Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Физико-Механический Институт

Работа допущена к защите
Руководитель ВШПМиВФ
_____ К. Н. Козлов
«_____» _____ 2022 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
РАБОТА БАКАЛАВРА
РАЗРАБОТКА МОДЕЛИ ПРОГНОЗИРОВАНИЯ ВРЕМЕНИ ЦВЕТЕНИЯ НУТА С ИСПОЛЬЗОВАНИЕМ ПЕРЕНОСА ЗНАНИЙ И СИМВОЛЬНОЙ РЕГРЕССИИ

по направлению подготовки 01.03.02 Прикладная Математика и Информатика
Направленность (профиль) 01.03.02_04 Биоинформатика

Выполнил
студент гр. 5030102/80401 Я.А.Тырыкин

Руководитель
доцент,
к. б. н. ВШПМиВФ ФизМех, СПБПУ К. Н. Козлов

Консультант
по нормоконтролю Л. А. Арефьева

Санкт-Петербург
2022

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-Механический Институт

УТВЕРЖДАЮ
Руководитель Образовательной программы
"Прикладная Математика и Информатика"
_____ К. Н. Козлов
«10» сентября 2021 г.

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

студенту Тырыкину Ярославу Алексеевичу гр.5030102/80401

1. Тема работы: Разработка модели прогнозирования времени цветения нута с использованием переноса знаний и символьной регрессии.
2. Срок сдачи студентом законченной работы: 04.06.2022.
3. Исходные данные по работе:
   Данные по времени цветения и однонуклеотидным полиморфизмам образцов дикого нута.
   Инструментальные средства:
   - Языки программирования C++, Python
   - Библиотеки numpy, sympy, scipy, multiprocessing
   - Среда разработки PyCharm
   - Система контроля версий git
4. Задействованные литературные источники:
   - Qi Chen, Bing Xue, Mengjie Zhang. Genetic Programming for Instance Transfer Learning in Symbolic Regression, IEEE Transactions On Cybernetics, 2020 г., 1–14
     https://ieeexplore.ieee.org/document/7257018
   - Kozlov, K., Samsonov, A.M., and Samsonova, M. (2016). A software for parameter optimization with Differential Evolution Entirely Parallel method. PeerJ Computer

Science 2, e74
https://peerj.com/articles/cs-74/
– Kozlov, K., Singh, A., Berger, J., Wettberg, E.B., Kahraman, A., Aydogan, A., Cook, D., Nuzhdin, S., and Samsonova, M. (2019). Non-linear regression models for time to flowering in wild chickpea combine genetic and climatic factors. BMC Plant Biology 19, 94.
https://doi.org/10.1186/s12870-019-1685-2
– Virgolin, M. (2021). Genetic programming is naturally suited to evolve bagging ensembles. In Proceedings of the Genetic and Evolutionary Computation Conference, (Lille France: ACM), pp. 830–839.

5. Содержание работы (перечень подлежащих разработке вопросов):
   5.1. Введение. Обоснование актуальности.
   5.2. Постановка задачи.
   5.3. Обзор существующих решений.
   5.4. Разработка модификаций.
   5.5. Применение.
   5.6. Результаты и их сравнительный анализ
   5.7. Выводы
   5.8. Заключение
6. Перечень графического материала (с указанием обязательных чертежей):
   6.1. Схема работы метода/алгоритма.
   6.2. Архитектура разработанной программы/библиотеки.
7. Дата выдачи задания: 10.09.2021.

Руководитель ВКР _____ К. Н. Козлов

Задание принял к исполнению

Студент _____ Я. А. Тырыкин

РЕФЕРАТ

На 42 с., 18 рисунков, 5 таблиц, 0 приложений

КЛЮЧЕВЫЕ СЛОВА: РАЗНОСТНАЯ ЭВОЛЮЦИЯ, СИМВОЛЬНАЯ РЕГРЕССИЯ, РЕГРЕССИОННЫЕ МОДЕЛИ, ГЕНЕТИЧЕСКОЕ ПРОГРАММИРОВАНИЕ, ТРАНСФЕРНОЕ ОБУЧЕНИЕ, C++, PYTHON.

Тема выпускной квалификационной работы: «Разработка модели прогнозирования времени цветения нута с использованием переноса знаний и символьной регрессии».

В рамках данной работы производилась разработка модели предсказания времени цветения дикого нута на основании имеющихся данных, полученных при проведении полногеномного поиска ассоциаций, а также измерения климатических показателей за период, в течение которого проходило исследование. Модели подбирались при помощи модификации метода трансферного обучения, основанного на экземплярах целевой области.

Цели данной работы - разработка модифицированного метода трансферного обучения на основе разностной эволюции и генетического программирования в символьной регрессии для построения моделей прогнозирования цветения нута. После реализации алгоритма требовалось применить полученные модели к реальным данным о генетических факторах исследуемых образцов и погодных факторах в месте их высадки, провести анализ результатов предсказания и сравнение этих показателей с уже существующими моделями.

В ходе проведения численных экспериментов были получены модели, по структуре отличающиеся от уже существующих моделей прогнозирования времени цветения, но при этом близкие к ним по качеству предсказания достижения нутом вегетационного периода.

Программные компоненты, разработанные в ходе работы, могут применены для построения моделей прогнозирования в других предметных областях, как связанных с биоинформатикой, так и не имеющих с ней никаких пересечений. Тем более, что рассмотренное сочетание методов разностной эволюции и генетического программирования в символьной регрессии показало себя достаточно эффективно, и имеет смысл далее искать области их приложения к задачам реального мира.

ABSTRACT

42 pages, 18 figures, 5 tables, 0 appendices

KEYWORDS: DIFFERENTIAL EVOLUTION, TRANSFER LEARNING, SYMBOLIC REGRESSION, REGRESSION MODELS, GENETIC PROGRAMMING, C++, PYTHON.

The subject of the graduate qualification work is «Development of a model for predicting Chickpea flowering time prediction model using knowledge transfer and symbolic regression».

Within the framework of this work, a model for predicting the flowering time of wild chickpeas was developed based on the available data obtained during a genome-wide association search, as well as climatic indicators for the period during which the study took place. The models were selected using a modification of the transfer learning method based on instances of the target area.

The purpose of this work is to develop a modified transfer learning method based on difference evolution and genetic programming in symbolic regression to build chickpea bloom prediction models. After the algorithm was implemented, it was necessary to apply the obtained models to real data on the genetic factors of the studied samples and weather factors at their landing site, analyze the prediction results and compare these indicators with existing models.

In the course of numerical experiments, models were obtained that differ in structure from existing models for predicting the flowering time, but at the same time are close to them in terms of the quality of predicting the achievement of the growing season by chickpeas.

The software components developed in the course of the work can be used to build forecasting models in other subject areas, both related to bioinformatics and having no intersections with it. Moreover, the considered combination of methods of difference evolution and genetic programming in symbolic regression has proved to be quite effective, and it makes sense to further look for areas of their application to real-world problems.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Нут - одна из наиболее широко распространенных зернобобовых культур, которая выращивается более чем в 50 странах мира, от субтропических экваториальных регионов Южной Азии, Восточной Африки и Австралии и заканчивая полярными регионами Австралии и умеренными районами Северной и Южной Америки.

В рамках ВКР реализуется и применяется метод машинного обучения для построения регрессионной модели, предсказывающей начало вегетационного периода у дикого нута двух видов - *Cicer echinospermum* и *Cicer reticulatum* [1]. Метод машинного обучения, выбранный для разработки модели прогнозирования - трансферное обучение на основе экземпляров (ITGP), фактически являющийся методом генетического программирования, связывающий в себе работу символьной регрессии и разностной эволюции.

Метод применяется на сведениях о 6 наиболее интересных нуклеотидных полиморфизмах, обнаруженных в исследуемых разновидностях нута, и агроклиматических данных, собранных в областях посадки исследованных ОНП.

Основная цель данной работы - применение модифицированного алгоритма трансферного обучения для нахождения оптимальной модели, прогнозирующей время цветения нута по имеющимся генетическим данным исследуемых растений и климатическим данным за период созревания подопытных экземпляров, для ее дальнейшего анализа, тестирования и, возможно, практического применения. Кроме того, результаты сравниваются с моделями. полученными в статье [2], где также был рассмотрен используемый датасет.

Данная задача сохраняет свою актуальность в современном мире, поскольку в дальнейшем результаты работы - метод оптимизации, модель и найденные закономерности - можно распространить как на смежные предметные области, так и на области, никак не связанные с генетикой и биоинформатикой. Кроме того, прогностическая модель может в дальнейшем оптимизировать временные затраты на выбор оптимального региона производства и финансовые затраты на выращивание рассматриваемой бобовой культуры.

В ходе решения поставленной задачи были решались следующие шаги:

- Реализовать модификацию метода ITGP [3] для получения модели(ей), дающей качественную точность предсказания, сопоставимую с первоисточниками.

- При помощи полученных модели(ей) спрогнозировать время цветения растений, по которым имеются данные.
- Провести сравнительный анализ качества модели(ей), а также проверить их устойчивость к стремительно меняющемуся климату.

ГЛАВА 1. ОБЗОР МЕТОДОВ ТРАНСФЕРНОГО ОБУЧЕНИЯ

1.1. Классическое трансферное обучение

Transfer learning (трансферное обучение) - подраздел машинного обучения, основное назначение которого - применение знаний, полученных из одной задачи, к другой целевой задаче, схожей по структуре к исходной.

Формально определение переноса обучения можно дать следующим образом:

Определение (Transfer Learning): Пусть заданы следующие пары: область обучения D_s - исходная задача обучения T_s и целевая область D_t - целевая задача обучения T_t, причем $T_s \neq T_t$ или $D_s \neq D_t$. Тогда процесс переноса обучения заключается в использовании знаний, полученных в исходной области, дл улучшения обучения модели в целевой области.

В контексте данного определения важно раскрыть понятия *задачи* и *области*.

