НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»

УДК №2123132123
Регистрационный №123123
Инв. №

УТВЕРЖДАЮ

Директор ООО «Рога и Копыта»

_____ И.И.Иванов
«__»_____ 2013 г.

ОТЧЁТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

«Оптимизация обучающих выборок с помощью генетического алгоритма»

Этап №1.1 «Обзор современного состояния торсионных наногенераторов»
(промежуточный)

Руководитель НИР,
инженер _____ Д.Е. Катаев

г. Москва, 2013

Список исполнителей

Научный руководитель,
доцент К.К.Петров _____

с.н.с, к.т.н,
Ж.Ж. Балбесов, _____

Реферат

Отчет 20 с, 3 таблица, 7 рисунок, 1 источник.

Содержание

Введение

Одним из ключевых моментов в реализации систем, использующих алгоритмы машинного обучения с учителем является подготовка обучающей выборки. В настоящее время, как правило, синтез обучающей выборки выполняется экспертом.ъ, который, однако, далеко не для всех задач в состоянии за разумное время выбрать и представить исходные данные задачи в оптимальном для обучаемой системы виде. Таким образмо имеется потребонсть в системах, которые способны автоматически формировать обучающие выборки для интеллектуальных систем из имеющихся данных таким образом, чтобы их обучение по автоматически сформированным выборкам было максимально эффективным. Естественно, подобные системы востребованы только в тез случаях, когда либо не существует, либо не сформулирован алгоритм формирования обучающей выборки. Таким образом наиболее иодходящим выглядит использование алгоритмов случайного поиска, например генетических.

Цель данной работы разработка новой версии ПО для генетической оптимизации обучющих выборок систем машинного обучения с учителем. Модификация алгоритма для устранения избыточности обучающих данных. А так же оценка класса решаемых алгоритмом задач.

1 Постановка задачи

Основной задачей являлась разработка новой версии программного обеспечения для оптимизации обучающих выборок при помощи генетического алгоритма на основе старой с реализацией ряда модификаций:

а) Модификация формата вывода данных для облегчения взаимодействия с пользователем

б) Устранение избыточности обучающих данных для снижения вычислительной мощности

в) Модификация алгоритма мутации для реализации рекурсивной мутации

1.1 Описание исходной системы

Пусть $f(S_h) = (S_t; S_v))$ - отображение f исторических данных S_h в совокупность обучающей S_t и валидационной S_v выборок. Пусть $N(S_t; S_v)$ - искусственная нейронная сеть прямого распространения, обученная по S_t и проверяемая по S_v, а $Err_v(N(S_t; S_v))$ - ее ошибка валидации, определяемая, как среднеквадратичное отклонение всех выходов сети от соответствующих эталонных значений при подаче на вход элементов валидационной выборки. Тогда

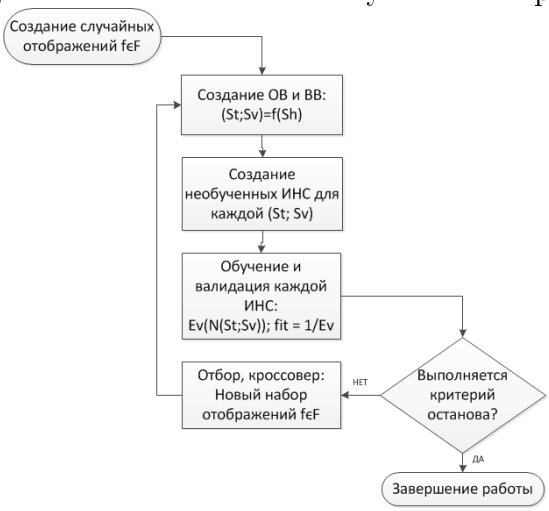
$$\exists \bar{f} : \bar{f}(S_h) = arg\{minErr_v(N(S_t; S_v)).\} \tag{1}$$

Требуется найти $\bar{f}$. Для этого реализован следующий алгоритм, представленный на рис. 1.1
Критерием останова в используемой реализации является прохождение заданного количества итераций. В текущей реализации используются ИНС прямого распространения в одним скрытым слоем, обучаемые методов эластичного распространения (RPROP)
Отображения f представляют собой конструкции следующего вида:

$$f(S_h) = \begin{vmatrix} (\lambda_{i_1} \circ \lambda_{i_2} \circ ... \circ \lambda_{i_m})(S_{h_{j_1}}) \\ ... \\ (\lambda_{i_{n-l}} \circ ... \circ \lambda_{i_n})(S_{h_{j_k}}) \end{vmatrix}, \lambda \in \Lambda, i,n,m,j \in N, \tag{2}$$

где Λ - множество доступных системе элементарных преобразований сигналов (дифференцирование, фильтры, добавление лагов и т.п.), S_{h_j} - произвольный сигнал из исторических данных. Атомарной единицей при кроссовере является $f(S_h)_i$, однако мутации возможны и внутри нее.

