org/10.15863/TAS.2018.03.59.24

РАЗРАБОТКА МЕТОДА КОМПОЗИЦИИ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ НА ПРИМЕРЕ ТЕХНОЛОГИИ ORACLE DATA MINING

Аннотация: *В статье рассмотрены существующие методы машинного обучения, решающие задачу прогнозирования, и обозначены их недостатки. Предложен и обоснован метод, решающий задачу прогнозирования посредством агрегирования результатов двух базовых алгоритмов машинного обучения, противоположных по природе. Проведено исследование метода на примере оценки рисков кардиологических заболеваний и прогнозирования динамики роста акций компании General Electronic, подтвердившее эффективность разработанного подхода.*

Ключевые слова: *машинное обучение, задача прогнозирования, классификация, регрессия, интеллектуальный анализ данных*

Введение

На сегодняшний день компьютерные технологии заняли одно из главных мест, как в повседневной жизни человека, так и в бизнесе. Благодаря быстрому развитию и совершенствованию аппаратной части и программного обеспечения цифровых устройств за пару десятков лет заметно снизилась стоимость вычислительных ресурсов, в том числе и параллельных, оперативной и постоянной памяти. Все это привело к накоплению больших объемов разнородных данных, исчисляющихся терабайтами, которые обозначаются термином большие данные. Такое количество информации человек не способен обработать вручную, кроме того, и традиционные программы, имеющие конечное число решений и состояний, также плохо справляются с задачами анализа и обработки больших данных. В связи с этим появилось новое направление в науке и технологиях - машинное обучение.

Так как анализ больших данных изначально является не искусственной задачей, а необходимостью современной жизни человека, существует множество различных алгоритмов машинного обучения, среди которых выделяются несколько главных групп по свойствам решаемых ими задач. При применении алгоритма

машинного обучения для решения конкретной задачи в большинстве случаев требуются дополнительные эвристики, эксперименты и различные модели для компенсации зависимостей и закономерностей, свойственных выбранному подмножеству данных и предположений.

Таким образом, с помощью алгоритмов и методов машинного обучения можно решать большое число практических задач в разнообразных сферах человеческой деятельности. Одной из таких задач является задача прогнозирования, которая заключается в предсказании значений некоторых параметров или свойств системы в будущем на основе зависимостей, которые могут быть найдены с помощью известных параметров и поведении системы в прошлом и настоящем. Например, в медицине можно предсказывать риски заболеваний по анализам пациента, в экономике – поведение рынка как ценных бумаг, так и потребительского, кроме того появляется возможность оценить потребительскую корзину и создать персонализированную рекламу. В банковской сфере – оценивать кредитные риски, в сфере безопасности – выявлять мошенников и преступников, в повседневной жизни – экономить время и беречь здоровье за счет предсказания времени прибытия общественного транспорта.

Цель работы

Для прогнозирования каких-либо значений, будь то курс криптовалюты в следующем месяце, вероятность развития сердечного заболевания через несколько лет или же состав потребительской корзины, требуется выявить закономерность по существующим данным, которые были собраны в прошлом и настоящем.

На сегодняшний день существует несколько алгоритмов машинного обучения, решающих данную задачу, каждый из которых ищет закономерность, основываясь на теоремах математической статистики, теории вероятности, дискретной математики или теории графов. Кроме того, базовые алгоритмы объединяют в композиции для получения более точной модели.

Но все же остается вопрос, какой алгоритм выбрать для решения задачи прогнозирования.

Целью данной работы является разработка универсального метода композиции алгоритмов машинного обучения для решения задачи прогнозирования на примере технологии Oracle Data Mining.

Алгоритмы машинного обучения, решающие задачу прогнозирования

Для начала рассмотрим подходы и методы решения задачи прогнозирования, существующие на сегодняшний день.

Данная задача является обобщением двух классических задач интеллектуального анализа данных: классификации и регрессии. Они относятся к прогностическому обучению, которое также называют обучением с учителем [1, с.2]. Его суть заключается в том, чтобы научиться связывать входные и выходные значения. При обучении прогнозирующей модели на вход подается тренировочный набор данных, на основе которого выводятся взаимосвязи или функции зависимостей, с помощью которых уже на реальных данных получают требуемые значения. Сам тренировочный набор данных представляет из себя множество экземпляров, каждый из которых описан значениями атрибутов или свойств.

