Прикладные исследования и разработки компании Форексис в области интеллектуального анализа данных

Воронцов Константин Вячеславович
ЗАО «Форексис», ВЦ РАН
http://www.forecsys.ru
http://www.ccas.ru/voron

Прикладные задачи

- Кредитный скоринг
  (Credit Scoring)
- Предсказание ухода клиентов
  (Churn Prediction)
- Анализ клиентских сред
  (на примере Web Usage Mining)
- Прогнозирование временных рядов
- Портфельная оптимизация
  (Online Portfolio Selection)
- Имитационное моделирование:
  биржевые торги, аэропорт, call-центр,…
- Анализ текстов:
  АнтиПлагиат

Задача кредитного скоринга физических лиц

Исходные данные:
анкеты заёмщиков [+ макроэкономические показатели]

Признаки:
сумма кредита, доходы и расходы, работодатель, образование, возраст, пол,…

Классы:
хороший / плохой заёмщик

Применяемые методы классификации:

- Скоринговая карта — бинаризация, затем SVM или LR
- Нейронные сети
- Решающие деревья
- Поиск логических правил-конъюнкций
  - Решающий список правил
  - Взвешенное голосование правил

Задача кредитного скоринга

Специфические особенности задачи:

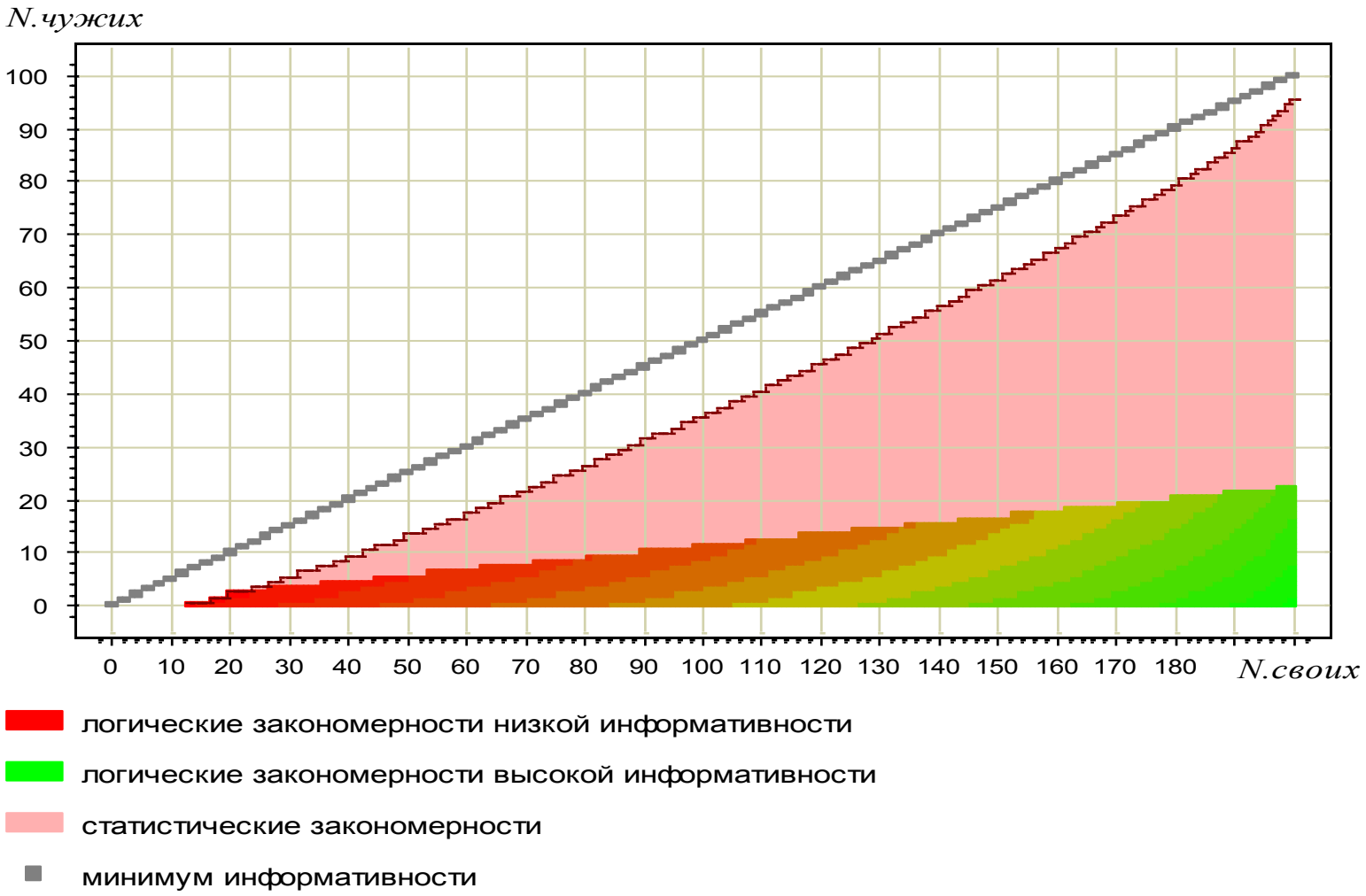
- Требование интерпретируемости классификаций (возможность выдать объяснение)
- Требование интерпретируемости алгоритма (возможность модифицировать его «вручную»)
- Обучаемость по малым выборкам
- Оценивание вероятности дефолта заёмщика
- Оценивание риска кредитного портфеля