Определение (Область): *Областью* знаний называется упорядоченная пара $D = \{\chi, P(X)\}$, где $\chi-$ пространство признаков, $P(X)$ - предельное распределение вероятностей ($X = \{x_1, x_2, \ldots, x_d\} \in \chi$ - элемент описанного пространства признаков, d - число признаков-измерений данного пространства).

Определение (Задача): *Задачей* знаний называется упорядоченная пара $T = \{\Upsilon, f(X)\}$, где $\Upsilon-$ пространство меток, $f(X)$ - функция предсказания по исходному набору факторов размера d. Условие неравенства исходной и целевой задач ($T_t \neq T_s$) означает либо различие пространств меток ($\Upsilon_t \neq \Upsilon_s$), либо различие функций предсказания ($f_t(X) \neq f_s(X)$), в то время как неравенство целевой и исходной областей ($D_t \neq D_t$) означает либо различие пространств признаков ($X_t \neq X_s$), либо различие их предельных распределений вероятностей ($P(X_t) \neq P(X_s)$).

В процессе машинного обучения с переносом знаний происходит параллельное решение задач оптимизации каждой из рассматриваемых областей (решение объединяет оба домена), оценка и селекция лучших индивидов происходит в два этапа, где как раз знания, полученные при итерировании на исходных данных переходит на целевую часть данных. Алгоритм выглядит перспективно с точки зрения эффективности, так как параллельно решаются две задачи оптимизации, на выходе алгоритма получается решение, обобщающее знания из обоих доменов.

Трансферное обучение имеет широкое количество различных вариаций в зависимости от того, какие данные подвергаются переносу от исходной области

к целевой. В данной работе для разработки алгоритма используется концепция передачи данных на основе экземпляров исходной области.

1.2. Трансферное обучение на основе экземпляров

Взвешивание экземпляров, как общий метод коррекции статистических отклонений, потенциально может исправить смещение исходных данных. Поэтому коррекция расстояния между исходной и целевой областями рассматривается как важная и перспективная разновидность переноса обучения. Обоснование взвешивания экземпляров в трансферном обучении заключается в том, что из-за различных распределений в исходной и целевой областях некоторые данные исходной области могут быть вредоносными, в то время как некоторые другие могут быть повторно использованы для обучения в целевой области путем повторного взвешивания для снижения предельной разницы распределений между двумя областями знаний.

Учитывая распределение (плотность) данных исходной области $P_s(x)$ и целевой области $P_t(x)$, соотношение этих характеристик $w(x) = \frac{P_s(x)}{P_t(x)}$ показало свою эффективность в уменьшении разницы в распределении между двумя доменами. Таким образом, он широко использовался в трансферном обучении.

1.3. Эволюционное трансферное обучение

Одной из интересных вариаций трансферного обучения можно считать эволюционный перенос обучения, подробно описанный в статье [4]. Представленная модификация классического трансферного обучения - MIST - использует метаэвристику, основанную на предположении, что каждое решение представляет подмножество экземпляров исходного домена. Качество решения оценивается по предполагаемой ошибке обобщения гауссовского процесса, который обучается на основе комбинации выбранных исходных и целевых обучающих данных. По большому счету, данная концепция эквивалентна той, чтобы была рассмотрена в статье [3] и рассматривалась как база для реализуемого в данной работе алгоритма оптимизации. Минусом же такого подхода можно считать предположение равнозначности каждого из экземпляров исходной области, используемых для обучения.

Существует несколько модификаций данного алгоритма, описанных в статьях [5] и [6], которые обеспечивают лучшую сходимость относительно MIST засчет развития концепции пропорционального взвешивания и распараллеливания эволюционного процесса на несколько компонент. Знания в них передаются через усложненные нейронные структуры, засчет чего и были достигнуты более высокие показатели эффективности.

1.4. Перенос обучения в символьной регрессии

Данная концепция пока слабо исследована, и в статье [3] авторы как раз пробуют восполнить пробел. В большинстве случаев перенос знаний в символьной регрессии заключался в прогоне символьной регрессии на данных исходной области, а набор моделей, составленный из лучших индивидов, полученных во всем эволюционном процессе, и популяции экземпляров, полученных в терминальном поколении, использовался как инициализирующее поколение моделей в задаче символьной регрессии на данных целевой области. В последствии методика переноса знаний в символьной регрессии была доработана путем добавления механизма детекции повторно используемых моделей, вредивших процессу сходимости.

1.5. Существующие решения рассматриваемой задачи

Помимо описания существующих методов, модифицирующих трансферное обучение и применяющих перенос знаний на основе дополнительных данных, стоит также упомянуть уже разработанные на текущий момент модели прогнозирования времени цветения растительных культур. Некоторые из данных моделей продемонстрировали свое качество на множестве экспериментов и на данный момент используются во многих агроклиматических задачах. Например, метод DSSAT [7], позволяющий строить динамические модели роста более чем 40 растительных культур, на данный момент нашел обширное применение в сельском хозяйстве. Аналогичным спросом на данный момент пользуются также модели SSM [8] и APSIM [9]. Но данные системы строят модели роста на основании большого количества факторов, характеризующих параметры почвы и климата в месте посадки, а также генетических особенностей рассматриваемых растений. Попытка разработать модель, не использующую почвенные параметры в качестве предикторов, предпринималась авторами статьи [2].

Интересный подход разработки и использования динамической модели, прогнозирующей стадию созревания экземпляров нута на основании погодных и генетических факторов, позволил получить ошибку прогнозирования времени достижения вегетационного периода в пределах одной недели.

Однако разработанная модель хорошо работала лишь для части данных, собранных в Турции использованных в процессе ее разработки (кросс-валидации) и не давала приемлемый прогноз на экземплярах, собраных в Австралии и подверженных дополнительному орошеню. Именно с целью устранения этого существенного недостатка в данной статье предпринимается попытка создания модели на основе переноса знаний.

ГЛАВА 2. РАЗРАБОТКА МЕТОДА ТРАНСФЕРНОГО ОБУЧЕНИЯ НА ОСНОВЕ ЭКЗЕМПЛЯРОВ

В данной главе подробно опишем модификацию алгоритма, описанного в статье, при помощи которого разрабатывалась модель прогнозирования времени цветения нута. Метод переноса обучения на основе экземпляров (Instance Transfer Learning) в условиях рассматриваемой задачи стохастической оптимизации представляет собой модификацию модификацию генетического программирования в символьной регрессии с измененным механизмом селекции наиболее приспособленных экземпляров.

2.1. Общая постановка задачи

В данной работе мы разрабатываем регрессионную модель $F(\overline{g}, \overline{c}; i)$ влияния генетических факторов $\overline{g}(i)$ на время цветения экземпляра дикого нута под индексом i с учетом погодных факторов $\overline{c}(i)$ в первые $N = 20$ дней после посадки. Данная функция подбирается в виде линейной комбинации нелинейных функций, задаваемых КС-грамматикой символьной регрессии. В данной работе все использованные элементарные нелинейные функции комбинируются из следующего перечня: $*, /, sin(\alpha), cos(\alpha), ln(\alpha), \frac{\alpha}{1+\beta^2}, \frac{\alpha}{\sqrt{1+\beta^2}}$, где α, β - некоторые факторы. Результирующая модель представляется в следующем виде:

$$F(\overline{g},\overline{c};i)=\sum_{j=0}^{M-1}\alpha_j * f_j\left(\overline{g},\overline{c};i\right) \quad (2.1)$$

где $i-$ индекс образца.

Результирующая модель прогнозирования обозначенного вида подбирается таким образом, чтобы минимизировать среднюю квадратичную ошибку MSE предсказания на данных как целевой, так и исходной области. Каким образом данная ошибка вычисляется, будет описано далее

2.2. Эволюционный процесс в трансферном обучении

Следующая блок-схема приближенно описывает работу алгоритма в процессе оптимизации и движение потоков данных внутри него:

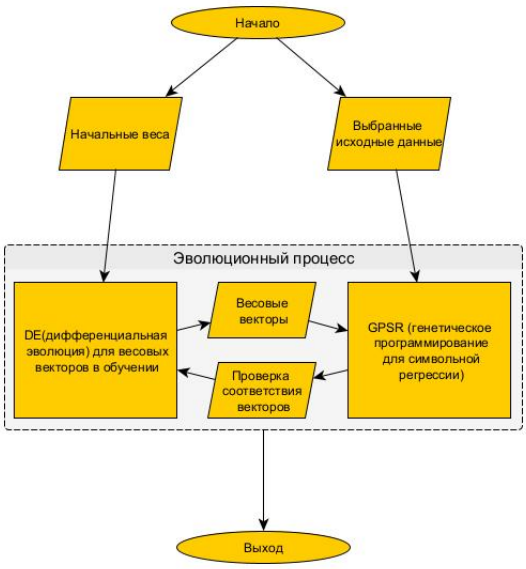

Рис.2.1. Общая схема работы алгоритма

Стоит отметить, что в оригинальной статье [3] на этапе инициализации начального взвешивания элементов исходной области используется подход *kernel mean matching*, который очень подробно описывается. В данной же работе в качестве начальной популяции векторов берутся случайные, нормально распределенные значения на отрезке от 0 до 1, в итоге получаем m векторов вида $w=\left(w_1,w_2,\ldots,w_{n_s}\right)$, где n_s - число значений исходной области, используемых в процессе оптимизации.

