Федеральное государственное бюджетное учреждение науки
Вычислительный центра им. А.А. Дородницына
Российской академии наук

На правах рукописи

СОЛОГУБ РОМАН АРКАДЬЕВИЧ

АЛГОРИТМЫ ПОРОЖДЕНИЯ И ТРАНСФОРМАЦИИ МОДЕЛЕЙ В ЗАДАЧАХ НЕЛИНЕЙНОЙ РЕГРЕССИИ

05.13.17 — теоретические основы информатики

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2014

Работа выполнена в федеральном государственном бюджетном учреждении науки Вычислительный центр им. А. А. Дородницына Российской академии наук.

Научный руководитель: кандидат физико-математических наук, **Стрижов Вадим Викторович**, Федеральное государственно бюджетное учреждение науки Вычислительный центр им. А.А. Дородницына Российской академии наук, научный сотрудник Отдела интеллектуальных систем.

Официальные оппоненты:
доктор физико-математических наук, **Хачай Михаил Юрьевич**, Федеральное государственное бюджетное учреждение науки Институт математики и механики им. Н.Н. Красовского Уральского отделения Российской академии наук, заведующий Отделом математического программирования;
кандидат физико-математических наук, **Гуров Сергей Исаевич**, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования Московский государственный университет имени М.В. Ломоносова, доцент Кафедры атематических методов прогнозирования Факультета вычислительной математики и кибернетики.

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт системного анализа Российской академии наук.

Защита диссертации состоится 13 ноября 2014 г. в 15:00 на заседании диссертационного совета Д 002.017.02 при Федеральном государственном бюджетном учреждении Вычислительный центр им. А.А. Дородницына Российской академии наук, расположенном по адресу: 119333, г. Москва, ул. Вавилова, д. 40.
С диссертацией можно ознакомиться в библиотеке федерального государственного бюджетного учреждения науки Вычислительный центр им. А. А. Дородницына Российской академии наук и на сайте http://www.ccas.ru.

Автореферат разослан

Ученый секретарь диссертационного совета
Д 002.017.02, д.ф.-м.н., профессор Рязанов В.В.

Общая характеристика работы

Диссертационная работа посвящена проблеме порождения и трансформации моделей в задачах нелинейной регрессии.

Актуальность темы

Данная работа направлена на решение проблемы автоматического создания и верификации количественных математических моделей. Модели предназначены для описания результатов измерений и прогнозирования экспериментов, составляющих неотъемлемую часть естественнонаучных исследований.

В работе исследуется фундаментальная проблема автоматического порождения моделей для решения задач анализа данных. Порождаемые модели предназначены для аппроксимации, анализа и прогнозирования результатов измерений. При порождении учитываются требования, предъявляемые экспертами-специалистами в предметной области к порождаемым моделям. Это дает возможность получения экспертно-интерпретируемых моделей, адекватно описывающих результат измерения.

Для создания адекватной модели измеряемых данных используются экспертно-заданные порождающие функции и набор правил порождения. Модель задается в виде суперпозиции порождающих функций. Правила порождения определяют допустимость суперпозиции и исключают порождение изоморфных моделей.

В работе предлагается развить существующие методы автоматического порождения моделей. В частности, при порождении моделей предлагается учитывать экспертные требования к виду моделей, ранжируя модели в соответствии с экспертными предпочтениями. Предлагаются новые методы поиска изоморфных суперпозиций, основанные на поиске изоморфных подграфов и подстановке подграфов по правилам. Исследуются методы и алгоритмы порождения моделей, их свойства, сложность и устойчивость.

В некоторых прикладных задачах моделирования оказывается, что сведения о структуре модели, в том числе экспертные оценки о виде искомых зависимостей являются недостаточными для применения методов, способных обеспечить необходимое качество модели. Недостаток числа независимых переменных для построения математических моделей делает применение методов порождения признаков и моделей перспективным для решения такого рода задач.

Идея метода порождения признаков заключается в создании дополнительных независимых переменных, являющихся образами исходных переменных относительно последовательно примененных наборов отображений. Такие отображения в рамках работы будут называться порождающими функциями.