Суть классификации состоит в определении зависимости принадлежности элемента к классу на основе его свойств по средству сопоставления входных и выходных данных [2, с.5]. Регрессия одновременно очень схожа и отличается от классификации. Главное отличие регрессии в том, что она позволяет обрабатывать как дискретные, так и непрерывные величины, что является значительным достоинством, поскольку большая часть измерений в реальном мире описываются законами или функциями [3]. Второе принципиальное отличие регрессии от классификации состоит в том, что результатом классификации является вероятность принадлежности элемента к некоторому классу, а регрессии – определенное значение, выбранного для прогнозирования параметра.

Задачу классификации решает большой круг алгоритмов машинного обучения. К нему относят байесовский классификатор, линейный классификатор, решающие деревья, решающие списки, логистическую регрессию, машину опорных векторов и их модификации [1, 4].

Кроме того, существует несколько разных подходов к составлению из данных алгоритмов композиции.

Первый подход заключается в обучении каждого алгоритма на случайном подмножестве тренировочного набора данных, причем каждый алгоритм получает свое подмножество на вход и данные подмножества могут пересекаться, и агрегации результатов данных алгоритмов путем простого или взвешенного голосования [5, с.587].

Второй подход заключается в последовательном обучении каждого алгоритма на той подвыборке тренировочного набора данных, на котором предыдущие алгоритмы показали недостаточно точный прогноз [1, с.556].

Поскольку задача регрессии очень схожа с задачей классификации, алгоритмы, решающие

эти две задачи, также схожи. Задачу регрессии решают алгоритм линейной регрессии, алгоритм нелинейной регрессии и метод опорных векторов [2].

Стоит отметить, что обобщенная линейная модель объединяет в себе как линейный классификатор, так и алгоритм линейной регрессии.

Таким образом, с помощью решения задачи классификации мы сможем спрогнозировать динамику некоторого параметра, например, динамику роста или спада спроса на кукурузу, или оценить риск появления урагана через месяц, а с помощью решения задачи регрессии предсказать значение параметра, например, цену на кукурузу или день появления урагана.

Проблемы существующих методов

Выбор алгоритма машинного обучения зависит от требований и условий поставленной задачи. Кроме того, точность решения задачи классификации и регрессии очень чувствительна к данным. Таким образом, при сильно зашумленных данных, или при малой и однотипной обучающей выборке, которая содержит дополнительные свойства, не распространяющиеся на все множество, для которого производится прогноз, невозможно получить качественный результат [6]. Кроме того, принципы, на которых построены алгоритмы машинного обучения, различаются, что позволяет им находить зависимости абсолютно разной природы: от линейных зависимостей до зависимостей очень сложной и причудливой формы.

Существует также проблема эффекта переобучения, который заключается в том, что на тренировочном наборе данных получается модель высокой точности, а на реальных данных параметры модели не удовлетворяют поставленной задаче [1, с.22]. В связи с этим требуется очень тщательно подбирать данные для обучения и тестирования, чтобы экземпляры тренировочного набора как можно шире были распределены в множестве всех возможных значений, а не представляли собой подмножество, имеющее особый признак относительно всего множества экземпляров.

Поскольку не всегда есть возможность проводить трудоемкий анализ данных для выявления свойств зависимостей, например, из-за недостатка квалификации сотрудников, требуется универсальный метод выбора алгоритма машинного обучения, который будет рассматривать данные с помощью разных подходов.

Универсальный метод композиции алгоритмов машинного обучения

Поскольку, как было сказано выше, задачи классификации и регрессии во многом схожи, предположим, что существует универсальный метод для их решения. В таком случае возьмем несколько алгоритмов, которые решают обе задачи, имеют различную природу и построим над ними композицию с помощью агрегирования результатов нескольких базовых алгоритмов, построенных на основе различных подмножеств элементов тренировочной выборки. Заметим, что, чем меньше алгоритмов участвуют в композиции, тем меньше используется вычислительных ресурсов и памяти.