После того, как начальные наборы моделей и весовых векторов заданы, начинается эволюционный следующий цикличный процесс:

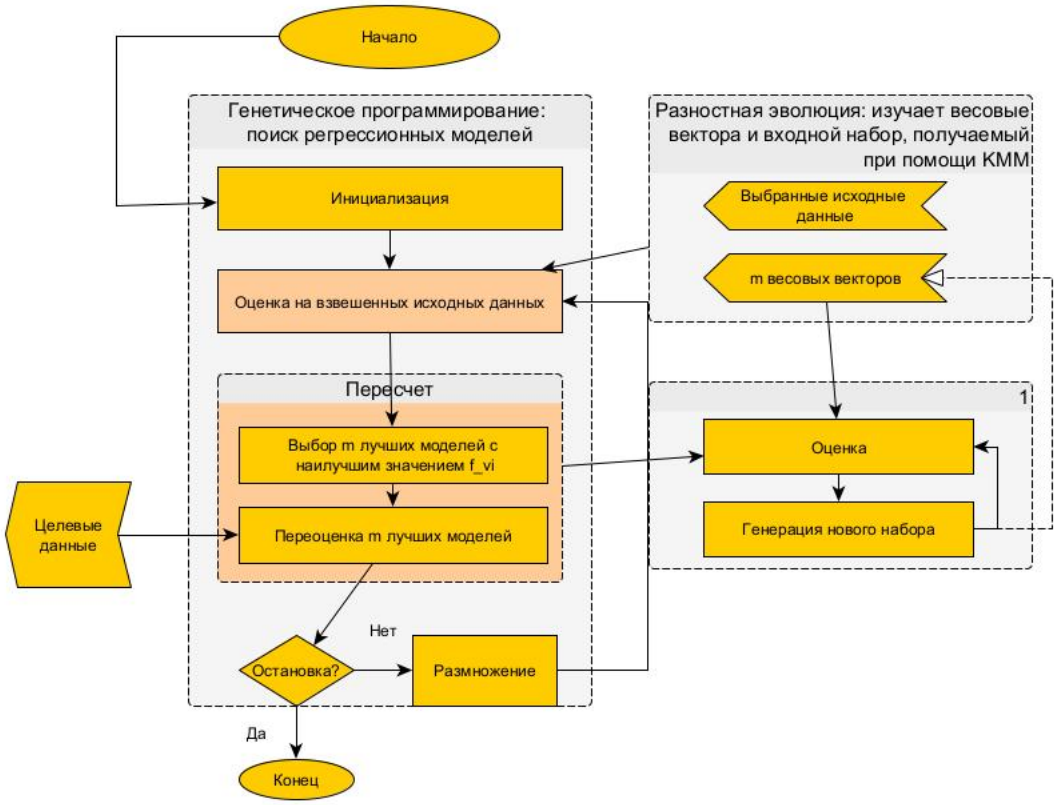

Рис.2.2. Подробное описание процесса оптимизации ITGP

A. Строится взвешенная оценка каждой модели для каждого весового вектора на всех данных исходной области.
B. Для каждого весового вектора выбирается соответствующая модель с наименьшим значением квадратичной ошибки.
C. Далее лучшие m моделей, полученных на предыдущем шаге, оцениваются уже на данных целевой области путем вычисления их среднеквадратических ошибок.
D. После этого весовые вектора и значения ошибок соответствующих им моделей на целевых данных передаются в разностную эволюцию, где происходит генерация нового набора весов для экземпляров исходной области.
E. Далее происходит выбор оставшихся $p - m$ моделей при помощи модифицированного оператора селекции, и очередные p моделей поступают на вход алгоритму символьной регрессии для построения новго поколения модели. Таким образом, основной алгоритм оптимизации вновь переходит к шагу A

Данный цикл прерывается либо по достижении лимита итераций, либо по достижении определенного значения ошибки прогнозирования.

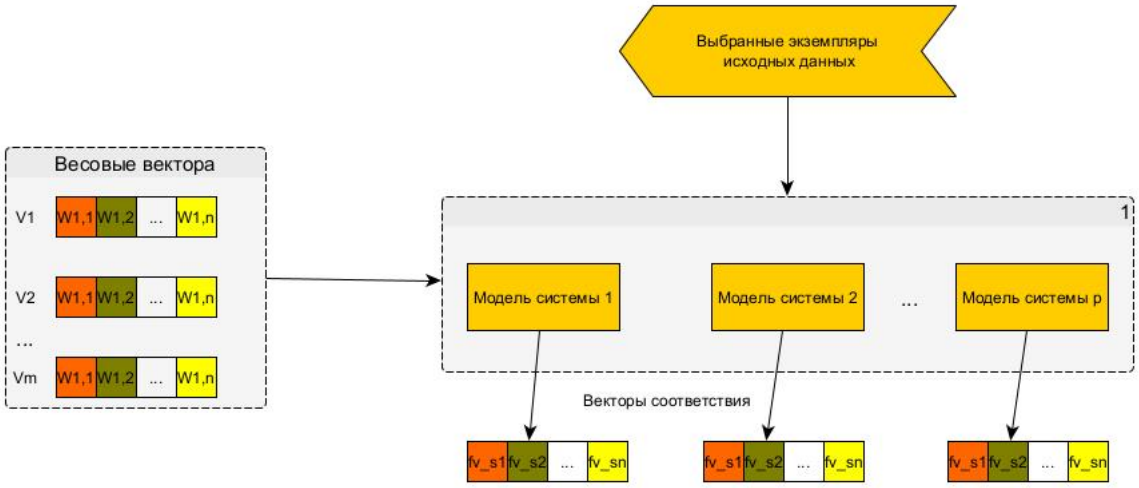

Рис.2.3. Схема связи моделей и весовых векторов в вычислении взвешенных ошибок момедей

Среднеквадратическая ошибка (MSE) и взвешенная квадратическая ошибка (WMSE) рассчитываются по следующим формулам:

$$WMSE = fv_i^s = \sum_{j=1}^{n_s} w_{ij}\left(f_j - y_j\right)^2, i = \overline{1,m} \tag{2.2}$$

где n_s - число экземпляров исходной области, m - размер популяции весовых векторов, w_{ij} - вес $j-$го экземпляра исходной области в $i-$ом векторе популяции весов, f_j - прогноз модели для $j-$го представителя исходной области, y_j - реальное время цветения растения-экземпляра.

$$MSE = fv^T = \frac{1}{n_t}\sum_{j=1}^{n_t}\left(f_j - y_j\right)^2 \tag{2.3}$$

где n_t-размер целевой области обучающих данных.

2.3. Реализация генетического программирования в символьной регрессии

Основной из двух компонент алгоритма следует считать символьную регрессию, так как непосредственно данный программный блок отвечает за подбор экземпляров-моделей, которые будут в дальнейшем прогнозировать время цветения дикого нута.

Символьная регрессия, как и любой генетический алгоритм, основывается на 4 важных эволюционных процессах - наследовании, мутации, скрещивании и естественном отборе.

Общий алгоритм строится следующим образом:

A. Случайным образом строится начальная популяция моделей-деревьев, представляющих собой набор функций от исследуемых факторов.
B. Далее в цикле повторяются следующие действия либо в течение фиксированного количества поколений, либо до достижения некоторого требуемого критерия качества представителей новой популяции:
    1. Последовательно выбираем каждого индивида из полученных в прошлом поколении алгоритма.
    2. Случайным образом происходит одна из трех мутаций - замена случайного листа (фактора) на другой фактор/случайную константу, замена узла (операции) на другую, соответствующую по арности или случайный обмен поддеревьями с другим индивидом из текущей популяции. Полученный мутировавший индивид добавляется в пробную популяцию.
    3. Когда популяция требуемого размера получена, происходит селекция наиболее приспособленных образцов. Вариантов это сделать достаточно - путем турнирной селекции определенного размера, при помощи попарного сравнения индвидов, стоящих в популяциях на идентичных позициях, и т. д.
    4. Если условие прекращения эволюционного процесса не выполнено, возвращаемся к шагу 1 в цикле, но уже с новым поколением, подверженным селекции.
C. По достижении предельного числа итераций алгоритма/достижения требуемого качества индивидов в популяции, алгоритм прекращает свою работу. Результатом может считаться как вся популяция индивидов на последнем поколении, так и лучший ее представитель с соответствующим значением критерия качества (как правило - некоторой целевой функции $F(\overline{x})$).

Реализация символьной регрессии, выбранная в качестве фундамента для создания модификации под трансферное обучение, использовался алгоритм, описанный в статье [10]. В данной реализации индивиды-деревья представляются именно в виде схем из терминальных символов в противовес классическому представлению в виде последовательности генов, каждый из которых кодирует один терминальный символ некоторого алфавита Σ (если описывать данную структуру в терминах Ахо [11]). Прогноз каждой из таких моделей для некоторого экземпляра нута при помощи синтаксического перевода данной схемы в цепочку терминальных

символов, а затем и в арифметическое выражение, которое уже после подстановки всех необходимых факторов дает предсказание.

В статье [3] компонента символьной регрессии использует для генерации моделей набор лишь из базовых арифметических операций - $+, -, *, /$. В рамках разработки модификации данного метода [12] был значительно расширен имевшийся алфавит, добавлены функции квадратного корня $\sqrt{\alpha}$, аналитических дробей $\frac{\beta}{1+\gamma^2}$ и $\frac{\beta}{1-\gamma^2}$, оператор домножения на константу $const * \omega$, где $\alpha, \beta, \gamma, \omega$ - некоторые цепочки из терминалов / нетерминалов над алфавитом Σ. Помимо них были задействованы уже реализованные, но не использованные в рамках тестирования ITGP в статье [3] тригонометрические функции $cos(\alpha)$, $sin(\alpha)$, натульный логарифм $ln(\alpha)$ и усложненная производная арктангенса $\frac{\alpha}{\sqrt{1+\beta^2}}$. Таким образом был расширен диапазон вариации функций.

2.4. Реализация разностной эволюции

Разностная эволюция — разновидность генетического алгоритма, использующаяся для многомерной стохастической оптимизации. Данный метод имитирует естественные эволюционные процессы, как и генетическое программирование в символьной регрессии.

Разностная эволюция используется в алгоритме для взвешивания экземпляров исходной области знаний, и, как следствие, для определения наиболее важных растений с точки зрения обучения моделей прогнозирования для применения на экземплярах целевой области.

Общий алгоритм разностной эволюции в целом совпадает с описанным ранее алгоритмом символьной регрессии за исключением формы представления индивидов в популяции. Здесь, как правило, индивиды - вектора с вещественными значениями.

