Оптимизация обучающих выборок с помощью генетических алгоритмов

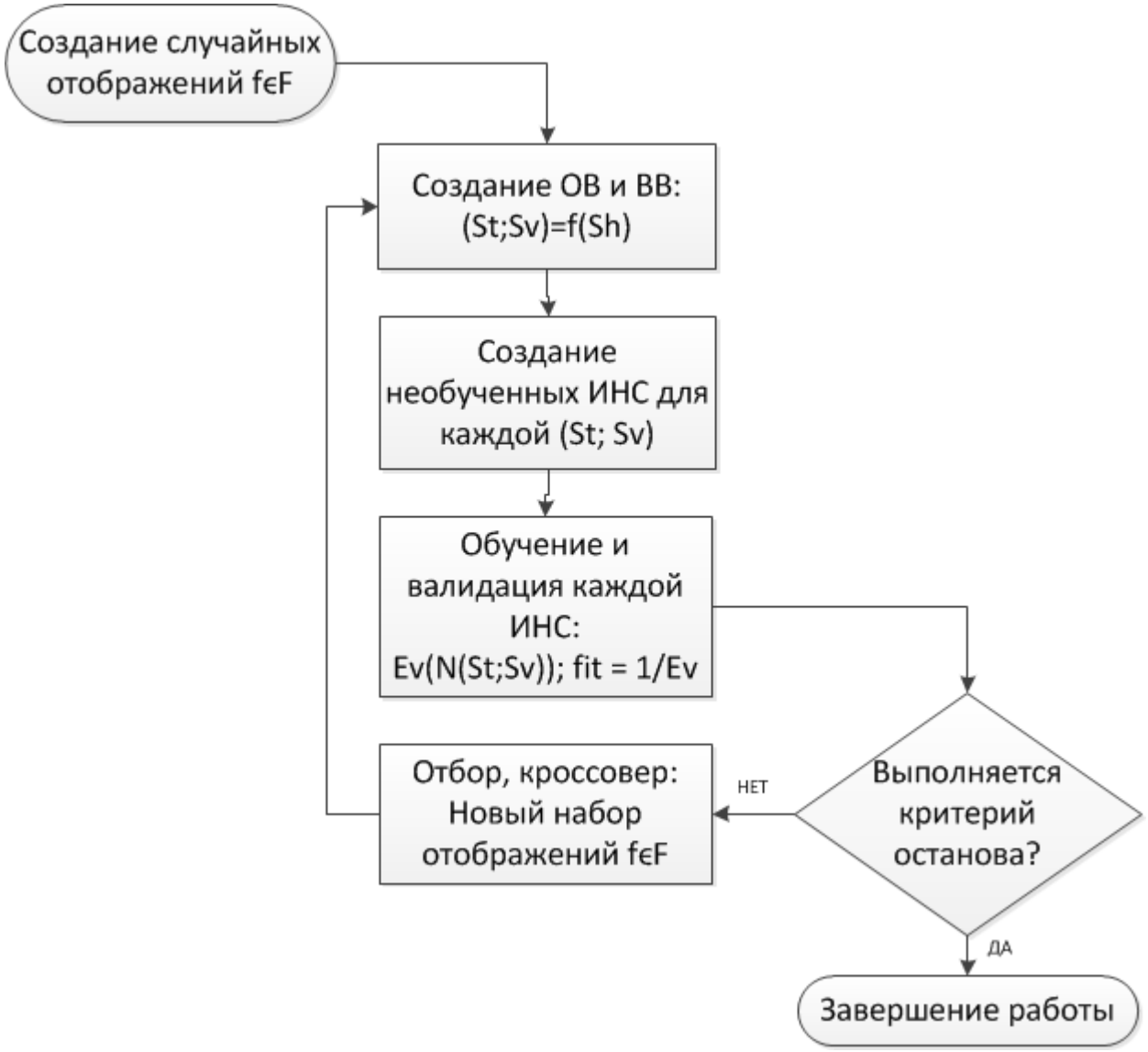
Построение обучающей выборки

- $\exists\overline{f}\colon\ \overline{f}(S_h) = \arg\{\min Err_v(N(S_t; S_v))\}$
- Отбор данных
- Определение и устранение шумов
- Вспомогательные преобразования
  - Нормировка
  - Обобщающие преобразования
  - etc