Ранее работы, выполненные в рамках данного подхода, являлись в основном прикладными. Для различных задач экономики и промышленности на основе экспертного описания проблемы выбирались порождающие функции и порождались наборы новых признаков для построения модели. При этом исследователями не ставился ворос существования набора, полноты или корректности существующего алгоритма.

В данной работе развивается теоретическое обоснование корректности и допустимости использования методов порождения суперпозиций для решения прикладных задач. Рассматриваются алгоритмы порождения суперпозиций, их сходимость, предлагаются различные методы оптимизации структуры моделей.

Индуктивное порождение моделей с помощью методов группового учета аргумента рассматривается в работах Г.Н. Ивахненко. В линейной модели предлагается генерировать новые признаки с помощью операции произведения. С помощью полиномов Колмогорова-Габора алгоритм целенаправленно порождает и перебирает модели-претенденты различной сложности по ряду критериев. В результате находится модель оптимальной структуры в виде одного уравнения или системы уравнений.

Важным этапом развития методов решения задач восстанов-

ления регрессии является ипользование нелинейной регрессии для решения прикладных задач. Данный подход широко описывается в работах Дж. Себера — рассматривается построение и оценка параметров нелинейных моделей. Для оценки моделей используется алгоритм Левенберга-Марквадта.

Для индуктивного порождения моделей в работых Дж. Козы и Н. Зелинки, связанных с генетическим программированием, используется символьная регрессия — метод построения регрессионных моделей путем перебора различных произвольных суперпозиций функций из некоторого заданного набора. Индуктивное порождение моделей рассматривается в приложении к задаче определения оптимальной формы антенны. В работах В. В. Стрижова идеи индуктивного порождения регрессионных моделей находят свое развитие в применении методов двухуровневого Байесовского вывода к процесу порождения и настройки моделей.

При анализе стркутуры моделей основным способом представления суперпозиций является граф-дерево. В связяи с этим к суперпозициям применимы методы трансформации графов, позволяющие описать формально методы структурной опитимизации суперпозиций. Рассматривается категорное представление трансформаций на графах и условия применимости правил Для преобразования деревьев выделяются некоторые элементарные графы, для которых строятся оболочки изоморфных им графов более сложной структуры.

Целью работы является исследование проблемы построения нелинейных регрессионных моделей как суперпозиций заданных параметрических функций.

Основные положения, выносимые на защиту:

1. Разработан алгоритм направленного порождения моделей, исследованы свойства порождаемых суперпозиций (с. 35-41).

2. Разработаны новые алгоритмы вычисления структурной

сложности порождаемых суперпозиций и алгоритмы вычисления расстояния между порождаемыми суперпозициями (с. 23-35).

3. Разработан метод обнаружения изоморфных суперпозиций. Разработан алгоритм поиска изоморфных подграфов, соответствующих порожденным суперпозициям (с. 47-68).

4. Введено формальное определение трансформаций графов, соответствующих суперпозициям. Доказана применимость трансформаций при использовании структуры двойного кодекартова квадрата (с. 51-62).

Научная новизна:

1. Предложен алгоритм индуктивного порождения регрессионных моделей, являющихся суперпозициями экспертно-заданных параметрических функций.

2. Предложен метод трансформации суперпозиций, представленных в виде категории на множестве направленных ациклических графов без самопересечений, соответствующих суперпозициям.

3. Предложен алгоритм упрощения регрессионных моделей, являющихся суперпозициями экспертно-заданных параметрических функций.

Практическая значимость

Предлагаемые в работе методы порождения моделей предназначены непосредствнно для применения на практике. Алгоритмы порождения суперпозиций могут использоваться для решения задач обучения по прецедентам в различных прикладных областях, включая техническую диагностику, социологию, экономические задачи, задачи финансового рынка.