В качестве модификации классического алгоритма разностной эволюции был переработан под решаемую задачу метод DEEP, подробно описанный в статье [13]. Основное, что стоит отметить - в нем помимо описанных выше механизмов мутации и скрещивания добавлены циклы пересчета вероятностей замены отдельных генов между индивидами и мощностей мутации генов с каждым новым поколением.

Пересчет вероятностей скрещивания выполняется по следующим формулам:

$$p_k = \begin{cases} -\left(NP * S_k^2 - 1\right) + \sqrt{\left(NP * S_k^2 - 1\right)^2 - NP * \left(1 - \rho_k\right)} & \text{если } \rho_k \geqslant 1 \\ p_{inf} & \text{если } \rho_k < 1 \end{cases} \quad (2.4)$$

Константы, масштабирующие мутацию индивидов, расчитываются по закону:

$$S_k = \begin{cases} \sqrt{\frac{NP*\left(\rho_k - 1\right) + p_k*(2-p_k)}{2*p_k*NP}} & \text{если } cond \geqslant 0 \\ S_{inf} & \text{если } cond < 0 \end{cases} \quad (2.5)$$

$$cond = NP * \left(\rho_k - 1\right) + p_k * (2 - p_k)$$

где NP - размер популяции индивидов, $p_{inf} = 0, S_{inf} = \frac{1}{\sqrt{NP}}$, а ρ_k - параметры вариации, связывающие две прошедших итерации алгоритма. Пересчет констант масштабирования и вероятностей мутации происходит через одну итерацию в порядке очередности.

2.5. Связующие алгоритмические компоненты метода

2.5.1. Селекция в ITGP

Селекция производится на заключительном этапе итерации построения нового поколения генетического алгоритма, после выбора лучшей модели для каждой популяции весовых векторов $w_1, w_2, \ldots, w_{n_s}$.

У нас уже имеются m оцененных на данных целевой области лучших моделей, каждая из которых соответствует одному весовому вектору. Требуется из общего числа моделей выбрать оставшиеся $p - m$ моделей для генерации нового поколения индивидов нужного размера.

Механизм селекции проходит в 3 этапа:

A. Случайным образом выбираются D измерений, по которым будет производиться турнирная селекция лучших моделей.

B. Применяется турнирная селекция. На каждом шаге алгоритма в качестве участников очередного турнира случайным образом выбираются T моделей и сравниваются между собой.

C. Из T участников выбирается модель, которая показывает показатели пригодности в выбранных D измерениях лучшие или аналогичные остальным

участникам турнира. Пригодность оценивается на взвешенных ошибках $fv_1^2, fv_2^s, \ldots, fv_m^s$ для каждой пары "модель-весовой вектор".

Выбор D измерений для оценки моделей производится при помощи *рулеточной селекции*, при этом вероятность выбрать каждый из весовых векторов равняется нормированному значению среднеквадратической ошибки лучшей модели, соответствующей данному вектору.

ГЛАВА 3. РАЗРАБОТКА, НАСТРОЙКА И ИСПОЛЬЗОВАНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Следует также остановиться на технической сторон работы. Основной алгоритм трансферного обучения был реализован согласно концепции метода, описанного в статье [3]. Модификация заключается в использовании модифицированных компонент символьной регрессии и разностной эволюции по сравнению с теми, которые были использованы для сравнения в оригинальном материале.

3.1. Реализация метода разностной эволюции

Одной из двух главных компонент использованного для разработки математической модели метода, как уже было упомянуто в прошлой главе, является разностная эволюция. Данный метод был адаптирован на языке программирования Python версии 3.10 на основе реализации, описанной в статье [13], с открытым исходным кодом. Фактически проводилась адаптация реализации метода DEEP под трансферное обучение. Оригинальная реализация метода на язике программирования *C* находистя в свободном доступе по ссылке [14].

3.2. Реализация генетического программирования в символьной регрессии

Для реализации ITGP, помимо разностной эволюции, использовался модифицированный алгоритм генетического программирования в символьной регрессии, который был адаптирован под решаемую задачу на основе программной интерпретации данного метода с открытым исходным кодом [12]. Метод разработан на языке программирования Python с применением библиотек numpy и scipy. Для адаптации метода под задачи данной работы было расширено множество функций,

используемых в эволюционном процессе, а также доработаны классы фитнес–функции и процедура генерации нового пробного поколения моделей-деревьев для ускорения вычислений.

3.3. Техническое обеспечение работы

Предобработка данных, разработка самой модели, численные эксперименты (промежуточное тестирование модели, кросс-валидация на имеющихся данных и т. д.) и анализ полученных результатов производились в среде программной разработки PyCharm на языках программирования Python и R.

В качестве вычислительного средства использовался пользовательский компьютер со следующими параметрами:

- Процессор: Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz 2.30 GHz
- Число ядер/потоков: 8/16
- Оперативная память: 12Гб

Каждый шаг процесса кросс-валидации производится на отдельном ядре вычислительной машины.

ГЛАВА 4. ПРИМЕНЕНИЕ РАЗРАБОТАННОГО МЕТОДА ОПТИМИЗАЦИИ ДЛЯ ПОИСКА МОДЕЛИ ПРОГНОЗИРОВАНИЯ

Реализация описанного ранее алгоритма обучения с переносом знаний производилась для достижения второй важной цели работы - разработки модели прогнозирования времени цветения дикого нута. В качестве факторов построения прогноза модель использует климатические и генетические данные, использованные для исследований в статьях [15] и [2].

4.1. Описание данных исследуемых объектов

Согласно статье [2], из которой были взяты данные для проведения исследований, данные содержат генетическую информацию для двух видов дикого нута, выращенных в 5 различных регионах Турции и 4 регионах Австралии, - *Cicer reticulatum* и *Cicer echinospermum*. Перечень имеющихся данных покрывает

широкий диапазон различных климатических зон, в частности особое значение имеет серьезное различие высот, на которых выращивались исследуемые образцы.

Для проведения исследований было повторено разделение данных тем же образом, как это было сделано в статье [2]. В исходную область были отправлены 2174 экземпляра, в то время как целевую область составили оставшиеся 2088 образца нута. В первую группу были отнесены данные по растениям, посаженным в осенний период (270, 290, 294, 305, 325 и 339 дни календарного года сооветственно), оставшиеся экземпляры были посажены в весенний период. Время цветения использованных растений - в пределах от 117 до 221.

Данные по каждому образцу составляются из двух частей - генетической $\overline{g}$ и клитматической $\overline{c}$.

В генетическую часть $\overline{g}$ входят информации о наличии/отсутствии каждого из 6 ассоциированных ОНП, выбранных как наиболее сильно влияющих на время цветения исследуемых сортов нута. В данных представляются для каждого растения в виде 18 бинарных признаков, каждый из которых определяет наличие для каждого SNP одного варианта из трех комбинаций аллелей: AA (Alternative/Alternative), AR (Alternative/Reference) или RR (Reference/Reference).

Климатические данные $\overline{c}$ представляют собой параметры погоды за первые 20 дней после посадки, каждый из которых описывается набором из следующих 5 факторов:

A. $rain$ - уровень осадков в выбранный день, в мм
B. $daylength(dl)$ - длина светового дня за выбранные сутки, в часах
C. s_{rad} - количество дневной солнечной радиации, в $\frac{\text{кВт}}{\text{м}^2}$
D. T_{min} - минимальная температура за выбранные сутки, в градусах
E. T_{max} - максимальная температура за выбранные сутки, в градусах

Суммарно можно сказать, что на время цветения нута как в рамках данной работы, так и в исходной статье [2], зависит от 119 факторов, последним из которых является метка региона посадки выбранного экземпляра.

Ниже приводится небольшой фрагмент датасета, как он представлен в формате csv-таблицы:

geo_id	b'snp1AA'	b'snp1AR'	b'snp1RR'	...
0.0	1.0	0.0	0.0	...
2.0	1.0	0.0	0.0	...
2.0	0.0	0.0	1.0	...

Таблица 4.1

Представление нуклеотидных данных в табличном виде

...	t_{min0}	t_{max0}	dl_0	$srad_0$	$rain_0$	...
...	8.9	16.05	10.10591224	1.8	0.0	...
...	2.0	1.0	10.97075033	0.0	0.0	...
...	2.0	0.0	0.0	0.0	0.0	...

Таблица 4.2

Представление погодных данных в табличном виде

...	$srad_{19}$	$rain_{19}$	response
...	0.0	10.22655597	114
...	0.0	10.22655597	131
...	0.6	9.597001465	115

Таблица 4.3

Завершающая часть набора данных

4.2. Описание параметров исследований

Для поиска подходящей модели прогнозирования производится кросс-валидация реализованного метода трансферного обучения на имеющихся данных со следующими параметрами:

A. Размер популяции моделей в каждом эксперименте - $p = 300$.
B. Размер популяции весовых векторов в каждом эксперименте - $m = 50$.
C. Предельное допустимое число поколений работы алгоритма - 100.
D. Параметры символьной регрессии:
E. Размер турнира в турнирной селекции - $T = 4$
F. Число измерений, используемых в качестве критериев пригодности участников турниров - $D = 10$
   - Вероятность скрещивания частей ветвей модели - 0.9.
   - Вероятность мутации терминалов в модели-дереве - 0.1.
   - Вероятность мутации нетерминалов в модели-дереве - 0.1.

- Используемые функции: $+, -, *, /$,*const*, *cos*, *sin*, *ln*, *sqrt*, $\frac{A}{\sqrt{1+B^2}}, \frac{A}{1+B^2}$, где $A, B-$ некоторые цепочки, состоящие как из терминалов, так и нетерминалов.
- Максимальная и минимальная глубина моделей - 4 и 12 соответственно.

G. Параметры разностной эволюции:

- Значения весов экземпляров исходной области - в пределах от 0 до 1.
- $\gamma = 0.9$.
- $p_{inf} = 0.0, p_{sup} = 1.0$.
- $S_{inf} = \frac{1}{\sqrt{m}} \approx 0.1415, S_{sup} = 2.0$, где $m-$ размер популяции весовых векторов.