Рисунок 1.1 — Схема используемого алгоритма



При хранении экземпляр кодировался линейной последовательностью исполняемых команд и списком команд, результаты выполенения которых попадают в обучающую выборку. Такая структура данных не способствовала тонкой настройке алгоритмов кроссовера и мутации.

1.2 описание стоящих задач

1.2.1 Модификация структуры хранения данных

В силу тяжести взаимодействия пользователь с исходным форматом вывода данных было решено сменить формат вывода данных. Целями смены формата были: повышение человекочитаемости, повышение легкости транспортировки и модифицирования данных эксперимента,по возможности сохранение легкости загрузки данных в приложение. Возможными структурами хранения при решении данной задачи были xml, json, исполняемый код или база данных. К плюсам xml при решении поставленной задачи следует отнести следующее:

Таблицаasd

Средство	Читаемость	Транспортабельность	Скорость
XML	Читаемый в сыром виде формат. Для прочтения в удобном виде достаточно элементарного браузера	Для передачи достаточно просто переслать интересующее поддерево, или целиком файл.	Парисng требует затрат времени. Затраты, однако, можно сократить оптимизацией, основанной на жесткости структуры.

Продолжение таблицы 1.1

Средство	Читаемость	Транспортабельность	Скорость
JSON	Читаемость формата в сыром виде несколько затруднена. Требуется специальный просмотровщик.	Для передачи достаточно переслать интересующий объект, или целиком файл.	Парсинг требует минимального времени - фактически это преобразование из строки в код.
Исполняемый код	Читаемость формата в сыром виде требует специальных навыков. Желательна подсветка синтаксиса и знание структуры программы.	Для передачи данных достаточно передать цельный кусок кода, что однако тоже требует специальных навыков.	Парсинг требует минимального времени - фактически это преобразование из строки в исполняемый код.
База данных	Читаемость в сыром виде практически отсутствует как класс. Требуются специальные просмотровщики. Цельность данных с точки зрения просмотра теряется.	Передача данных как таковых требует либо объединенной БД для нескльких экземпляров системы, между которыми идет передача, либо выгрузки из базы данных	Скорость работы зависит от СУБД и скорости соединения, если база данных удалена, однако скорость работы любой базы данных на локальной машине - выше, чем скорость преобразования текстовых файлов.

По результатам выбора между представленными альтернативами был выбран формат данных xml.

1.2.2 Устранение избыточности обучающих данных

В предыдущей версии возникла проблема избыточности размера экземпляров, которая достаточно критично влияла на объем вычислений, производимых системой. Задачей является сокращение роста экземпляров, при сохранении возможности такого роста.

Устранить избыточность обучающих данных можно было при помощи нескольких модификаций:

а) Выкидывание на каждой мутационной фазе алгоритма одного элемента из экземпляра и проверка направления изменения фитнесс-функции

б) Жесткое ограничение на размер экземпляра

в) Изменения распределения вероятностей в пользу уменьшения, а не увеличения размера экземпляра

Вариант с жестким ограничением на размер экземпляра был отвергнут в связи с невозможностью в общем случае предугадать рельную правильную длину. Вариант с изменением алгоритма мутации был исключен в силу увеличения времени работы каждого шага почти в два раза. Реализуемым вариантом стал вариант с изменением распределения вероятностей. Конкретным изменением послужило увеличение возможного интервала размеров на следующем шаге в сторону уменьшения.

1.2.3 Модификация алгоритма мутации для реализации рекурсивной мутации

Для осуществления возможности вносить большее разнообразие в экземпляры и большей настраиваемости хода работы алгоритма требовалось введение возможности мутировать не только атомарные объекты но и структуры их объединяющие.

Для реализации рекурсивной мутации в каждой сущности, вплоть до атомарного элемента - функции ЦОС, была модифицирована процедура мутации. На фазе мутации каждая сущность верхнего уровня с заданной вероятностью мутировала полностью - генерировался полностью новый экземпляр этой сущности, в обратном же случае она вызывала мутацию всех своих сущностей-детей.