Финансовый рынок в целом характеризуется большим количеством ложных регрессий между ценами различных инструментов. В связи с этим, в качестве входных переменных,

использование которых в модели ценообразования инструмента финансового рынка оправдано, может выступать лишь небольшая группа основных факторов рынка. В то же время цены и волатильности различных производных инструментов финансового рынка имеют ярко выраженную существенно-нелинейную зависимость от независимых переменных. В таких условиях применение алгоритмов порождения суперпозиций является оптимальным инструментом решения прикладных задач, стоящих перед экспертами финансового рынка.

Достоверность изложенных в работе результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных алгоритмов на реальных задачах регрессии; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК.

Апробация работы. Основные результаты работы докладывались на:

- Интеллектуализация Обработки Информации ИОИ-08 (Украина, Алушта, 2008);
- SIAM Conference on Financial Mathematics & Engineering 2008 (USA, New Bruinswick, 2008);
- Математика. Компьютер. Образование. 2009 (Россия, Пущино, 2009);
- EURO 2009 conference (Germany, Bonn, 2009);
- Математические Методы Распознавания Образов ММРО-14 (Россия, Суздаль, 2009);
- EURO 2010 conference (Portugal, Lisbon, 2010);
- EURO 2012 conference (Lithuania, Vilnus, 2012).

Публикации. Основные результаты по теме диссертации изложены в 6 печатных изданиях, 3 из которых изданы в журналах, рекомендованных ВАК, 3 — в тезисах докладов.

1. Сологуб Р.А. Алгоритмы порождения нелинейных регрессионных моделей // Информационные технологии, 2013. No 5. С. 8 – 12.

2. Сологуб Р.А. Порождение регрессионных моделей поверхности волатильности биржевых опционов // Информационные технологии, 2012. No 8. С. 47 – 52.

3. Стрижов В.В., Сологуб Р.А. Индуктивное порождение поверхности волатильности опционных торгов // Вычислительные технологии, 2009. No 5. С. 102—113.

4. Sologub R., Strijov V. The inductive generation of the volatility smile models // SIAM Financial Modeling 08 conference proceedings. P. 21.

5. Sologub R. Inductive generation of foreign exchange forecast models // 23rd European Conference On Operational Research proceedings. P. 162.

6. Sologub R. Model generation for equity-futures spread forecasting // 24th European Conference On Operational Research proceedings. P. 168.

Личный вклад. Результаты получены автором самостоятельно при научном руководстве к.ф.-м.н. В.В. Стрижова. Личный вклад автора в работах с соавторами заключается в следующем: в работе [3] алгоритмы порождения моделей и экспериментальная часть работы созданы лично автором. Работы [1] и [2] выполнены автором целиком лично.

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и приложения. Полный объем диссертации **92** страницы текста с **5** рисунками и 2 таблицами. Список литературы содержит **53** наименования.

Содержание работы

Во введении обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме. Рассматривается переход от порождения линейных моделей к индуктивному порождению моделей, описываются возникающие проблемы и ранее разработанные методы их решения. Также формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава

В данной главе рассматривается постановка задачи нелинейной регрессии в контексте построения моделей-суперпозиций заданных функций.

Пусть задана выборка $\mathfrak{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, $\mathbf{x} \in \mathbb{R}^m$. Требуется построить функцию регрессии $\varphi(\mathbf{x}, \mathbf{w}) \mapsto \mathbf{y}$. Требуется определить модель f — отображение из декартова произведения множества свободных переменных $\mathbf{x} \in \mathbb{R}^n$ и множества параметров $\mathbf{w} \in \mathbb{R}^m$ в $\mathbb{R}^1$. Модель соответствует функции φ при заданном значении $\mathbf{w} = \mathbf{w}_0$. Для модели оценивается набор параметров $\mathbf{w}_0$, доставляющие минимум внешнему критерию качества модели — квадратичной ошибке

$$S(\mathbf{w}|\mathfrak{D}, f) = ||f(\mathbf{x}, \mathbf{w}) - y||_2.$$

Задано множество G порождающих функций $g(\mathbf{w}, \mathbf{x})$. Для каждого элемента данного множества g_i определены области аргументов $\mathbf{w} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^n$ и значений, при этом область значений принадлежит $\mathbb{R}^1$. В множество порождающих функций обязательно входит не имеющая аргументов функция $\mathrm{id}(\mathbf{x})$, значение которой тождественно значению свободной переменной, а также функция константы Const.