4.3. Анализ данных

Прежде чем начинать строить модель, необходимо проанализировать, насколько сильно отличаются группы значений каждого из рассматриваемых факторов. В случае, если группы мало отличимы между собой, то есть принята нулевая гипотеза о равентсве средних данных выборок, нет смысла использовать их всех для прогнозирования.

Проверка проводится путем применения теста Тьюки для имеющегося набора данных. Алгоритм теста заключается в следующем:

A. Поочередно исследуем влияние каждого фактора (в данном случае таких 119, 18 аллельных, а остальные - климатические) на время цветения образцов.

B. Попарно берутся все группы значений факторов, для данной пары вычисляется критерий Тьюки:

$$q = \frac{\overline{x_A} - \overline{x_B}}{SE}$$

$$SE = \sqrt{\frac{MS_\omega}{2}\left(\frac{1}{n_B} + \frac{1}{n_A}\right)}$$

где A, B - исследуемые группы значений фактора, $\overline{x_A}, \overline{x_B}$ - их выборочные средние, n_A, n_B - размеры данных групп, MS_ω - внутригрупповая дис-

персия, или средневзвешенное арифметическое групповых дисперсий, вычсляемое по формуле:

$$MS_{\omega} = \frac{\sum_{i=1}^{N} \sigma_i^2 n_i}{\sum_{i=1}^{N} n_i}$$

где $N-$ число групп значений исследуемого фактора.

C. Далее, если полученное значение попадает в область отклонения нулевой гипотезы с уровнем значимости $\alpha = 0.05$, то H_0 отвергается, в дальнейшем полагаем, что данные группы значений имеют разные распределения генеральных совокупностей. В противном же случае принимаем гипотезу H_0 и полагаем данные группы выборками из одной генеральной совокупности.

В результате применения данного алгоритма было получено, что нулевая гипотеза H_0 была отвергнута для всех погодных факторов и территориальной метки места посадки образца, группы значений данных факторов распределены неодинаково, и соответственно представляют наибольший интерес для использования в прогнозе. Отличный от этого результат был получен при исследовании генетических признаков образцов, использованных для измерения времени цветения: одинаковое распределение наличия /отсутствия аллелей AR и AA показали все 6 нуклеотидных полиморфизмов, в то время как группы образцов с присутствующей /отсутствующей аллелью RR каждого из ОНП распределены по разным случайным законам.

4.4. Процесс кросс-валидации

Для построения моделей на разных частях сисходных и целевых данных производятся следующие манипуляции:

A. Как целевые, так и исходные данные делятся на 4 фолда случайным образом.

B. Далее выбираются 3 из 4 фолдов (также случайным образом) в качестве обучающих наборов целевых и исходных данных, оставшиеся четверти остаются на валидацию.

C. Алгоритм производит оптимизацию на выбранных обучающих наборах данных, после чего происходит валидация популяции моделей на

отобранном на прошлом шаге части датасета, отдельно функции валидируются на целевом и исходном наборах данных.

D. Модель, показавшая лучшие результаты в процессе валидации, выбирается как итог работы алгоритма.

Описанные шаги повторяются 100 раз, после чего производится оценка результатов по критерию Манна-Уитни. Необходимо добиться, чтобы распределения значений ошибок прогнозирования на тренировочных и валидационных данных совпадали по обозначенному критерию с уровнем значимости $\alpha = 0.05$.

4.5. Результаты экспериментов

Параметры проведенной кросс-валидации:

- Число фолдов - 4.
- Количество проведенных экспериментов - 50.
- Размер одного фолда экземпляров исходной области - 543.
- Размер одного фолда экземпляров целевой области - 108.

Результаты проведения кросс-валидации данных представлены в виде следующих гистограмм:

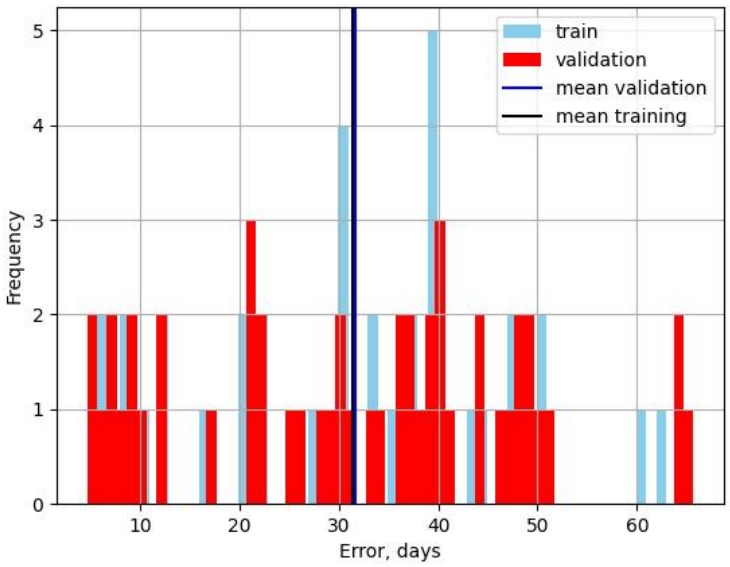

Рис.4.1. Статистическая диаграмма ошибок моделей, полученных в ходе кросс-валидации данных исходной области

Значени и $p - value$ статистики Манна-Уитни на данных, продемонстрированных на гистограммах, имеют следующие значения:

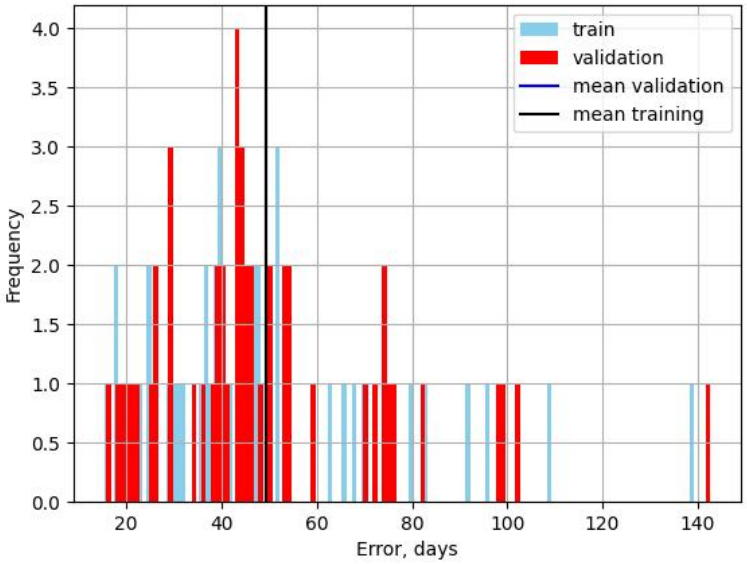

Рис.4.2. Статистическая диаграмма ошибок моделей, полученных в ходе кросс-валидации данных целевой области

– Для целевых (target) данных: $stat = 1242, p_{value} \approx 0.9587$
– Для исходных (source) данных: $stat = 1235, p_{value} \approx 0.92037$

Видно, что значения $p - value$ сильно превышает уровень значимости $\alpha = 0.05$ для обеих областей данных, следовательно гипотеза H_0 отвергается в обоих случаях, и распределения ошибок прогноза на тренировочном и валидационном наборах одинаковы как для исходного, так и для целевого домена.

Столбцы, расположенные наиболее близко к началу координат графика, считаются лучшими, соответственно наиболее удачными считаем модели, соответствующие данным показателям.

В целом видно, что средняя ошибка прогнозирования как для данных исходной, так и целевой области, достаточно велика и находится в пределах месяца. Но лучшие индивиды показывают достаточно хорошие результаты, так как их ошибки прогнозирования находятся в пределах двух недель.

Ниже приводятся лучшие модели, показавшие наименьшую ошибку на обучающих и валидационных данных, в таблие приводятся показатели ошибок рассматриваемых функций:

$$F(\overline{x}) = 14.49109 * dl_4 - 2.438 * t_{min18} + 5.94384 * log(Abs(
(log(Abs(t_{min18}/sqrt(dl_9^2 + 1) - 0.766/(dl_4^2 * t_{min15}^2 + 1)))-
1.605)/(rain_{13}^2 + 1))) + 5.94384 * log(Abs(log(Abs((dl_7+
t_{max1}/sqrt(rain_3^2 + 1))/(t_{min13} * t_{min18})))+
4.966)) + 5.94384 * cos(41.27313 * dl_4 - 6.94384 * t_{min18} + 14.49109*
log(Abs(dl_9)) + 2.438 * cos(rain_{13}) + 14.49109 * cos(5.94384 * dl_4 - t_{min18}+
2.438 * log(Abs(dl_9)) + 2.438 * cos(rain_{13})) + 14.09983/((1 + 0.58676/
(t_{min13}^2 * cos(rain_{13})^2 + 1)^2) * log(Abs(log(Abs(t_{max1})/Abs(sqrt(rain_3^2 + 1)))))))+
5.94384 * cos(41.27313 * dl_4 - 6.94384 * t_{min18} + 14.49109 * log(Abs(dl_9))+
2.438 * cos(rain_{13}) + 14.49109 * cos(5.943844 * dl_4 - t_{min18}+
2.438 * log(Abs(dl_9)) + 2.438 * cos(41.27313 * dl_4-
6.94384 * t_{min18} + 14.4911 * log(Abs(dl_9)) + 2.438 * cos(0.766/(t_{min10}^2 * t_{min15}^2 + 1))+
14.4911 * cos(5.94384 * dl_4 - t_{min18} + 2.438 * log(Abs(dl_9)) + 2.438 * cos(rain_{13}))+
14.09983/((1 + 0.58676/((rain_3^2 + 1)^2 * (t_{min15}^2 * log(Abs(t_{min13}))^2+
1)^2)) * log(Abs(log(Abs(t_{max1})/Abs(sqrt(rain_3^2 + 1)))))))) + 14.09983/((1+
0.58676/(dl_4^2 * t_{min15}^2 + 1)^2) * log(Abs(log(Abs(t_{max1})/Abs(sqrt(rain_3^2+
1)))))))+5.78336/((0.00476*dl_4^2/(((log(0.26269*Abs(t_{max1})/Abs(sqrt(0.069-dl_4^2)))+
log(Abs(dl_9)) + cos(5.94384 * dl_4 + 2.438 * log(Abs(dl_9)) + 2.438 * cos(rain_{13})-
0.766))^2 + 1)^2 * ((-rain_{16} + 0.16824 * t_{min18} - 0.41017 * log(Abs(dl_9))-
0.41017 * log(Abs(log(Abs(t_{max1})/Abs(sqrt(log(0.64045*
Abs(t_{max1})/Abs(sqrt(0.41017-
log(Abs(dl_9))^2)))^2 + 1))))) - 0.41017 * cos(cos(rain_{13})) - 0.3991/(1+
0.58676/(rain_{13}^2 * t_{min15}^2 + 1)^2))^2 + 0.02831)^2) + 1) * log(Abs(log(Abs(t_{max1})/
Abs(sqrt(rain_3^2 + 1)))))) \quad (4.1)$$