Метод агрегирования результатов нескольких алгоритмов композиции рекомендуется использовать с нечетным количеством базовых алгоритмов, что связано с проблемой состояния неопределенности, в случае, когда результаты базовых алгоритмов различаются в равной степени. Однако, если произвести некоторую модификацию правила выбора результирующего значения, то можно построить композицию над двумя алгоритмами. Модификация заключается в использовании вероятности отнесения данного объекта к выбранному классу. То есть результирующим является тот ответ, у которого вероятность больше. В случае, если вероятности равны, выберем результат первой базовой модели. Из-за того, что, значения вероятностей принадлежат бесконечному множеству действительных чисел в интервале от 0 до 1, в отличие от конечного множества результатов классификации, то вероятность получения ситуации неопределенности стремится к 0. При решении задачи регрессии будем выбирать среднее арифметическое значение из двух результатов.

Базовые алгоритмы композиции

Как уже было сказано, для универсальности метода требуется выбрать алгоритмы разной природы. Такими алгоритмами являются линейный классификатор и машина опорных векторов с нелинейным гауссовским ядром [2, с.54, 4, с.247]. Данные алгоритмы решают обе задачи и строят разделяющую поверхность классов разными методами: классическими градиентными методами, минимизируя ошибку, и средствами квадратичного программирования, максимизируя зазор между классами, соответственно. Таким образом, алгоритмы способны компенсировать друг друга. Кроме того, компенсация будет производиться за счет построения индивидуальной обучающей выборки для каждого базового алгоритма путем случайного выбора элементов из тренировочного набора.

Следует отметить также, что к достоинствам линейной регрессии можно отнести быстроту и

простоту создания модели, результат для которой может быть получен аналитически. Также данная модель позволяет сделать дополнительные выводы о характере зависимости предикторов и отклика по коэффициентам регрессии. К тому же, данный алгоритм хорошо изучен: известны его проблемы и методы их решения.

Главным преимуществом метода опорных векторов является сведение обучения машины опорных векторов к задаче квадратического программирования, которая имеет единственное решение и эффективное вычисление, в том числе в случаях больших объемов данных обучающей выборки. Кроме того, оптимальное положение разделяющей гиперплоскости зависит только от опорных векторов, которые составляют малую долю всех объектов выборки.

Таким образом, в результате составления композиции на основе данных алгоритмов получим метод, который позволит прогнозировать любые показатели на основе любых данных, не требуя предварительного аналитического и статистического анализа.

Реализация

Постановка задачи прогнозирования для реализации

Для реализации, описанного ранее универсального метода выбора алгоритма машинного обучения для решения задачи прогнозирования, были сформулированы следующее задания:

- оценка рисков кардиологических заболеваний;
- прогнозирование динамики фондового рынка.

Оценка рисков кардиологических заболеваний на основе анализов пациента является одной из задач медицины, которая хорошо поддается математическому и интеллектуальному анализу [7]. Кроме того, существует аналог ее решения, который называется формулой Фременгхэма [8]. Данная формула используется на практике врачами и основана на подсчете суммы баллов, которые присваиваются или изымаются в зависимости от показаний пациента. Стоит отметить, что данная формула гарантирует свой результат лишь на 30%.

Прогнозирование динамики курса акций является актуальной задачей в сфере экономики и финансов как со стороны участника торгов, который должен принять решение о приобретении или продаже акций, так и со стороны организаторов фондового рынка для управления над ним. Несколько лет назад данное прогнозирование производилось с помощью технического анализа, но он не охватывает всех параметров и объектов, влияющих на цены акций [9]. Именно поэтому, с популяризацией и массовым внедрением программных средств, использующих алгоритмы машинного обучения, прогнозирование динамики фондового рынка выходит на новый уровень, качественно превосходящий технический анализ.