Искомую модель f мы будем искать среди множества супер-

позиций функций $g \in G$. При этом накладываются ограничения на структуру суперпозиции:

Определение 1. Допустимой называется суперпозиция, удовлетворяющая следующим требованиям:

1. Элементами суперпозици f могут являться только порождающие функции g_j и свободные перменные $\mathbf{x}$.

2. Количество аргументов элемента суперпозиции равно арности соответствующей ему функции g_j.

3. Порядок аргументов элемента суперпозиции соотвествует порядку аргументов соответствующей функции g_j.

4. Для элемента s_i, аргументом которого является элемент s_j, область определения соответствующей порождающей функции g_i содержит область значений порождающей функции аргумента g_j: $\mathrm{dom}(g_i) \supseteq \mathrm{cod}(g_j)$;

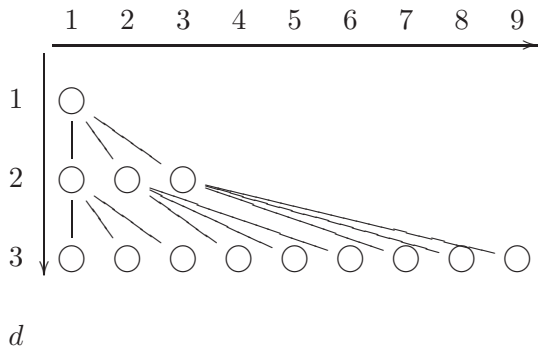
Условимся считать, что каждой суперпозиции f сопоставлено дерево Γ_f, эквивалентное этой суперпозиции и строящееся следующим образом:

- В вершинах v_i дерева Γ_f находятся соответствующие порождающие функции g_j.

- Число дочерних вершин у некоторой вершины v_i равно арности соответствующей ей функции g_j.

- Порядок дочерних вершин вершины v_i соотвествует порядку аргументов соответствующей функции g_j.

- Листьями дерева Γ_f являются свободные переменные x_i либо числовые параметры w_i.



(1)

Заранее задается максимальная арность вершин $a_{\max}$. Максимальное дерево, которое можно построить при таком условии, будет иметь 1 вершину на первом слое, $a_{\max}$ на втором, $a_{\max}^2$ на третьем и так далее. Пример координатной сетки для деревьев с $a_{\max} = 3$ представлен на диаграмме (1).

В дальнейшем при изменении моделей будет модифицироваться именно их структура, поэтому помимо расстояния как нормы разности функций следует каким то образом определить структурное расстояние, позволяющее оценить степень схожести и различия моделей.

Определение 2. Дерево Γ_0 называется общим поддеревом деревьев Γ_1 и Γ_2, если в них существуют поддеревья Γ_1' и Γ_2', изоморфные дереву Γ_0.

Определение 3. Общее поддерево двух деревьев называется наибольшим, если в нем содержится наибольшее число вершин среди других общих поддеревьев. Наибольшее поддерево деревьев $\Gamma_i(V_i)$ и $\Gamma_j(V_j)$ обозначается как $\Gamma_{ij}(V_{ij})$. Символом p будет обозначаться количество элементов в множестве V: $|V_i| = p_i$, $|V_j| = p_j$, $|V_{ij}| = p_{ij}$.

В работе будет использоваться функция расстояния r между Γ_i и Γ_j, зависящая от их размеров и наибольшего общего подграфа,

$$r_{ij} = p_i + p_j - 2p_{ij}.$$

Теорема 1. Для функции r_{ij} выполняются условия метрики.