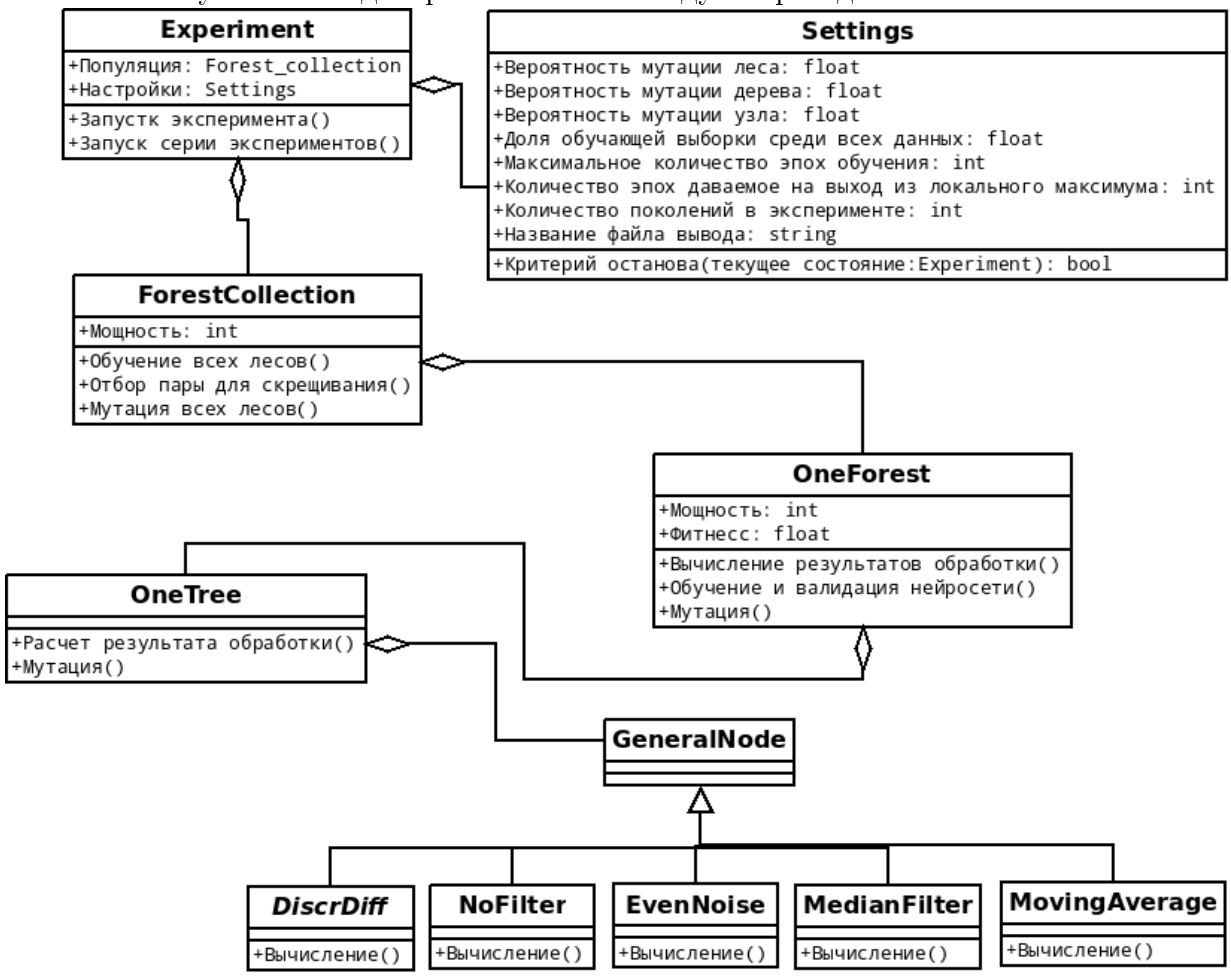
2 Описание результата разработки

Текущая версия ПО представляет из себя пакет скриптов на языке python версии 2.7, содержащий инструменты для проведения вычислений и анализа результатов.

Алгоритм по существу не менялся по сравнению с прошлой версией. Однако изменению подверглись его подалгоритмы и методы реализации.