Задача кредитного скоринга

Оценка риска кредитного портфеля:

- Оценивается
  вероятность дефолта каждого заемщика =
  = частота ошибок правила на обучении +
  + поправка на переобученность правила
- Строится эмпирическая функция распределения возможных потерь по всем заемщикам

Что такое «информативность логической закономерности» ?



Предсказание ухода клиентов

Сферы применения:

телеком, дисконтные программы, интернет-магазины,…

Исходные данные:

анкеты клиентов + протоколы поведения клиентов

Признаки поведения:

частота покупок, тариф, скидки, пакет услуг,…

Классы:

уходящий клиент / лояльный клиент

Специфика задачи:

- Огромные обучающие выборки
- Неизбежно высокий уровень ошибок
- Требуется выдача рекомендаций по способу воздействия на клиента

Задачи анализа клиентских сред

Сферы применения:

- е-коммерция, видеопрокат, телеком, дисконтные программы,…

Исходные данные:

- протоколы действий клиентов + анкеты (если есть) преобразуются в частотную матрицу «клиенты–товары»

Решаемые задачи:

- Персонализация предложения
- Каталогизация предлагаемых ресурсов (услуг, товаров)
- Анализ потребительских корзин и поиск правил ассоциации
- Маркетинговые исследования:
  сегментация, кластеризация, классификация

Задачи прогнозирования

Сферы применения:

- Объёмы продаж товаров в сети супермаркетов
- Объёмы ж/д перевозок
- Оптовые цены электроэнергии

Специфика задачи прогнозирования объёмов продаж:

- Огромное число временных рядов
  (порядка 10^7: 150 магазинов × 70,000 товаров)
- Ряды нестационарные, сильно зашумленные
- Существенность внешних факторов:
  праздники, промо-акции, действия конкурентов,…
- Требуется оперативность прогнозов
- Нестандартный функционал потерь

Адаптивные композиции алгоритмов прогнозирования

Гипотеза.

Ряд может переходить из одного состояния в другое,
и в каждом состоянии его поведение неплохо описывается
одной из стандартных моделей.