Теорема 2. При замене в дереве $\Gamma_0(V_0)$ поддерева $\Gamma'_0(V'_0)$ поддеревом $\Gamma'_0(V'_0)$ дерева $\Gamma_1(V_1)$ получается дерево $\Gamma_2(V_2)$. Расстояния между деревьями $r_{02} = r(\Gamma_0(V_0), \Gamma_2(V_2))$ и $r_{01} = r(\Gamma_0(V_0), \Gamma_1(V_1))$:

$$r_{02} \leqslant p'_0 + p'_1,$$
$$r_{12} \leqslant p_1 + p_0 - p'_0 - p'_1.$$

Теорема 3. Расстояние между исходным деревом $\Gamma_0(V_0)$ и порожденным деревом $\Gamma_1(V_1)$, $|V_0| = |V_1| = p_0$, полученным с помощью операции замены одной вершиной дерева, не более чем $2(p_0 - \frac{p_0-1}{k+1})$, где k — максимальное число аргументов среди порождающих функций g, составляющих деревья Γ_0 и Γ_1.

Для ограничения размера порождаемых моделей предлагается определение сложности суперпозиции, позволяющее штрафовать суперпозиции с большим числом вложенных функций.

Определение 4. Сложность C суперпозиции f равна сложности дерева Γ, соответствующего ей, и определяется как сумма количества элементов во всех поддеревьях дерева Γ.

Теорема 4. Сложность C_{Γ_1} дерева Γ_1, полученного заменой в дереве Γ_0 поддерева Γ'_0 на поддерево Γ'_1 с корнем в вершине (d, i). В случае подобной замены сложность C_{Γ_1} дерева Γ_1 будет равна

$$C_{\Gamma_1} = C_{\Gamma_0} + d(C_{\Gamma'_1} - C_{\Gamma'_0}).$$

Вторая глава

Для решения прикладных задач машинного обучения недостаточно определить свойства моделей. Требуется указать ал-

горитм построения регрессионных моделей. Рассматриваются различные алгоритмы порождения моделей и их свойства.

Алгоритм $\mathcal{F}$ порождения суперпозиций будет итеративным. Перед первым шагом построим начальное значения множества моделей F_0 (индекс множества соответствует шагу, на котором получены соответствующие модели):

$$F_0 = \{X, \text{Const}\},$$

где X соответствует выборке $\mathfrak{D}$, а Const соотвествует порождающей функции константы. Далее на каждом шаге для множества F_i построим вспомогательное множество U_i, состоящее из суперпозиций, полученных в результате применения функций $g_i \in G$ к элементам $f_j \in F_{i-1}$:

$$U_i = \{g_i(f_{j_1}, \dots f_{j_k}), \mid g_i \in G_u, f \in F_{i-1}\}.$$

Тогда множество F_i

$$F_i = F_{i-1} \cup U_i.$$

Теорема 5. Алгоритм $\mathcal{F}$ породит любую допустимую суперпозицию ограниченной сложности за конечное число шагов.

Теорема 6. Справедлива следующая оценка количества суперпозиций, порожденных алгоритмом $\mathfrak{F}$ после k-ой итерации:

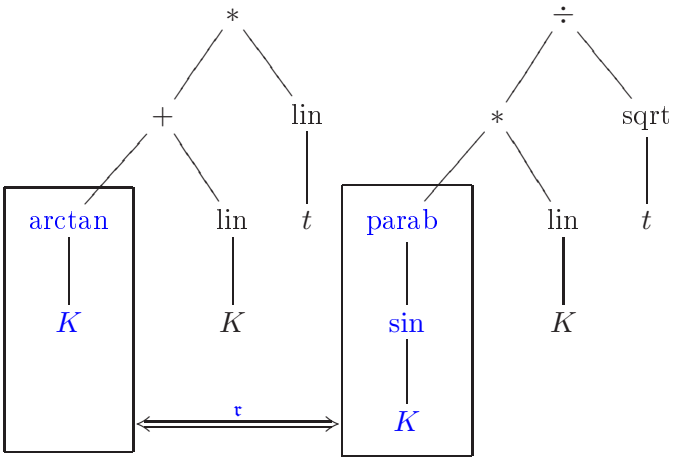
$$|\mathcal{F}_k| = \mathcal{O}(l_p^{(p^k-1)/(p-1)} n^{p^k}).$$

Следует заметить, что количество моделей растет более чем экспоненциально, поэтому алгоритм $\mathfrak{F}$ имеет слишком высокую вычислительную сложность даже при невысокой сложности C порождаемых моделей и поэтому не подходит для решения прикладных задач.