Модуль вычислений описан в диаграмме классов 2.1

Рисунок 2.1 — диаграмма классов модуля проведения вычислений



Для инициализации вычислений требуется представить входные данные для обработки и входные данные для вычислительного модуля. Стандартным методом получения данных для обработки является набор csv файлов со значениями. Входные данные для вычислительного модуля фактически представляют из себя экземпляр класса настроек. Выходные данные вычислительного модуля представлены в XML формате, и выгружаются в файл с настраиваемым названием.

Модуль обработки результатов вычислений представляет из себя набор скриптов, позволяющих собирать информацию о ходе и результатах генетического алгоритма как с одного, так и с набора XML файлов. По результатам работы строит csv файл содержащий в себе выбранные и обработанные значения.

Используемые библиотеки: PyBrain - для создания и обучения искуственных нейросетей. SciPy и NumPy - для ускорения вычислений.

3 Эксперименты

3.1 Используемые тестовые сценарии

Для тестирования использовалось два типа систем - линейная и нелинейная. Под сценарием эксперимента подразумевается набор из входных и целевых данных для нейросети.

3.1.1 Линейная система

Пусть существует бассейн, который является объектом управления. К нему подведены две трубы: через одну в бассейн может поступать вода с произвольной скоростью (в данном тестовом сценарии не будут учитываться физические ограничения), через вторую трубу вода может уходить из бассейна с произвольной скоростью. Через эти две трубы осуществляется управление бассейном. Ограничения на то, что объем воды в бассейне должен быть неотрицательным нет, но, стоит отметить, что в данном сценарии не возникает соответствующей ситуации.

Пусть существуют следующие экспериментальные данные:

$X^{(0)}$ – объем поступившей за Δt воды;

$X^{(1)}$ – объем откачанной за Δt воды;

$X^{(2)}$ – объем воды, находящейся в бассейне в данный момент времени;

$X^{(3)}$ – посторонний сигнал, отличающийся от остальных;

Реально бассейн реагирует на управляющие воздействия следующим образом:

$$X_i^{(2)} = X_{i-1}^{(2)} + X_{i-1}^{(0)} - X_{i-1}^{(1)}; X_0^{(2)} = 10 \tag{3}$$

Пусть испытуемая система не имеет априорной информации об объекте управления и имеет доступ к следующим данным в качестве исходных:

$X^{(0)\Delta t}$ – объем поступившей за Δt воды, на этот сигнал наложен импульсный шум (inflow_noizd);

$X^{(1)\Delta t}$ – объем откачанной за Δt воды (outflow);

$X^{(3)}$ – посторонний сигнал (trash).

3.1.2 Нелинейная система

Пусть $X^{(0)}, .., X^{(5)}$ – действительные временные ряды

из которых $X^{(3)}, .., X^{(5)}$ не оказывают на систему никакого реального воздействия.

$X^{(6)}$ - выход системы

Реально система реагирует на управляющие воздействия следующим образом:

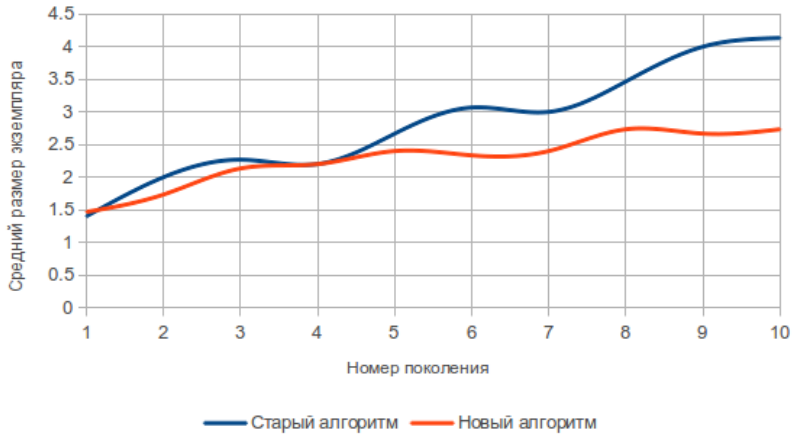
$$X_i^{(6)} = |X^{(0)} - X^{(1)}| + X^{(2)}; \tag{4}$$

Перед подачей в вычислительный модуль на $X^{(0)}$ был наложен постоянный шум с амплитудой 0.5, на $X^{(1)}$ наложен шум с вероятностью 2.5% прибавляющий 1 и с такой же - вычитающей (условно такой шум можно считать импульсным). Остальные сигналы зашумлению не подвергались.