В ходе данной работы были построены следующие 4 модели, использующие предложенный метод композиции алгоритмов машинного обучения, на примере технологии Oracle Data Mining:

- модель, прогнозирующая присутствие или отсутствие кардиологического заболевания пациента;
- модель, прогнозирующая время, через которое у пациента появится кардиологическое заболевание;
- модель, прогнозирующая рост или спад цен акций компании General Electronic;
- модель, прогнозирующая цены закрытия акций компании General Electronic.

Таким образом, модели, построенные с помощью одного метода, решают, как задачи классификации, так и задачи регрессии на двух различных наборах данных из разных областей человеческой деятельности.

Технология для реализации

В наши дни как уже было сказано ранее накоплены экзабайты данных ретроспективного характера. Только половина этих данных структурирована и может быть подвержена интеллектуальному анализу. Самый распространенный способ структуризации и хранения данных, это организация баз данных. Именно в них находиться большая часть структурированных данных. В связи с этим, для обеспечения большей производительности и защиты, а также снижению накладных расходов на передачу данных, проведение интеллектуального анализа средствами базы данных является более предпочтительным. В данной работе используется технология системы управления реляционными базами данных Oracle Enterprise Edition версии 12c Oracle Data Mining.

В ходе данной работы при построении моделей был использован PL/SQL API, который реализован пакетом DBMS_DATA_MINING [10].

Исходные данные

В качестве исходных данных для моделей, оценивающих риски кардиологических заболеваний, в данной работе был использован набор «Framingham Heart Study», который был запрошен в Национальном Университете Сердца, Легких и Крови. Этот набор представляет собой данные о пациенте, такие как возраст, пол,

индекс массы тела, показатели давления, наличие или отсутствие диабета и так далее [11]. Показатели для каждого пациента представлены 3 записями, поскольку были сняты 3 раза на протяжении 12 лет через равные промежутки времени. Из данного набора было выбрано случайным образом 10000 строк, которые позднее были разбиты на две части: обучающую и тестовую, которые составляют 7000 и 3000 записей соответственно.

В качестве входных данных для моделей, прогнозирующих динамику курса акций, было выбрано подмножество данных «Huge Stock Market Dataset», содержащее информацию о акциях компании General Electronic. Данный набор находится в открытом доступе на платформе для соревнований в области науки о данных Kaggle [12]. Он представляет 10000 записей, содержащих дату, цены начала и окончания торгов, максимальную и минимальную цены акции и волатильность. Данный набор также был разделен на 2 части: обучающую и тестовую, которые составят 7000 и 3000 записей соответственно.

Далее в соответствии с рекомендациями Oracle были преобразованы типы данных в обоих наборах данных следующим образом:

- категориальные переменные, например, идентификатор пациента, преобразовываются к строковому типу данных;
- даты преобразовываются к численному типу данных посредством вычитания из сегодняшней даты – даты, указанной в наборе.

Причем за один объект набора «Framingham Heart Study» был взят пациент в определенный период сдачи анализов и показаний, для этого переменные идентификатора пациента и периода были объединены в одну с помощью операции конкатенации. Все остальные переменные, которые должны быть предсказанными, были удалены, так как в случае реальных данных их значения не определены, то есть они присутствуют только в обучающем и тестовом наборах. Таким образом, остается только информация о заболеваниях инфаркта миокарда и коронарной болезни сердца, которые представлены одной переменной. Также показания о холестерине были отброшены, поскольку неопределенных значений в 5 раз больше определенных. Это связано с тем, что этот показатель измерялся только в последнем периоде из трех.

Таким образом, были сформированы обучающие и тестовые выборки, в той форме, в которой модель способна правильно их трактовать.

Построение моделей

Для реализации метода композиции базовых алгоритмов машинного обучения потребовалось написать на языке PL/SQL код-надстройку для вызова процедуры создания модели пакета DBMS_DATA_MINING. Это связно с тем, что помимо непосредственного вызова процедуры создания модели требуется подготовка индивидуальных, для каждого базового алгоритма композиции, выборок и получение прогнозируемого значения на основе результатов базовых алгоритмов.