Алгоритм последовательного порождения моделей $\mathfrak{G}$ работает итерационно, пока не будет достигнут необходимый уровень значения ошибки S или через определенное число итераций.

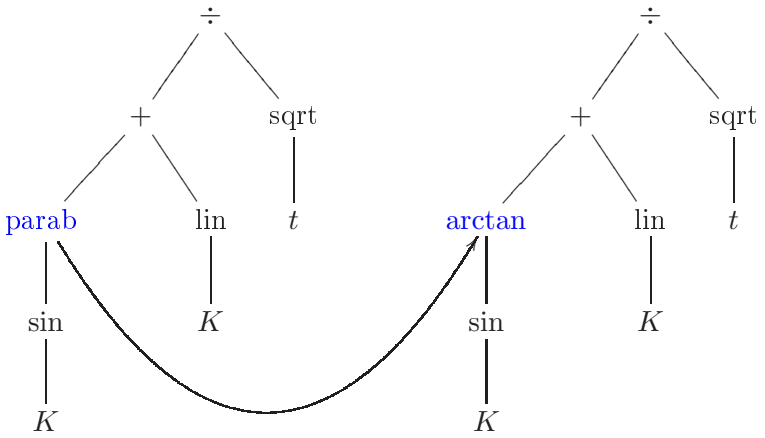
1. Некоторым методом оптимизации параметров минимизируются функции ошибки $S_i(w)$ для каждой модели f_i. Отыскиваются параметры $\mathbf{w}$ и вычисляется значение функции ошибки S каждой модели.

2. Заданы следующие правила построения производных моделей f_j. Для модели f_j строится тождественная ей модель f'_j. В дереве Γ'_j, соответствующем модели f'_j произвольно выбирается вершина v_k. Выбирается произвольная модель f_m из множества F и произвольная вершина v_l её дерева Γ_m. Модель f'_j модифицируется путем замещения в дереве Γ'_j поддерева, корнем которого является вершина функции v_k на поддерево дерева Γ_m с корнем в вершине v_l. Измененная модель f'_j добавляется в множество F.



3. С заданной вероятностью p_0 каждая модель $f_j \in F$ подвергается изменениям. В дереве Γ'_j, соответствующем модели f'_j в соответствии с некоторой заданной функцией распределения выбирается вершина v_k. Соответствующая ей порождающая функция g_k заменяется случайным образом выбранной функцией g_m той же арности из множества

порождающих функций G.



4. Модели из множества F сортируются в соответствии со значениями функции ошибки S_i. Заданная доля наилучших моделей используется в дальнейших итерациях.

Алгоритм $\mathfrak{G}$ последовательного порождения моделей является рандомизированным. Таким образом, достижение некоторой оптимальной модели не может быть гарантированно. Однако возможно показать рост числа моделей обладающих определенными признаками. Для исследования процедуры замены поддеревьев следует ввести некоторые понятия.

Определение 5.
Схемой называется дерево, содержащее функции из множества $G \bigcup \{=\}$ и свободные переменные из множества $T \bigcup \{=\}$, где G и T — множества порождающийх функций и свободных переменных соответственно. Порождющая функция $\{=\}$ означает произвольный символ, который может быть любой функцией или свободной переменной. Порядком схемы $O(H)$ называется количество вершин в ней, не являющихся $\{=\}$.

Определение 6. Гиперсхемой называется дерево, содержащее функции из множества $G \bigcup \{=\}$ и свободные переменные из

множества $T \bigcup \{=, \#\}$. Порождющая функция $\{=\}$ определяется также, как и для схемы, а свободная переменная $\{\#\}$ означает любое допустимое дерево.