3.2 Устранение избыточности

В качестве эксперимента для проверки работоспособности решения по устранению избыточности было решено провести пару тестов на одинаковых входных данных для старого и нового алгоритмов кроссовера. В качестве метрики для сравнения был выбран размер экземпляра на определенной итерации, усредненный среди нескольких экспериментов, для устранения возможности влияния случайных факторов.

Рисунок 3.1 — график средних размеров экземпляров по поколениям для старого и нового алгоритма



Средняя производная по таким графикам различается в два раза в пользу новой версии алгоритма кроссовера, что говорит о решении данной поставленной задачи по крайней мере в ограниченном объеме.

3.3 Модернизация алгоритма мутации

Для оценки качества модификации алгоритма мутации было проведено равное количество экспериментов при одинаковых начальных условиях для старого и нового алгоритма мутации, для разных вероятностей мутации. Для моделирования старого алгоритма (Плоский случай) новому задавались нулевыми вероятности мутации всех элементов кроме атомарных - функций ЦОС. Для моделирования аналогичной вероятности с использованием рекурсивной мутации вероятности выставлялись одинаковыми и равными приведенной вероятности рассчитаной по формуле. (5)

$$P_n = \sqrt[3]{p-1} + 1 \tag{5}$$

Где p - вероятность для плоского случая (приведенная вероятность), а P_n - аналогичная ей вероятность для рекурсивного случая.

На основе экспериментов проведенных с линейной системой построены графики приведенные на рисунках 3.2 и 3.3. На основе графика 3.3 можно сказать, что после реализации новой версии мутации уменьшилось время сходимости алгоритма, а так как на графике 3.2 видно, что фитнесс-функция оптимального экземпляра не только не ухудшилась но и улучшилась при таком переходе, можно сказать что мы не столкнулись с явлением преждевременной сходимости.

Рисунок 3.2 — средняя фитнесс-функция для приведенных вероятностей мутации

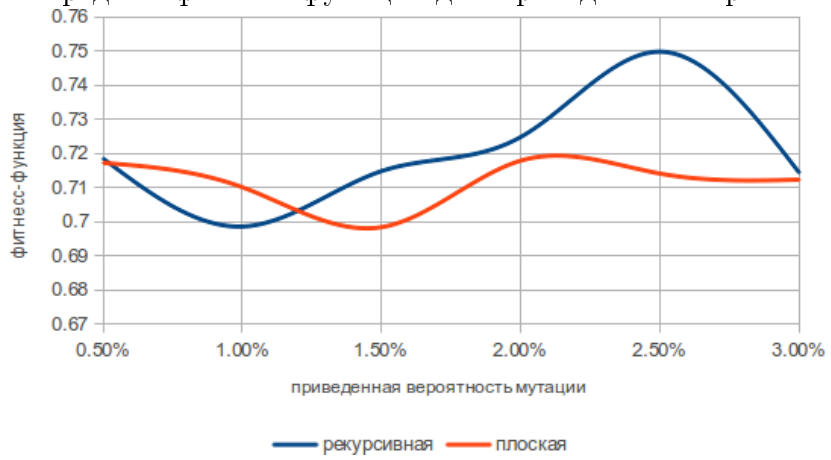
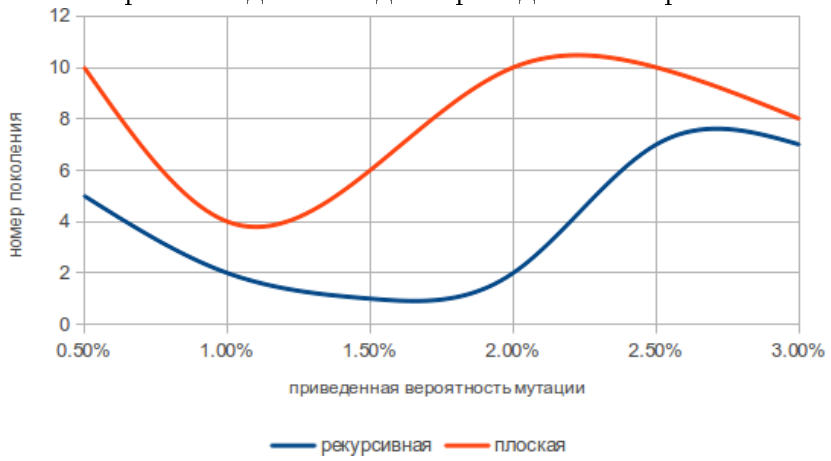


Рисунок 3.3 — время сходимости для приведенных вероятностей мутации



На основе экспериментов проведенных с нелинейной системой построены графики приведенные на рисунках 3.4 и 3.5.