$$F(\overline{x}) = 13.41024*dl5+7.324*srad_8/(t_{min0}^2+1)+3.662*cos(dl_3+cos(3.662*cos(dl_7+ cos(0.04803 * t_{min0})) - 3.662 * cos(dl_5 + 3.662 * rain_2 * (13.41024 * dl_7^2 * snp_{3RR}^2/ (snp_{1AR}^2 * (cos(log(Abs(log(Abs(log(sqrt(Abs(dl_3)))))))))^2 + 1)^2) + 1)/srad_8- sqrt(Abs(t_{min0} + 4.38))))) + 3.662 * cos(3.662 * dl_{10} + cos(srad_8) + 4.671)+ 3.662*cos(3.662*dl_{10}-cos(0.069/(423.07258*(-0.65197*dl_{10}-0.65197*rain_{17}/srad_4- 1+0.19707 * (1-3.662 * t_{max19}^2)/t_{max19})^2 * log(Abs(3.662*rain_2*(t_{min0}^2+1)/srad_8- sqrt(Abs(t_{min0} + 4.38))))^2 + 1)) + 5.6168) + 3.662 * cos(3.662 * dl_{10}+ cos(3.662*dl_0-13.41024*dl_{10}+3.662*t_{max0}-3.662)+4.671)+3.662*cos(3.662*dl_{10}+ cos(13.41024*dl_{10}+3.662*dl_7-3.662*srad_6+3.662*t_{min19}+3.662*cos(dl_5-dl_7)+ 8.04963 + 10.9091/(0.27307 * snp_{6AA}2/((cos(13.41024 * dl_{10} - 3.662 * dl_7+ 3.662 * t_{min19} + 3.662 * cos(dl_5 - dl_7) + 9.80741 + 10.9091/ (1.13689 * snp_{6AA}^2/(rain_9^2 + 1)^2 + 1))^2 + 1)^2 * (Abs(t_{min12}/ (((rain_{18} + srad_2)^2 + 1) * (cos(cos(3.662 * dl_{10} + dl_5 - dl_7 + t_{min19}+ cos(dl_5-dl_7)-1.9798))^2+1)))+0.27307))+1))+4.671)+3.662*cos(-3.662*dl_{10}+ 3.662 * rain_2 * (t_{min0}^2 + 1)/srad_8 + cos(t_{min0}) + 2.35657) + 17.1052 \quad (4.2)$$

$$F(\overline{x}) = -0.10253 * (-0.376 * srad_{10} * t_{max4}/(cos(t_{max1})^2 + 1) - 0.376 * t_{max17} - 0.376 * t_{max17} * cos(cos(0.376 * t_{min5}))/(cos(t_{max1})^2 + 1) - 1.5149 * t_{max2} + 0.376 * log(Abs(0.376 * rain_{13}/sqrt(t_{min5}^2 + 1) + 0.376 * cos(0.30351 * (0.12702 * dl_6 + 0.12702 * rain_1 + 0.51175 * snp_{1AR} + 0.12702 * t_{max17} + 0.12702 * t_{max17} * cos(cos(0.376 * srad_{19} * t_{min14} * (dl_4 + t_{min17}) + 0.376 * log(0.376 * Abs(srad_{11})) - 0.376 * sqrt(Abs(dl_{13} * dl_3))/sqrt(Abs(sqrt(snp_{3RR}^2 * srad_{14}^2 + 1)))))/(cos(t_{max1})^2 + 1) + 0.51175 * t_{max2} - 0.12702 * t_{min17} - 0.12702 * log(Abs(0.376 * rain_{13}/sqrt((srad_{19} * t_{min14} * (0.376 * rain_{13}/sqrt((dl_{16} - Abs(dl_{13} * dl_3)^{1/4}/Abs(sqrt(snp_{3RR}^2 * srad_{14}^2 + 1))^{1/4})^2 + 1) + 0.376 * cos(t_{min17})) - log(0.376 * Abs(srad_{11})) + sqrt(Abs(dl_{13} * dl_3))/sqrt(Abs(sqrt(snp_{3RR}^2 * srad_{14}^2 + 1))))^2 + 1) + 0.376 * cos(t_{min17}))) + 0.12702 * sqrt(Abs(dl_{19})) + 0.12702 * sqrt(Abs(t_{max0})) - 0.63357 + 0.12702 * (dl_{16} + geo\_id) * sqrt(Abs(log(Abs(t_{max1})))/sqrt(1 - 0.376 * t_{min18}^2)) * Abs(srad_7)/Abs(sqrt(rain_1^2 + 1))))) + 0.376 * log(Abs(cos(t_{min17}))) - 0.376 * sqrt(Abs(dl_{19})) - 0.376 * sqrt(Abs(rain_4)) - 0.376 * Abs(log(Abs(rain_{12})/Abs(sqrt((-0.376 * snp_{6AA}/sqrt(rain_1^2 + 1) + t_{max17} * log(Abs(rain_{12})))^2 + 1))))^{1/4} + 2.2133 - 0.376 * (geo\_id + t_{min5}) * sqrt(Abs(log(Abs(t_{max1})))/sqrt(1 - 0.376 * t_{min18}^2)) * Abs(srad_7)^{3/2}/(sqrt(Abs(sqrt(1 - 0.376 * cos(sqrt(Abs(log(Abs(srad_2))))))^2))) * Abs(sqrt(rain_1^2 + 1))) \quad (4.3)$$

во всех приведенных функциях $\overline{x} = (x_0, x_1, \ldots, x_{118})$ - многомерная переменная, включающая в себя все факторы как независимые компоненты.

	Error, source (days)	Error, target (days)
(4.1)	10.5488	14.6116
(4.2)	8.2062	20.1637
(4.3)	6.8441	20.1472

Таблица 4.4

Наиболее точные модели прогнозирования, полученные в ходе кросс-валидации

Для каждой модели были проведены исследования влияния каждого из 119 факторов, участвовавших в генерации моделей, при помощи *теста перестановок*

Алгоритм теста:

A. Перебираются все предикторы-столбцы исходного датасета.
B. Значения выбранного столбца-фактора случайным образом перемешиваются, други изменений датасета не производится.
C. Считается предсказание модели для каждого экземпляра датасета с учетом одного перемешанного фактора.
D. Вычисляем ошибку - насколько сильно изменилось абсолютное значение предсказания модели.
E. Шаги Б-Г повторяются 100 раз, по ним считается среднее значение изменения прогноза по всем перестановкам и по всем растениям при некорректном значении данного фактора.

Функция	(4.1)	(4.2)	(4.3)
log	27	12	10
sin	0	0	0
cos	9	6	19
$sqrt$	16	27	33
$+$	59	48	38
$-$	60	76	38
$*$	51	27	46
$/$	29	26	18
pow	24	20	18

Таблица 4.5

Использование элементарных функций в полученных моделях

По итогам проверки влияния каждого из факторов, входящих в модели, были получены следующие нормированные гистограммы изменений прогнозов:

По гистограммам первоначально делаем вывод, что температурные показатели играют ключевую роль в прогнозировании времени цветения нута в экспериментально полученных моделях. Зависимость от аллелей SNP хоть и присутствует в двух из трех моделей, но влияние данных Факторов на предсказание минимально.

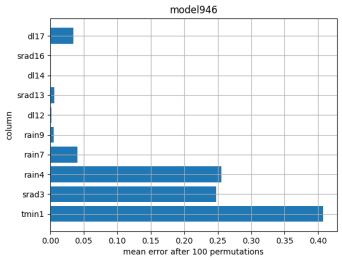

Рис.4.3. Гистограмма влияния признаков, использующихся в модели (4.1)

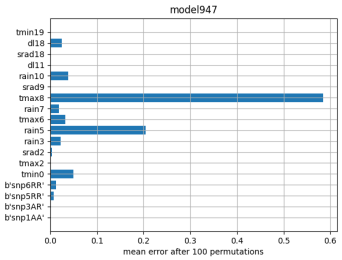

Рис.4.4. Гистограмма влияния признаков, использующихся в модели (4.2)

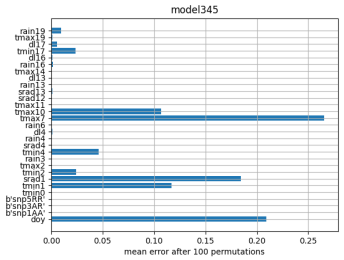

Рис.4.5. Гистограмма влияния признаков, использующихся в модели (4.3)

4.6. Проверка модели на устойчивость к климатическим изменениям

В ходе подбора наиболее подходящей модели прогнозирования для описанного перечня факторов была получена куда более точная модель, дающая ошибку прогноза ≈ 11.5 дней на всем имеющемся наборе данных.

Параметры ITGP, использованные в процессе стохастической оптимизации:

- Максимальные и минимальные количества уровней деревьев - 4 и 13.
- Вероятности скрещивания моделей/мутации листьев/мутации узлов - 0.8 & 0.2 & 0.2.
- Размер популяции весовых векторов - $m = 60$.
- Размер популяции моделей - $p = 300$.
- Максимальный размер дерева - 320 узлов и листьев.
- Использованные в ходе эволюции элементарные функции - $+, -, *, /, const$.