Рассмотрим операцию замены поддерева со следующим ограничением: топологическая структура родительских вершин заменяемого и заменяющего поддерева равны.

Определение 7. Операция замены поддерева $\mathfrak{r}$, для которой топологическая структура родительских вершин заменяемого и заменяющего поддерева равны, будет называться операцией замены поддерева с сохранением структуры.

Для данной операции формулируется аналог теоремы схем.

Теорема 7. Для операции $\mathfrak{r}_c$ замены поддерева с сохранением структуры оценивается вероятность сохранения схемы H:

$$\alpha(H,t) = (1 - p_{x0})p(H,t) + p_{x0}\alpha_{x0}(H,t),$$

где

$$\alpha_{x0}(H,t) = \sum_{h_1}\sum_{h_2} \frac{p(h_1,t)p(h_2,t)}{NC(h_1,h_2)} \sum_{i \in C(h_1,h_2)} \delta(h_1 \in U(H,i))\delta(h_2 \in L(H,i)),$$

при этом

p_{x0} — вероятность проведения операции замены поддерева,

$p(H,t)$ — вероятность выбора вершины из схемы H,

суммы проходят по всем деревьям из набора,

$NC(h_1,h_2)$ — количество вершин с равной структурой родительских вершин,

$L(H,i)$ — гиперсхема, получаемая из H заменой всех вершин от корня до вершины i вершинами типа $=$, а всех поддеревьев, выходящих из этих вершин — $\#$,

$U(H, i)$ — гиперсхема, получаемая из H заменой поддеревьев ниже точки i вершинами типа $\#$.

Третья глава

В данной главе исследуется структура порождаемых моделей, предлагается алгоритм структурного упрощения моделей. Рассматривается категорное представление правил трансформации графов и анализируются условия применимости правил.

Определение 8. Модель f_2 с вектором параметров $\mathbf{w}_2$ называется обобщающей для модели f_1 с вектором параметров $\mathbf{w}_1$, если для любого вектора $\mathbf{w}_1$ найдется такой вектор $\mathbf{w}_2$, что для любого $\mathbf{x} \in D$ значения функций $f_1(\mathbf{w}_1, \mathbf{x})$ и $f_2(\mathbf{w}_2, \mathbf{x})$ равны:

$$\mathbf{x} \in D \Rightarrow f_1(\mathbf{w}_1, \mathbf{x}) = f_2(\mathbf{w}_2, \mathbf{x}).$$

Определение 9. Модели f_1 и f_2 с векторами параметров $\mathbf{w}_1$ и $\mathbf{w}_2$ называются эквивалентными, если каждая из них является обобщающей для другой.

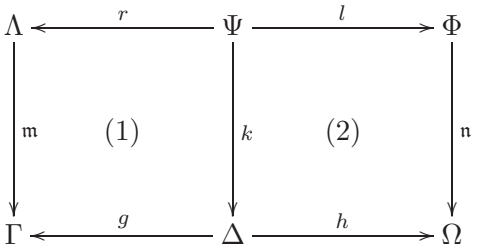
Алгоритм упрощения модели $f(\mathbf{w}, \mathbf{x})$ минимизирует сложность суперпозиции, соответствующей её дереву, при условии, что результирующая модель $f'(\mathbf{w}', \mathbf{x})$ является обобщающей моделью для исходной модели $f(\mathbf{w}, \mathbf{x})$. При проведении данной операции какие-либо вершины и ребра из дерева, соответствующего трансформируемой модели f будут удалены, и будут построены другие вершины и ребра вместо них. Обобщим процедуру упрощения на орграфы любого вида, а не только на деревья. Далее для каждого графа подразумевается, что это орграф.

Определение 10. Подграф L, удаляемый из графа G в алгоритме упрощения, будет называться заменяемым подграфом.

Определение 11. Создаваемый подграф R, помещаемый в граф G в алгоритме упрощения, называется замещающим подграфом.

Для рассмотрения трансформации графов необходимо ввести понятие правила.