Рисунок 3.4 — средняя фитнесс-функция для приведенных вероятностей мутации

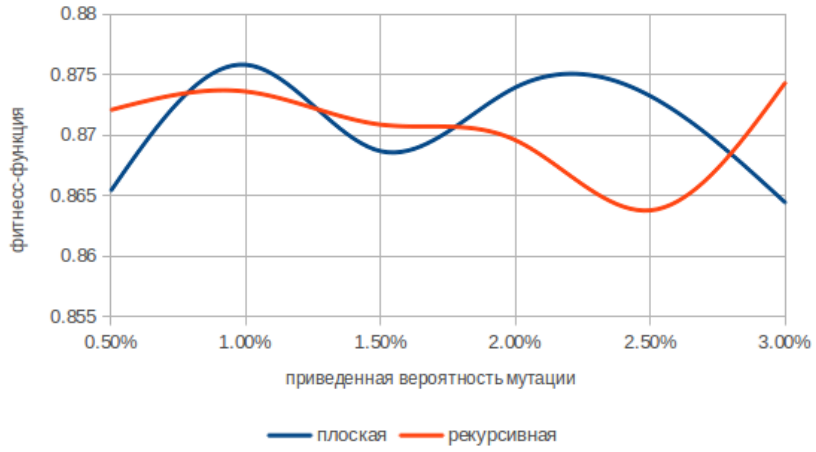
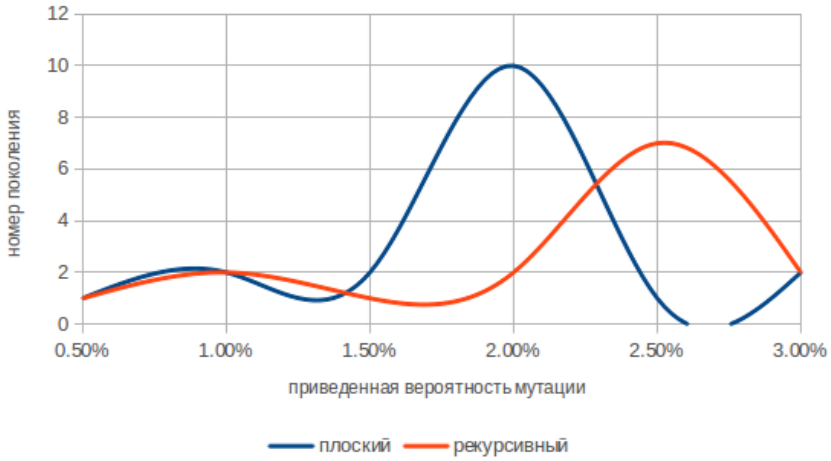


Рисунок 3.5 — время сходимости для приведенных вероятностей мутации



4 Постановка задачи

В прошлой версии программного обеспечения, реализованного в дипломной работе научного руководителя присутствовала избыточность обучающих данных и слабая сопровождаемость кода.

Основной задачей была реализация новой версии ПО на python с попутным внесением технических и алгоритмических изменений Основной задачей было избавление от избыточности обучающих данных. В качестве метода решения этой задачи была выбрана модификация алгоритма кроссовера, а точнее смещение вероятности при выборе размера нового экземпляра в сторону уменьшения. В качестве критерия оценки избыточности была выбрана следующая метрика: средний размер экземпляра в популяции { на последней итерации || на момент максимума || в течение всех поколений }

В качестве оценок качества кода и сопровождаемости были выбраны следующие эксперименты, метрики и критерии:

а) Валидация PEP8 & PEP257

б) Проведение ряда экспериментов работавших в прошлой версии ПО, как средство проверки правильности практической реализации алгоритма

в) Сравнение характеристик сходимости при старом и новом алгоритме мутации

5 Решение поставленных задач

В прошлой версии программного обеспечения, реализованного в дипломной работе научного руководителя присутствовал ряд проблем:

а) Избыточность обучающих данных вызванная алгоритмом кроссовера.

б) Слабая сопровождаемость кода.

6 разнообразная хуита из шаблона.

В отчётах могут быть и таблицы - см.табл. 6.1.

Таблица 6.1 — Пример таблицы

Название 1	Название 2	Название 31
Это	пример	данных
помещённых	внутрь	таблицы

Заключение

Список использованных источников

1 Грицанов А.А. и др. Новейший философский словарь. Мн.: Книжный Дом., 2003.