Создание индивидуальной выборки

Поскольку метод основывается на композиции, агрегирующей результаты нескольких базовых алгоритмов, построенных на основе различных подмножеств элементов обучающей выборки, первым этапом построения модели является создание этих выборок. В нашем случае их две, поскольку по описанным ранее причинам в методе используется два базовых алгоритма. Отметим, что каждая такая выборка должна содержать подмножество элементов обучающей выборки, которые могут повторяться. В следствии чего, строки тренировочного набора были пронумерованы, каждый раз генерировалось псевдослучайное число и в текущую подвыборку была добавлена та строка из обучающей выборки, номер которой совпадал с полученным числом. Данная операция была повторена столько раз, сколько строк требовалось в подвыборке. В данной работе такое число 5000, поскольку оно должно быть меньше, чем количество строк исходной выборки, но все же строк должно быть достаточно для устойчивости к переобучению. Опытным путем было выяснено, что среди 4000, 5000 и 6000 записей в индивидуальной выборке, 5000 показывают наилучший результат работы метода.

Для повышения производительности было принято решение каждый раз добавлять не по одной строке, а по 1000. Для этого сначала было сгенерировано 1000 псевдослучайных чисел, а затем в таблицу каждой индивидуальной выборки были добавлены те строки из обучающей выборки, номера которых совпадали с полученными числами. Поскольку псевдослучайные числа повторялись, то каждый раз в выборку добавлялось меньше 1000 строк. Данное свойство было учтено, вследствие чего алгоритм генерации псевдослучайных чисел и их добавления в подвыборку был запущен на 1 раз больше. Таким образом, для получения 5000 строк потребовалось 6 запусков алгоритма, при этом полученная выборка содержала около 5500 строк.

Выбор параметров модели

Далее потребовалась настройка каждой модели. Идентификатор задачи, которую решает модель, было решено передавать в процедуру, реализующую метод композиции базовых алгоритмов, в качестве параметра. Для него осуществлена проверка на соответствие одному из двух допустимых значений. Этот параметр также передается обоим базовым алгоритмам на вход. Далее для каждой из моделей был включен параметр ADR, который позволяет автоматически производить нормировку входных данных при построении модели базовым алгоритмом.

В соответствии с реализуемым методом первая базовая модель была основана на обобщенном линейном алгоритме. Такая модель способна выявлять линейные зависимости и закономерности. Второй базовой модели соответствовала машина опорных векторов с гауссовским ядром, а также было разрешено применение активного обучения, что позволяет сократить использование ресурсов памяти.

Остальные параметры принимали значения по умолчанию. Также стоит отметить, что большая часть параметров принимают значения автоматически, основываясь на входных данных [10].

Таким образом, была получена таблица настроек для каждой из двух моделей, которые могут быть поданы на вход процедуре построения модели CREATE_MODEL пакета DBMS_DATA_MINING.

Описание процедуры, реализующей метод композиции базовых алгоритмов машинного обучения

Как уже было сказано ранее, реализация предложенного в данной работе метода требует надстройки над PL/SQL API, который поставляется технологией Oracle Data Mining. Для этого была создана хранимая PL/SQL процедура, принимающая на вход название модели, имя таблицы с обучающей выборкой, идентификатор решающейся задачи, имя ключевого атрибута и имя атрибута, который требуется спрогнозировать.

Кроме того, данная процедура принимает на вход флаг, который отвечает за создание индивидуальных выборок для каждого базового алгоритма. В таком случае модель можно пересоздать, например, если требуется спрогнозировать другой атрибут или проанализировать работу метода при разных параметрах, не пересоздавая и не меняя при этом сами данные. Создание индивидуальной выборки в данной работе реализовано отдельной хранимой процедурой PL/SQL, которая принимает на вход название таблицы с обучающей выборкой, количество выборок, которое должно получиться на выходе, и количество тысяч строк, которое должно быть в каждой из выходных выборок.

В зависимости от значения флага в процедуре, реализующей универсальный метод, вызывается или не вызывается процедура создания двух подвыборок на 5000 записей. Далее вызывается процедура создания базовой модели на основе обобщенного линейного алгоритма для первой индивидуальной выборки, а затем на основе машины опорных векторов для второй.