Пути решения

- Труд эксперта
  - Временные затраты и качество результата слабо поддаются прогнозированию
  - Требуется априорная информация о задаче
  - Человеческий фактор
  - Существует ряд методов построения ОВ и метрик качества ОВ
- Алгоритмы случайного поиска
  - Огромная вычислительная сложность
  - Легко распараллеливается
  - Возможно прогнозировать временные затраты
  - Не требуется никакой априорной информации
  - Возможна полностью автономная работа

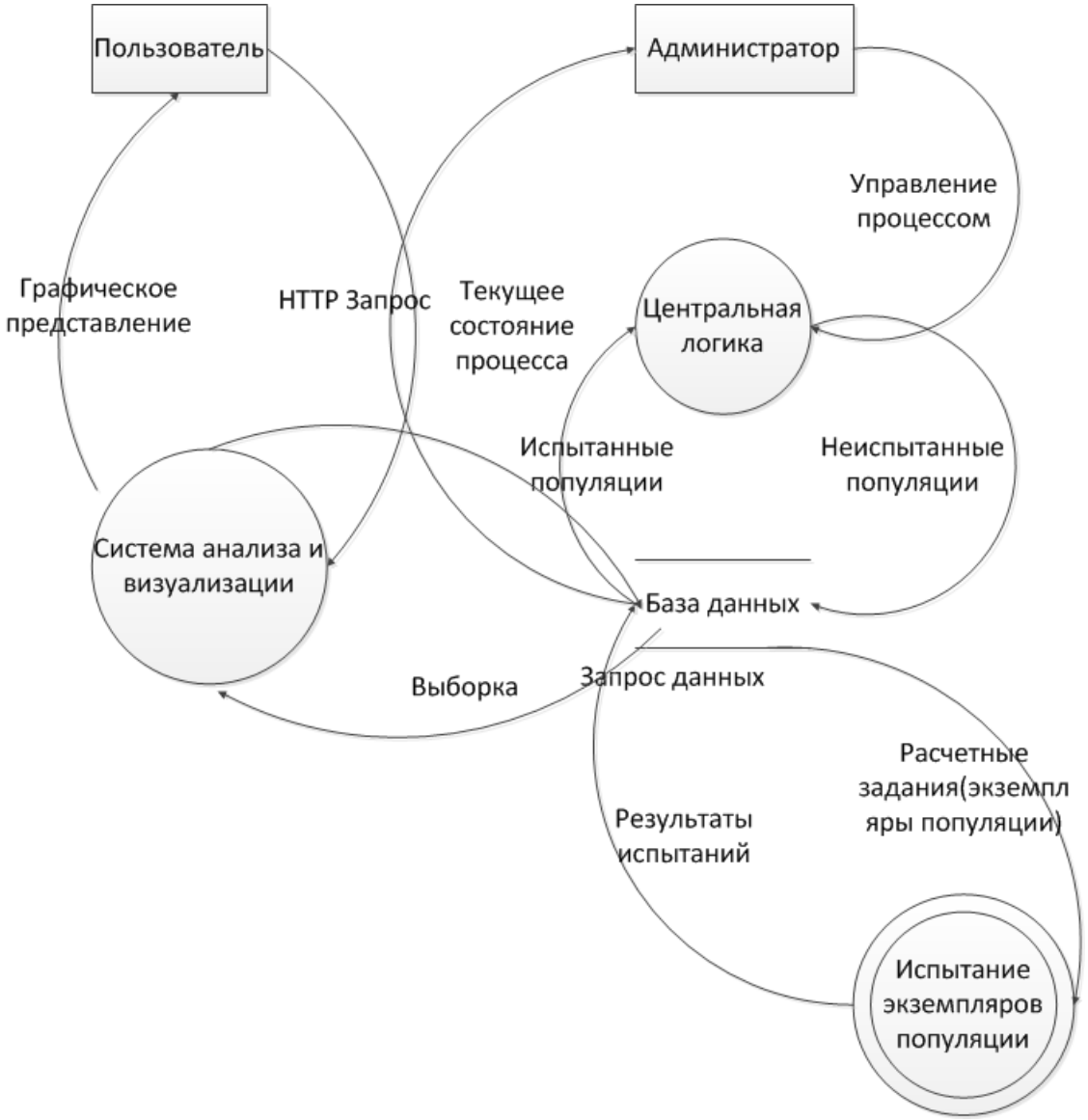
Используемый алгоритм



Используемые внутренние алгоритмы

- Кодирование – деревья функций
- Отбор – метод рулетки
- Фитнесс функция – обратна ошибке валидации
- Останов ГА – кол-во итераций / ручное прерывание
- Обучение ИНС – RPROP
- Останов обучения – достижение градиентом ошибки порогового значения