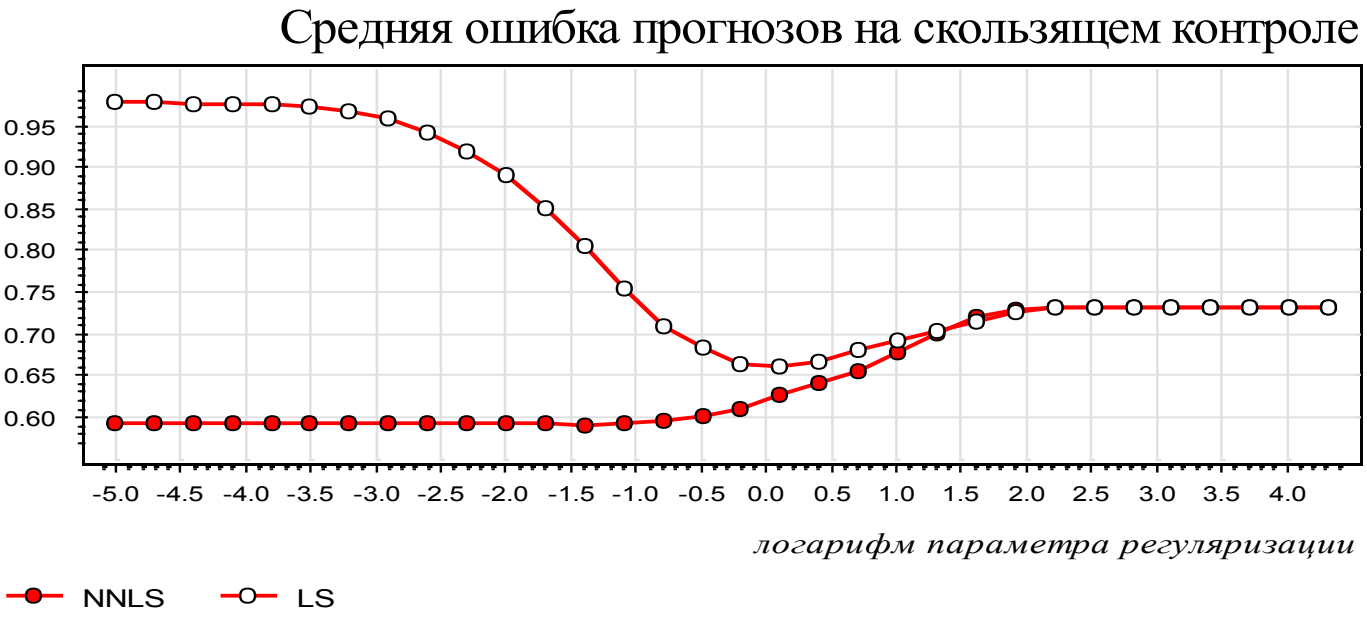
Адаптивные композиции алгоритмов прогнозирования

Композиции с адаптацией по последним T отсчетам:

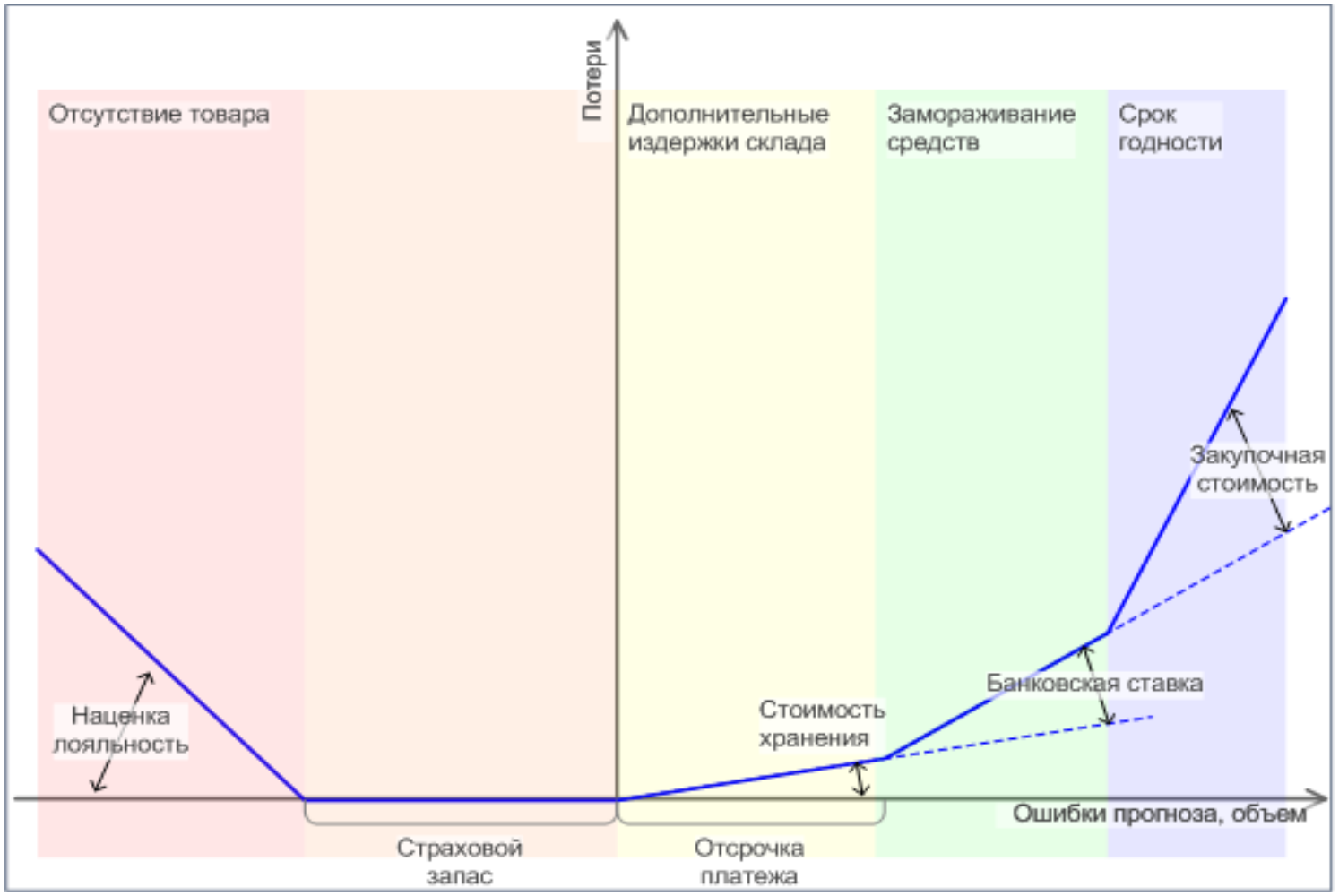
MS(T) — выбор лучшей модели …… 0.596

LS(T) — МНК с регуляризацией …… 0.659

NNLS(T) — МНК с регуляризацией и неотрицательными весами …… 0.590



Несимметричный неквадратичный функционал потерь



Задача оптимизации инвестиционного портфеля

Немного истории:

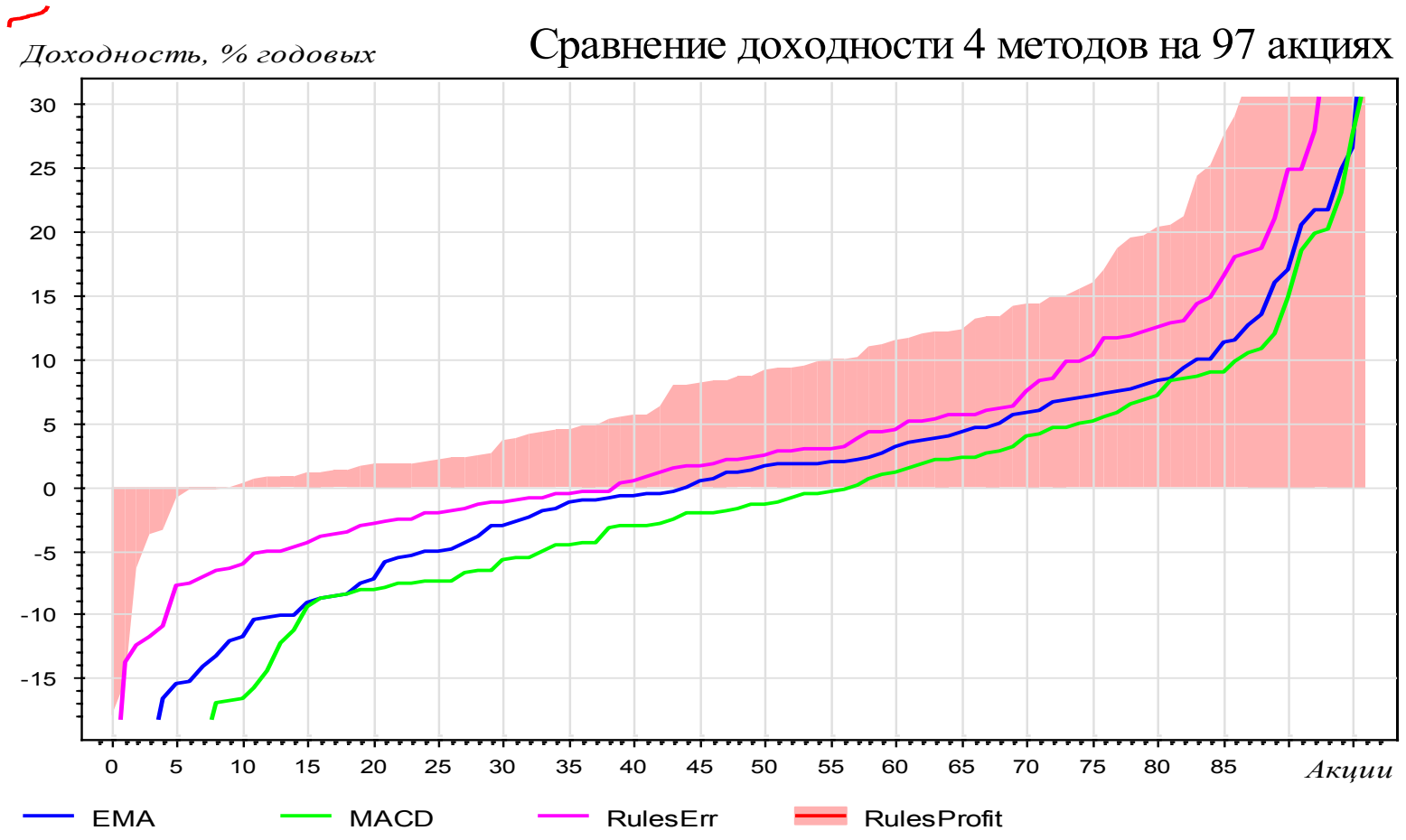
- Технические индикаторы рынка
- Теория Марковица [Markowitz 1952], [Sharpe 1963]
- Сравнительная теория портфелей (Competitive Theory of Portfolio Selection):
  - CBAL [Cover 1980]
  - Universal Portfolio [Cover 1991]
  - EG [Helmbold 1996]
  - ANTICOR, DELTA [Borodin 2000, 2004]