Затем создается представление, которое содержит два столбца: идентификатор объекта и спрогнозированное значение. При решении задач классификации выбирается то значение из двух, у которого вероятность больше. В случае решения задачи регрессии спрогнозированным значением становится среднее арифметическое результатов работы двух базовых моделей.

Анализ полученных результатов

После построения моделей, прогнозирующих риски кардиологических заболеваний и динамику роста или спада цен акций компании General Electronic, были оценены ошибки этих моделей на основе тестовых выборок.

Для моделей, решающих задачу классификации, была посчитана доля неправильных ответов по формуле (1):

$$Q(a, X) = \frac{1}{m}\sum_{i=1}^{m}[a(x_i) \neq y_i], \qquad (1)$$

где x_i – i-ый объект тестовой выборки, y_i – исходный идентификатор класса объекта x_i, a – спрогнозированный идентификатор класса объекта x_i моделью, m – количество объектов тестовой выборки X. Для моделей, решающих задачу регрессии была посчитана среднеквадратичная и абсолютная средняя ошибки по формулам (2) и (3) соответственно:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \qquad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \qquad (3)$$

где y_i – исходный идентификатор класса объекта, $\hat{y}_i$ – спрогнозированный идентификатор класса объекта, n – количество объектов тестовой выборки.

Оценки ошибок для универсального метода, представлены в строке с названием алгоритма «Универсальный метод» в таблице 1. Для проведения сравнительного анализа в ходе данной работы были также построены модели на основе тех же обучающих выборок, что и

предложенный метод, с помощью реализованных алгоритмов машинного обучения в технологии Oracle Data Mining, решающих задачи классификации и регрессии. Кроме того, был построен прогноз с помощью встроенной процедуры для прогнозирования технологии Oracle Data Mining, которая является частью пакета DBMS_PREDICTIVE_ANALYTICS.

Стоит отметить, что данная процедура принимает на вход только название таблицы, содержащей обучающую выборку, и подбирает параметры модели, в том числе и алгоритм, автоматически.

Таблица 1

Результаты оценки ошибок моделей

Алгоритм	Классификация		Регрессия			
	Болезни	Акции	Болезни		Акции	
			RMSE	MAE	RMSE	MAE
Линейная обобщенная модель	0,131	0,233	1766,771	1288,382	0,213	0,142
Линейная обобщенная модель с гребневой регрессией	0,132	0,520	1766,848	1288,504	0,280	0,181
Дерево решений	0,132	0,480	-	-	-	-
Наивный байесовский классификатор	0,153	0,520	-	-	-	-
Машина опорных векторов с линейным ядром	0,127	0,000	1852,090	1309,244	0,231	0,153
Машина опорных векторов с гауссовским ядром	0,130	0,520	2292,842	1948,004	22,655	19,927
Универсальный метод	0,128	0,172	1789,762	1390,731	11,488	10,146
DBMS_PREDICTIVE_ANALYTICS. PREDICT	0,289	0,382	1685,300	1070,462	0,207	0,150

Из таблицы 1 видно, что для решения задачи классификации универсальный метод находиться на втором месте по точности, уступая машине опорных векторов с линейным ядром. В случае прогнозирования динамики фондового рынка на примере цен акций компании General Electronic мы видим нулевую ошибку, что на первый взгляд кажется идеальным результатом, но на самом деле это пример проблемы переобучения. В данном случае, закономерность, найденная алгоритмом машинного обучения, распространяется не только на обучающую выборку, но также и на тестовую. В данном случае, прогноз все равно является точным, но не для всех данных он будет абсолютно безошибочным.

Также отметим, что процедура DBMS_PREDICTIVE_ANALYTICS.PREDICT, которая выбирает и настраивает модель автоматически, основываясь на обучающей выборке, показывает менее точный результат, чем полученная модель.

Теперь рассмотрим оценки ошибок для моделей, решающих задачу регрессии.

Наилучший результат у встроенной процедуры технологии Oracle Data Mining DBMS_PREDICTIVE_ANALYTICS.PREDICT, а наихудший – у модели, построенной на основе машины опорных векторов с гауссовским ядром.