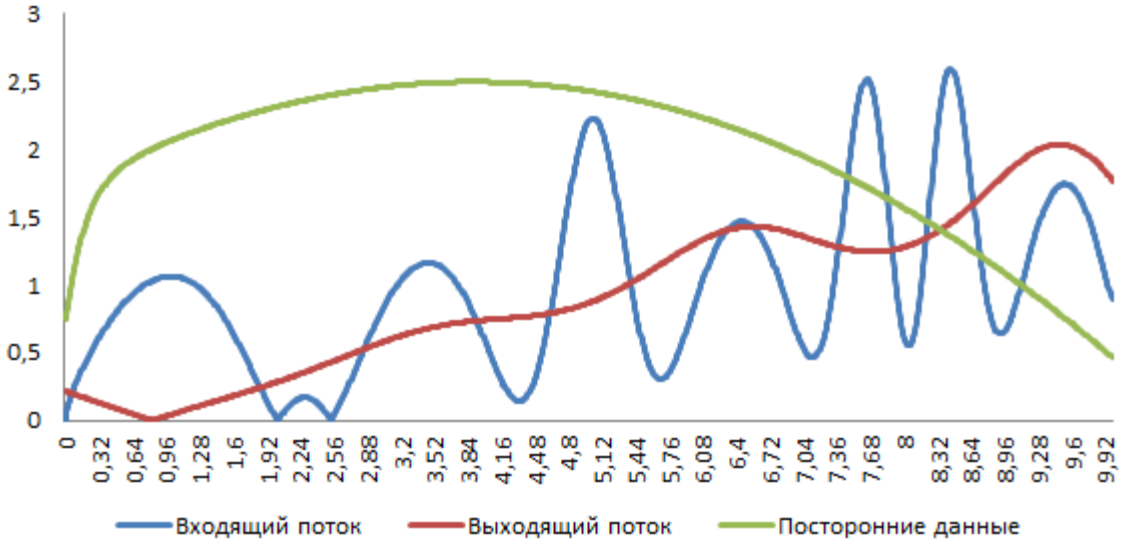
Прототип GRID системы, генерирующей ОВ

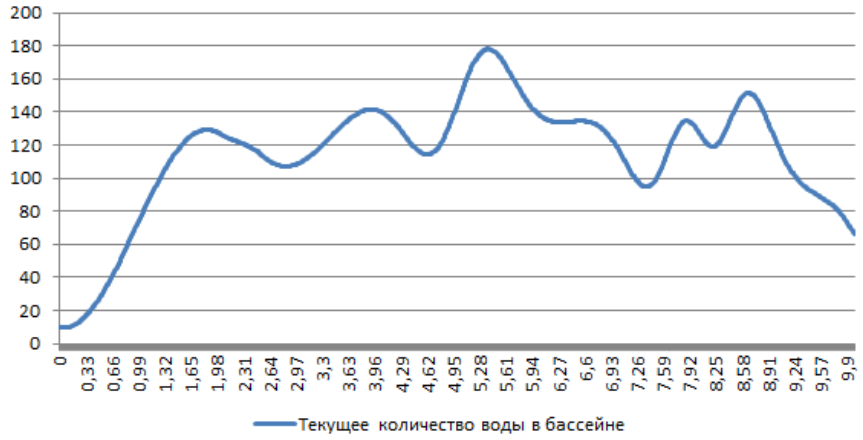


Описание реализованной системы

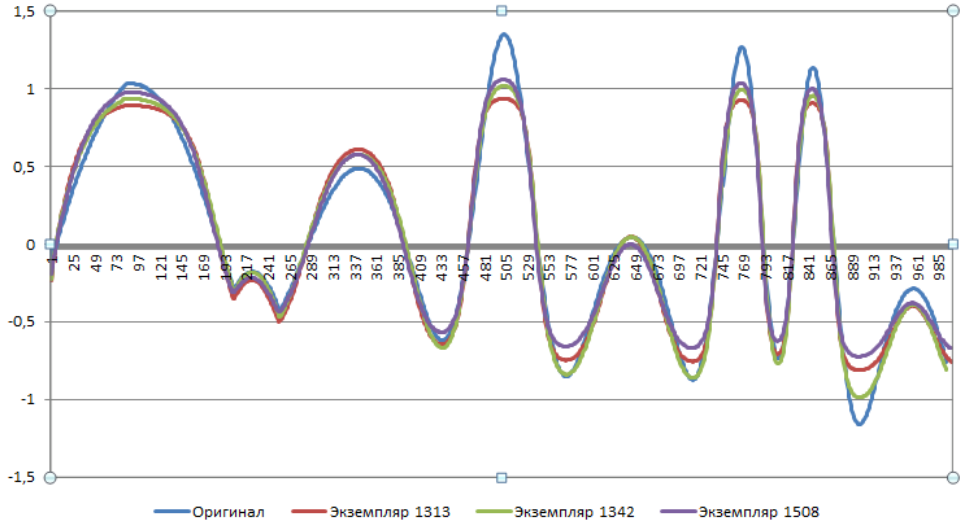
- Иерархическая GRID система
- Модули:
  - База данных. Роли: межпрограммный интерфейс, централизованное хранилище данных. Реализована на ОРСУБД PostgreSQL 9.0
  - Центральный логический модуль. Роли: генератор популяций для ГА, административный модуль. Реализован на языке F# 2.0