Задача оптимизации инвестиционного портфеля

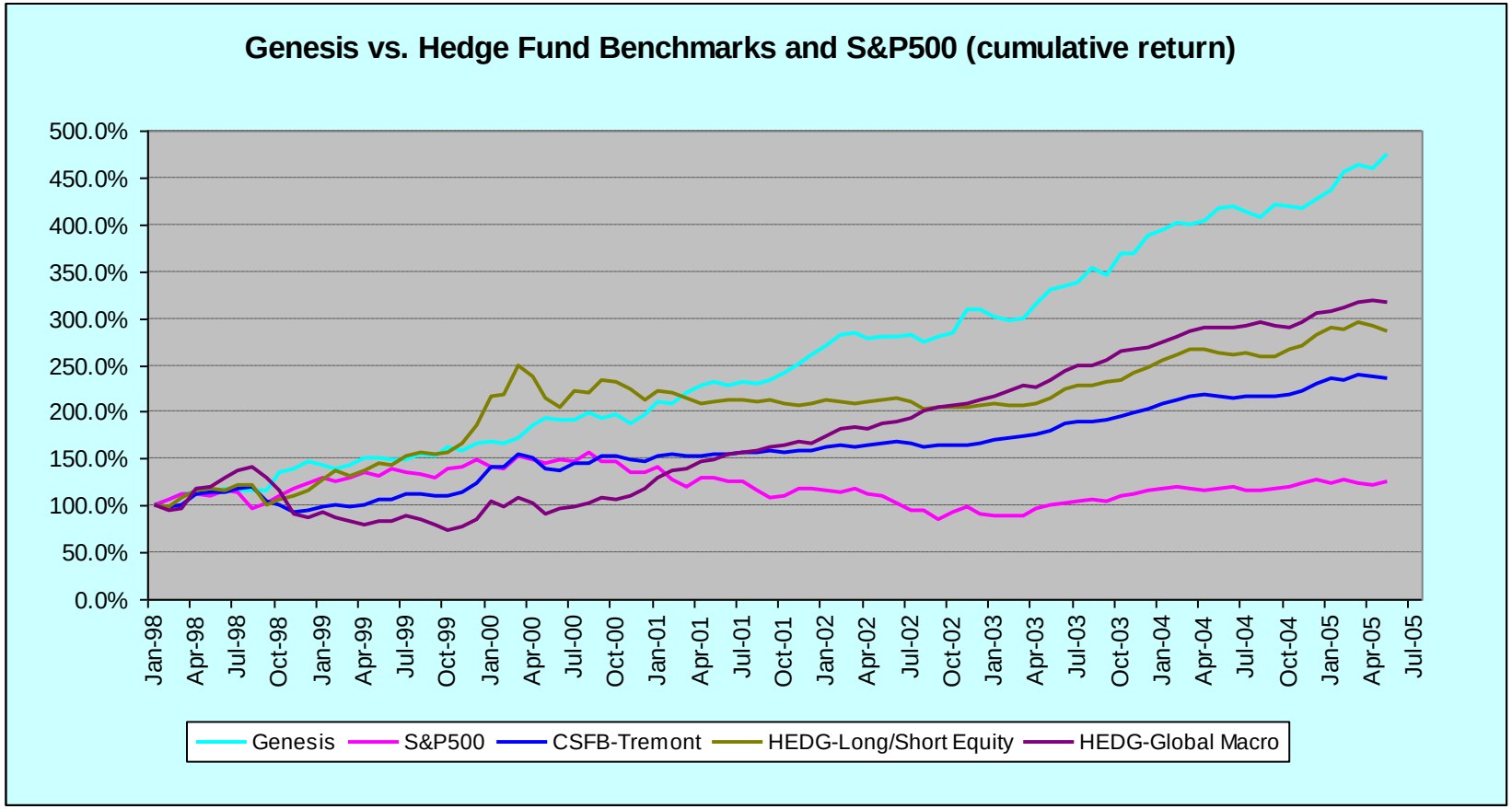
Подход Forecsys:

- Портфель собирается из «equities»:
  equity = (акция, технический индикатор)
- Портфельный алгоритм:
  - байесовский генетический алгоритм
  - тщательный подбор эвристик
- Скользящий контроль для снижения переобучения
- **Know-how:**
  построение индикаторов на основе поиска логических закономерностей в бинарных временных рядах

Инвестиционный портфель: результаты логического индикатора



Инвестиционный портфель: результаты портфельного алгоритма



Задачи имитационного моделирования

- Биржевые торги
- Аэропорт
- Центр обработки вызовов (call-center)
- Автотранспортные потоки

Имитационное моделирование биржевых торгов

Исходные данные:

детальный персонифицированный протокол торгов

Задача обучения:

построить модель игрока по протоколу его операций

Алгебраический подход:

- *Композиция*:

  стратегия игрока строится как композиция элементарных эвристических мотивов

- *Корректность*:

  абсолютно точное воспроизведение реальных торгов

Тесты адекватности:

проверка реакции модели на внешние воздействия