Полученная модель имеет следующий вид:

$$F(\overline{x}) = dl_1 * t_{min5} - dl_1 * (dl_8 - srad_4 - t_{max2} + t_{min8})/((-1.478 * rain_{17} - 0.974002/(rain_{13} + t_{min11})) * (srad_9 - t_{max8}/t_{max13} - (rain_{13} - t_{min8})/srad_{12})) - rain_{12} - rain_{13} * t_{max5} * t_{min11}/(t_{max0} * t_{min8}) + t_{max4} + 0.659/(-dl_1 * (dl_8/srad_4 - t_{max2} + t_{min8})/((-1.478 * rain_{17} - 1.478 * t_{min10}) * ((rain_{11} + rain_{13} + (3.9 * rain_{10} * rain_{17}/t_{max11} - 3.9 * snp_{2AA} + 3.9 * t_{min12})/(t_{min5} * (rain_{13} + t_{min11}))) * (-srad_{12} + t_{min7} + 2.507) - (srad_{12} - srad_{15} - 2.109) * (t_{max5} - t_{min6} + 1.733) - 1.648 + (srad_{11} + srad_8 + (rain_{11} + rain_{13} + (3.9 * dl_5 * rain_{17}/t_{max11} + 15.21)/(t_{min5} * (rain_{13} + t_{min11}))) * (-srad_{12} + t_{min7} + 2.507))/dl_1)) + rain_{12} + snp_{6AA} * t_{max5} * (0.90827 * dl_{16} * srad_2 + dl_8 - 0.95459)/(t_{min8} * (t_{min11} + 0.659/(rain_{12} + rain_{13} * t_{max5} * t_{min8}/(srad_{11} * t_{max0}) - srad17 + tmin11 - t_{min7} * (-1.648 * t_{max5} + 1.648 * t_{min6} - 2.855984)/(rain_{14} * (-rain_{13} + srad_{12} - srad_{15}) * (t_{max5} - t_{min6} + 1.733) * (-dl_{15} + (dl_{19} - t_{min12} - 4.89) * (rain_{14} + rain_{17} + (-dl_{17} + 0.13928 * rain_{16}/(dl_{19} - 4.111 * rain_{11} * t_{max7} - 4.111 * rain_6 * snp_{5RR} + snp_{5AA}^2 - t_{max12} * t_{max16} * (dl_8 * (rain_{13} + rain_5)/t_{max0} - 1.03) - t_{min12} + 3.403) - 0.214)/(t_{min11} * t_{min5})) - 1.648 + (srad_{11} + srad_4 + srad_8)/dl_1))))) - srad_{17} + t_{min11}) \quad (4.4)$$

Тестирование модели с помощью перестановок значений факторов дает следующие результаты:

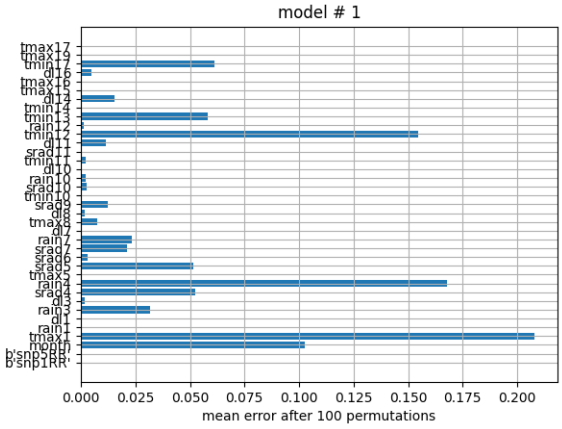

Рис.4.6. Нормированная гистограмма влияния признаков в модели (4.4)

Стоит отметить, что данная модель вычислялась на модифицированном датасете, который в качестве факторов содержал год и месяц посадки образцов нута, поэтому данный предиктор появился в итоговой модели. В остальном видно, что так же, как в моделях, полученных путем кросс-валидации, на первый план выходят максимальные температурные показатели за исследуемые дни.

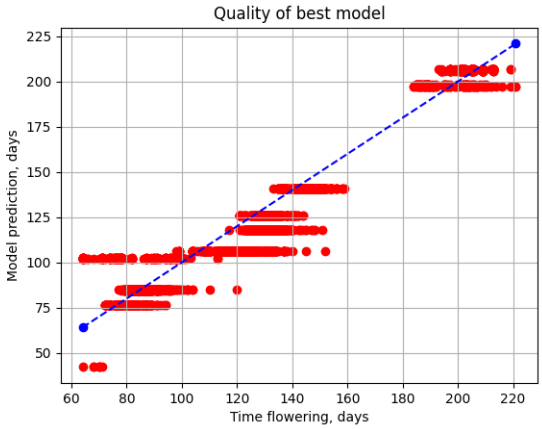

Рис.4.7. Сравнение прогноза модели (4.4) с точными значениями времени цветения

4.7. Проверка устойчивости модели к климатическим изменениям

Чтобы проверить, насколько актуальной останется полученная модель в условиях динамически меняющейся климатической ситуации в мире, попробуем внести изменения в имеющиеся климатические данные. Будем проверять изменение прогноза в условиях трех экспериментов:

- Имитация глобального потепления - значения максимальной и минимальной дневных температур изменяются на полижительную величину от 0 до 2 градусов по Цельсию с шагом 0.1 в каждый из 20 дней, по которым модель вычисляет предполагаемое время цветения.
- Симуляция засухи в начале периода цветения. Данный эксперимент производится путем увеличения солнечной радиации на величину от 0 до 3 $\frac{\text{кВт}}{\text{м}^2}$ с шагом 0.1.

4.7.1. Проверка в условиях глобального потепления

При вычислении прогноза в условиях глобального потепления для каждого растения было получено более 15 тысяч графиков, с которыми также можно

ознакомиться в репозитории [16]. Все полученные кривые можно разделить на группы, характерные примеры приводятся ниже:

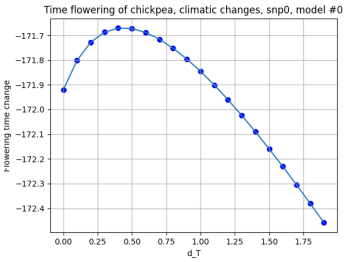

Рис.4.8

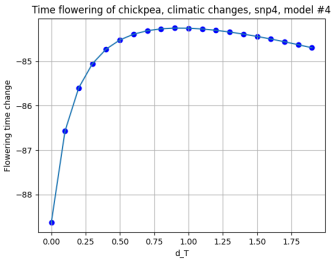

Рис.4.9

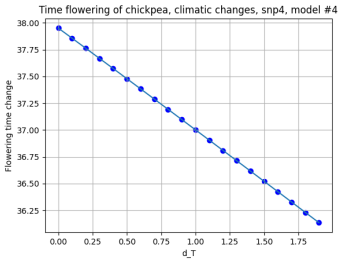

Рис.4.10

На представленных графиках (4.8), (4.9), (4.10) видно, что во многих случаях изменение времени цветения хоть и уменьшается, но либо значение прогноза уходит в отрицательные значения, либо изменения критически большие. Кроме того, для многих образцов не удается предсказать даже физически возможное время цветения. На основании этих показателей можно сделать предположение о том, что модель не может применяться при изменении температурных показателей в соответствующий сезон в будущем. Косвенно подтверждают данное предположение показатели тестов перестановок, согласно которым максимальная и минимальная суточные температуры имеют значительное по сравнению с остальными факторами влияние на прогноз модели, поэтому их малое изменение ведет к неустойчивости модели.

4.7.2. Проверка в условиях имитации засухи

В результате прогнозирования при фиксированном уменьшении количества осадков получены также 15000 графиков, иллюстрирующих отсутствие какого-либо однозначного влияния изменения данного фактора на прогноз:

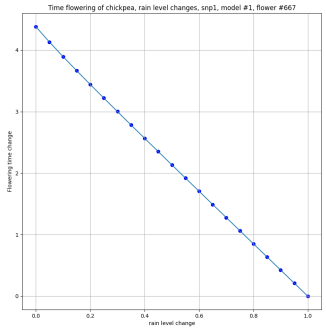

Рис.4.11

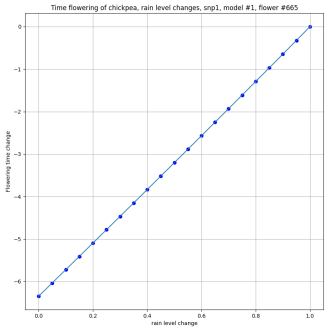

Рис.4.12

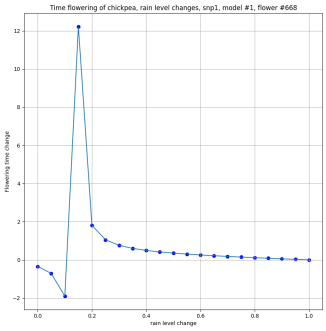

Рис.4.13

На графиках (4.11), (4.11) и (4.13) приведены наиболее характерные кривые изменения прогноза времени цветения образцов при уменьшении числа осадков в первые 5 дней после посадки на величину от 0 до 100% от начального объема с шагом в 5%. Некоторые результаты отличаются лишь положением кривых вдоль оси ординат, что объясняется переменчивым начальным скачком изменения предсказания у разных образцов. Наблюдаются как монотонные линейные убывания и возрастания времени цветения с уменьшением числа осадков (графики (4.11), (4.11)), так и константное изменение на некоторую величину с разрывом в определенном значении (график (4.13)), последнее свидетельствует о некоторой гиперболической зависимости от объема осадков. Следовательно, делаем вывод, что хоть отклонения прогноза находятся в пределах реальных значений, модель не является устойчивой в условиях частичной засухи и требует доработки.