Отметим, что в наихудшем случае величина среднеквадратичной ошибки примерно равна среднему значению прогнозируемой величины, которое составляет 22,465 доллара, в случае прогнозирования конечной цены акций компании General Electronic. Заметим, что второй по возрастанию величины ошибки является модель, построенная с помощью линейного обобщенного алгоритма.

Таким образом, в случае данных об акциях предложенный в данной работе метод основан на модели с одним из лучших показателей и самым худшим и его точность является примерно средним значением точностей базовых алгоритмов. Это происходит за счет того, что каждый раз значением прогноза становится среднее значение из результатов двух базовых алгоритмов, и тем самым значение с большим отклонением компенсируется значением с меньшим, и значение с меньшим отклонением отклоняется еще больше за счет значения с большим.

Данная проблема, также, наблюдается и в случае оценки рисков кардиологических заболеваний, но в менее явном виде. Это связано с тем, что величина среднеквадратичной ошибки в случае модели, с наихудшим алгоритмом – машиной опорных векторов с гауссовским ядром, составляет всего 30% прогнозируемой величины,

что равно 7498,530 дням, а не находится с ней в равенстве.

Таким образом, метод композиции алгоритмов для решения задачи прогнозирования, основанный на двух противоположных по природе алгоритмах машинного обучения, таких, как обобщенная линейная модель и машина опорных векторов с гауссовским ядром, увеличивает точность базовых алгоритмов в случае решения задачи классификации, и компенсирует точность базового алгоритма с меньшим показателем точности за счет второго базового алгоритма, точность которого выше.

Кроме того, данный метод сохраняет свойства композиционного агрегирования результатов нескольких базовых алгоритмов, построенных на основе различных подмножеств элементов обучающей выборки, такие как устойчивость к исходным зашумленным данным и к переобучению.

References:

1. Murphy, Kevin P. (2012) Machine learning: a probabilistic perspective. – The MIT Press, 2012. – 1104 p.
2. Vorontsov K.V. (2011) Mathematical methods of learning by use of precedents // Machine learning theory. – 2011 – 141 p, URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (Data of access 13.02.18)
3. Kashnitsky Y. (2017) Classification and regression liner model // Open machine learning course on Habrahabr, 2017, URL: https://habrahabr.ru/company/ods/blog/323890/ (Data of access: 10.02.18)
4. Segaran, Toby (2008) Programming Collective Intelligence. – O'Reilly Media, 2008. – 368 p.
5. Hastie T., Tibshirani R., Fiedman J. (2009) The Elements of Statistical Learning 2nd edition. – Springer-Vrlag, 2009. – 763 p.
6. Kaftannikov I.L., Parasich A.V. Problems of Training Set's Formation in Machine Learning Tasks. Bulletin of the South Ural State Univercity. Ser. Computer Technologies, Automatic Control, Radio Electronics, 2016/ vol 16, no3, pp.15-24.
7. Fuster-Parra P., Tauler P., Bennasar-Veny M., Ligeza A., Lopez-Conzalez A.A., Aguilo A. (2016) Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk // Computer Methods and Programs in Biomedicine, 2016 // no126, pp.128-142
8. Systematic Evidence Review From the Cholesterol Expart Panel, 2013 // Managing Blood Cholesterol in Adults – U.S. Department of Health and Human Service National Institutes of Health – 2013, p.239
9. Schwager J.D. (1995) Schwager on Futures: Technical Analysis – Wiley, p.800
10. Oracle DBMS_DATA_MINING Online Documentation, URL: https://docs.oracle.com/cd/B19306_01/appdev.102/b14258/d_datmin.htm#BAJBGHGD (Data of access: 13.03.18)
11. Framingham Heart Study. Longitudinal Data Documentation. URL: https://biolincc.nhlbi.nih.gov/static/studies/teaching/framdoc.pdf?link_time=2018-03-13_11:44:27.687606 (Data of access: 15.12.17)
12. Marjanovic B (2017) Overview of Huge Stock Market Dataset on Kaggle, URL: https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs (Data of access: 16.12.17)
13. Witten, Ian H., Frank (2005) Data mining: practical machine learning tools and techniques – 2-nd ed. – Elsevier, 2005 – 560 p.