  - Вычислительные модули. Роли: вычисление значения фитнесс функции. Реализован на языке C# 4.0
  - Модуль анализа и визуализации экспериментальных данных. Роль: графический визуализатор с аналитическими возможностями. Реализован на языке Python
- Входные данные – набор доступных исторических данных и указание целевой функции
- Выходные данные – обученная модель ИНС, решающая задачу прогнозирования целевой функции, побочные результаты работы ГА

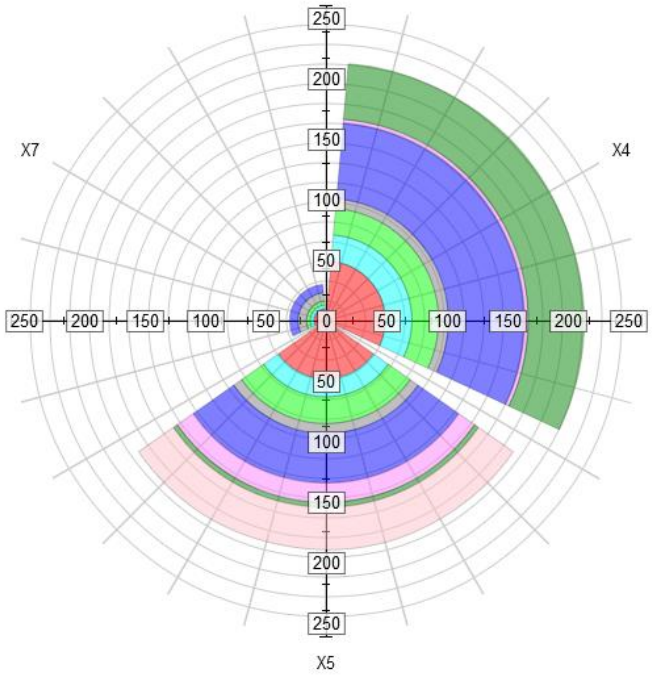
Тестовый сценарий «Бассейн». Исходные данные