4.7.3. Проверка в условиях увеличения солнечной активности

В результате прогнозирования при фиксированном увеличении солнечной радиации были еще раз получены 15000 графиков, иллюстрирующих характерное возрастание времени цветения нута вида:

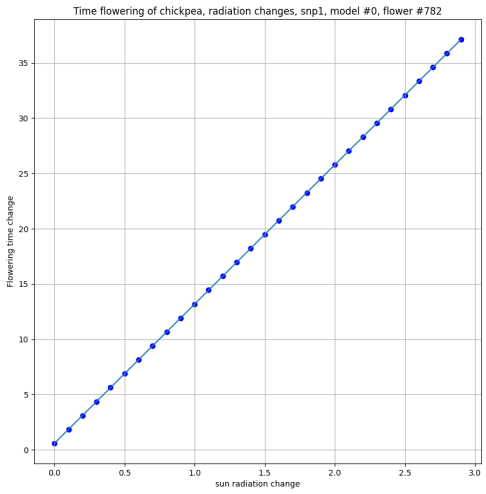

Рис.4.14. Изменение прогноза времени цветения моделью (4.4) при увеличении получаемой солнечной радиации в первые 5 дней после посадки образца

Для всех образцов наблюдается подобная линейная зависимость, кривые отличаются друг от друга только положением относительно оси ординат, изменение прогноза происходит в разумных пределах, максимальное изменение прогноза при максимальном увеличении солнечной радиации составляет 85 дней и в целом

редко превышает 30 дней. На основании этого можно выдвинуть предположение о том, что модель в целом может применяться в условиях повышения интенсивности солнечного излучения до определенного уровня.

4.8. Сравнение с моделями из использованных источников

В статье [2] модели получаются схожим способом с тем, который используется в данной работе - стохастической оптимизацией с применением символьной регрессии и разностной эволюции. Обучение моделей проводилось на большем числе факторов, а также число шагов в кросс-валидации вдвое превышало проведенные в рамках ВКР эксперименты, поэтому и качество полученных индивидов выгодно смотрится на фоне полученных результатов.

Показатели ошибок лучших экземпляров, полученных в процессе кросс–валидации, статистически описывается гистограммой:

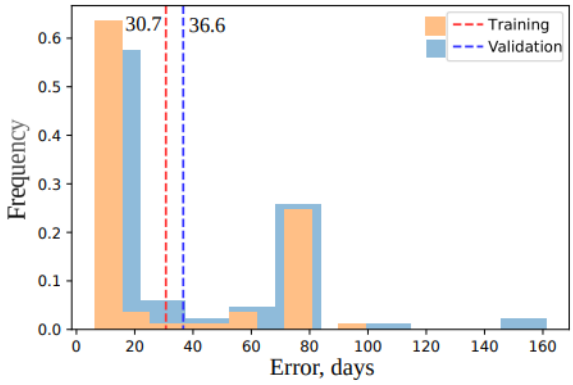

Рис.4.15. Результаты кросс-валидации, проведенные в статье [2]

По гистограмме и приведенным в статье численным результатам видно, что лучшие индивиды в среднем допускают ошибку прогноза в два раза меньшую, чем лучшие из полученных моделей (6.085 против 14.6116). Аналогичный вывод можно сделать касательно сравнения средних значений ошибок лучших моделей в обих исследованиях (36.6 и 30.6 против 49.812 и 48.113 соответственно). Однако стоит уточнить, что модель в оригинальной статье обучается постепенно и большее количество дней, параллельно получая все новые климатические данные и оценивая степень созревания растений, в то время как полученная модель использует фиксированный набор параметров, ограниченный первыми 20 днями после посадки экземпляров. Это обеспечивает более ранний прогноз временной границы достижения вегетационного периода (пусть и менее точный) и, как следствие,

позволяет более четко спланировать дальнейший план действий по завершении цветения.

4.9. Доступ к результатам работы

Все конечные и промежуточные данные, полученные в ходе работы, расположены в репозитории Github, обратиться к которому можно по ссылке [16]. Можно ознакомиться с графиками сходимости алгоритма в ходе проведения кросс-валидации, гистограммами зависимости прогнозов моделей от факторов, гистограммами результатов, а также некоторыми промежуточными результатами - перестановками, использованными при исследовании влияния факторов на предсказания модели, всеми моделями (в формате для чтения и в сериализованном виде, для загрузки в программу и дальнейшего исследования), которые были получены в ходе всех экспериментов.

ЗАКЛЮЧЕНИЕ

В ходе проведенной работы был изучен и реализован метод машинного обучения с переносом знаний, связывающий символьную регрессию с разностной эволюцию. Для этого были исследованы и доработаны классические реализации данных компонент трансферного обучения с учетом как рационального использования временных и пространственных ресурсов, так и достижения высокого качества получаемых моделей прогнозирования.

После теоретической работы реализованный алгоритм был использован для прогнозирования времени цветения нута на имеющихся данных в ходе проведения кросс-валидации. В целом результаты, полученные в ходе экспериментов, в несколько раз уступили результатам, представленным в статье.

По результатам проведения экспериментов была проведена визуализация и сравнительный анализ данных, для чего были изучены и применены соответствующие методы.

Выводы, сделанные по результатам работы:

- Реализованные надстройки над классическими методами генетических алгоритмов учитывают больше статистических условий в процессе обучения.
- Полученные модели прогнозирования в целом хуже разработанных и описанных в рассмотренной статье, но с учетом упрощения данных, используемых в качестве факторов предсказания, оказались достаточно приемлимыми.
- Также к достоинствам полученных моделей можно причислить тот факт, что они используют куда меньше факторов, влияющих на объекты, в сравнении с моделями, полученными в статье [2]. Несмотря на то, что прогноз оказывается не очень точным, его можно сделать на более ранних стадиях созревания посаженных образцов.
- Реализованный метод достаточно быстро и качественно работает с большими объемами данных.
- В лучших полученных моделях отмечается сильное влияние температурных характеристик и длины светового дня, в то время как влияние генетических различий незначительно и слабо влияет на прогноз.

СЛОВАРЬ ТЕРМИНОВ

DE — Differential Evolution — Разностная эволюция, метод многомерной стохастической оптимизации.

SRGP — Symbolic Regression in Genetic Programming — Генетическое программирование для символьной регрессии.

ITGP — Instance Transferring Genetic Programming — Генетическое программирование с передачей экземпляров.

ОНП — однонуклеотидный полиморфизм (также обозначается как SNP, Single Nucleotide Polymorphism) - различие ДНК-последовательностей гомологичных участков хромосом величиной в 1 нуклеотид A|C|G|T.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Abbo S.*, *Berger J.*, *Turner N. C.* Viewpoint: Evolution of cultivated chickpea: four bottlenecks limit diversity and constrain adaptation // Functional Plant Biology. — 2003. — Vol. 30. — P. 1081–1087. — DOI https://doi.org/10.1016/S1161-0301(02)00108-9.

2. Simulation Model for Time to Flowering with Climatic andGenetic Inputs for Wild Chickpea / A. Ageev [et al.] // Agronomy, MDPI journal division. — 2021. — DOI https://www.mdpi.com/article/10.3390/agronomy11071389/s1.

3. *Chen Q.*, *Xue B.*, *Zhang M.* Genetic Programming for Instance Transfer Learning in Symbolic Regression // IEEE Transactions on cybernetics. — 2020. — DOI https://ieeexplore.ieee.org/document/9007621.

4. Knowledge Transfer Through Machine Learning in Aircraft Design / A. T. W. Min [et al.] // IEEE Computational Intelligence Magazine. — 2017. — DOI https://ieeexplore.ieee.org/document/8065136.

5. *Gupta A.*, *Ong Y.-S.*, *Feng L.* Multifactorial Evolution: Toward Evolutionary Multitasking // IEEE Transactions on Evolutionary Computation. — 2016. — DOI https://ieeexplore.ieee.org/document/7161358.

6. Evolutionary Multitasking via Explicit Autoencoding / L. Feng [et al.] // IEEE Transactions on Cybernetics. — 2019. — DOI https://ieeexplore.ieee.org/document/8401802.

7. Putting mechanisms into crop production models / K. J. Boote [et al.] // Plant, Cell Environment. — 2013. — Vol. 18. — P. 1–13. — DOI https://onlinelibrary.wiley.com/doi/10.1111/pce.12119.

8. Modeling chickpea growth and development: Leaf production and senescence / A. Soltani [et al.] // Field Crops Research. — 2006. — Vol. 99. — P. 14–23. — DOI https://doi.org/10.1016/j.fcr.2006.02.005.

9. An overview of APSIM, a model designed for farming systems simulation / B. Keatinga [et al.] // European Journal of Agronomy. — 2003. — Vol. 18. — P. 267–288. — DOI https://doi.org/10.1016/S1161-0301(02)00108-9.

10. *Virgolin M.* Genetic programming is naturally suited to evolve bagging ensembles // ACMDL, Digital Library. — 2018. — DOI https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-019-1685-2.

11. *Ахо А. В.*, *Ульман Д. Д.* Теория синтаксического анализа. Т. 1. — Москва, И-110, ГСП, 1-й Рижский пер., д.2: Издательство "Мир", 1972. — С. 104—123, 163—176. 613 с. — (Сер.: 4).

12. SRGP Implementation. — URL: https://github.com/marcovirgolin/SimpleGP (visited on 28.05.2022).

13. *Kozlov K.*, *Samsonov A. M. S.*, *Samsonova M.* A software for parameter optimization with Differential Evolution Entirely Parallel method // PeerJ Computer Science. — 2016. — DOI https://peerj.com/articles/cs-74/.

14. DEEP method implementation. — URL: https://gitlab.com/mackoel/deepmethod (visited on 28.05.2022).

15. Non-linear regression models for time to flowering in wild chickpea combine genetic and climatic factors / K. Kozlov [et al.] // BMC Plant Biology. — 2018. — DOI https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-019-1685-2.

16. Algorithm implementation, datasets parsing and analysis. — URL: https://github.com/YaroslavAggressive/DE-and-SRGP-in-Transfer-Learning/tree/master (visited on 01.06.2022).