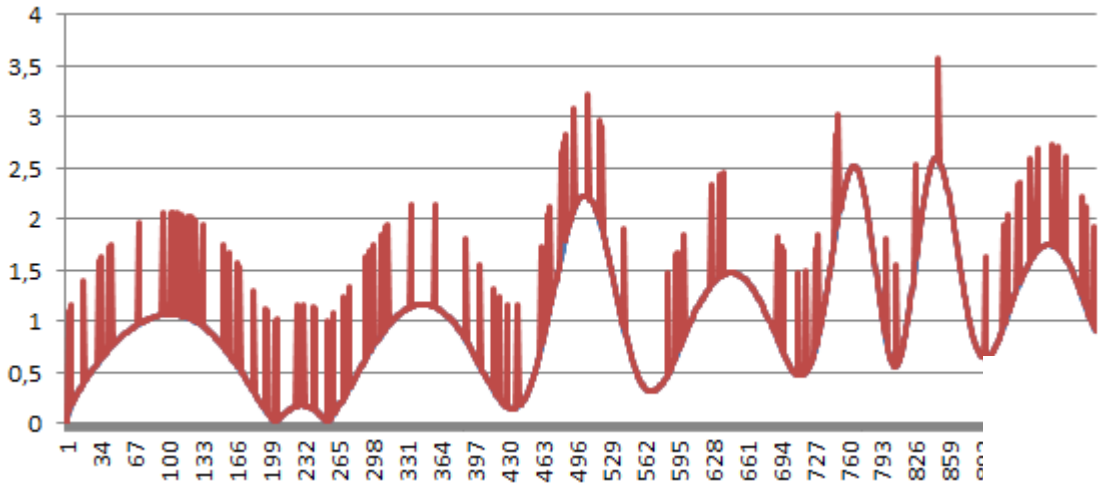
Тестовый сценарий «Бассейн». Результаты

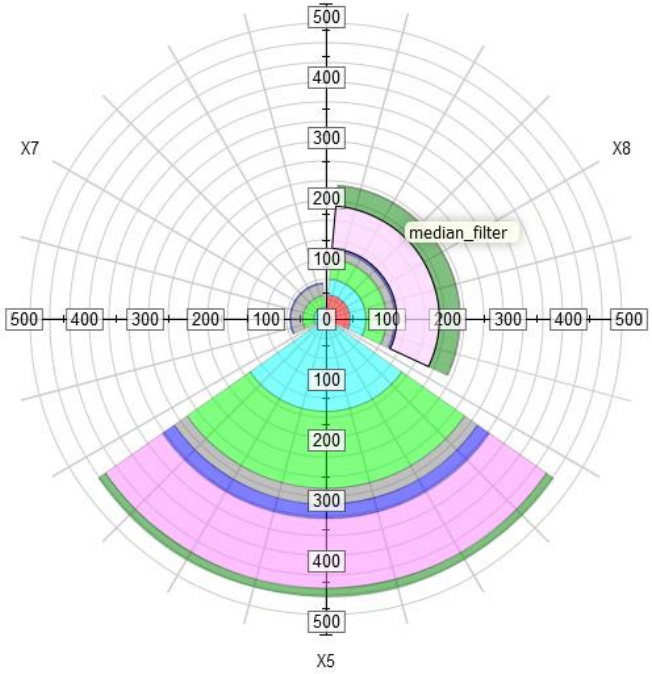




У лучшего прогноза
ширина доверительного интервала 0,02
Среднее по модулю значение сигнала 0,47

Тестовый сценарий «Бассейн. Зашумленный»





У лучшего прогноза
ширина доверительного интервала 0,14
Среднее по модулю значение сигнала 0,47

Финансовый тестовый сценарий

- Прогнозирование DJI по другим индексам, ценам на нефть и золото
- Лучший полученный результат: прогнозирование знака производной на 70 дней вперед с точностью 94,2%
- Получена сходная с действительной картина взаимосвязей предложенных исходных данных с DJI

Итоги работы

- Подтверждена способность системы к формированию эффективной обучающей выборки в условиях отсутствия информации о природе исходных данных
  - Обнаружение и ликвидация шумов
  - Обнаружение посторонних данных
- Подтверждена целесообразность анализа побочных результатов работы системы

Спасибо за внимание