Определение 12. Правило — это тройка $p = (\Lambda, \Psi, \Phi)$, где Λ и Φ являются заменяемым и замещающим подграфами и граф Ψ является общей частью подграфов Λ и Φ, то есть их пересечением. Заменяемый, или начальный подграф Λ называется условием применения правила, замещающий, или конечный подграф Φ — итогом его применения. Подграф Ψ описывает часть графа, необходимую для применения правила, но неизменную в процессе применения. Множество $\Lambda \setminus \Psi$ является удаляемой частью графа, вместо неё создается множество $\Phi \setminus \Psi$.



Определение 13. Процедура поиска $\mathfrak{m}$ — отображение из Λ в Γ, ставящая в соответствие заменяемому графу эквивалентный ему подграф. При этом процедура $\mathfrak{m}$ сохраняет структуру графа Γ.

Определение 14. Трансформация графа — это пара, элементами которой являются правило p и процедура поиска $\mathfrak{m}$. Процедура трансформации графа Γ в граф Ω с помощью правила p и процедуры поиска $\mathfrak{m}$ будет также обозначаться как $\Gamma \overset{p,\mathfrak{m}}{\to} \Omega$.

Сформулируем условие существования графов Ψ и Δ при трансформации графа. Для этого вводим следующие определния:

Определение 15. Точки соединения — вершины и ребра в Λ, которые не удаляются при применении правила p.

Определение 16. Точки обнаружения — вершины и ребра в Λ, образы которых относительно $\mathfrak{m}$ имеют более одного прообраза.

Определение 17. Подвешенные вершины — вершины в Λ, образы которых относительно $\mathfrak{m}$ в Γ имеют входящие или выходящие ребра, не содержащиеся в Λ.

Теорема 8. Пусть дано правило $p = (\Lambda \leftarrow \Psi \rightarrow \Phi)$, граф Γ и процедура поиска $\mathfrak{m} : \Lambda \rightarrow \Gamma$. Вершины графов обозначаются буквой V, ребра — E. Тогда правило p с процедурой поиска $\mathfrak{m}$ удовлетворяет условию соединения если все точки обнаружения и подвешенные вершины также являются точками соединения.

Теорема 9. Любой трансформации $t = (\Lambda_t, \Psi_t, \Phi_t)$ графа соответствует набор правил $p_t = (\Lambda_{p_t}, \Psi_{p_t}, \Phi_{p_t})$, удовлетворяющих условию соединения, такой что любое применение трансформации t аналогично применению одного из правил p_t.

Для рассмотрения случаев применения нескольких трансформаций необходимо определить условие, при котором трансформации могут применяться последовательно и параллельно. Введем понятия параллельно и последовательно независимых трансформаций.

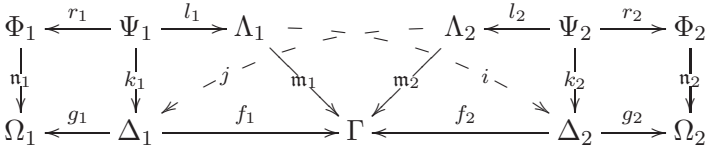
Определение 18. Две трансформации графов $\Gamma \overset{p_1,m_1}{\rightarrow} \Omega_1$ и $\Gamma \overset{p_2,m_2}{\rightarrow} \Omega_2$ являются параллельно независимыми, если все вершины и ребра, попадающие в образ обоих морфизмов поиска, являются соединительными:

$$\mathfrak{m}_1(\Lambda_1) \bigcap \mathfrak{m}_2(\Lambda_2) \subseteq \mathfrak{m}_1(l_1(\Psi_1)) \bigcap \mathfrak{m}_2(l_1(\Psi_2)).$$

Теорема 10. Две трансформации графов-деревьев $\Gamma \overset{p_1,m_1}{\rightarrow} \Omega_1$ и $\Gamma \overset{p_2,m_2}{\rightarrow} \Omega_2$ являются параллельно и последовательно независимыми, если образы корней v_1 и v_2 деревьев $\mathfrak{m}_1(\Lambda_1)$ и $\mathfrak{m}_2(\Lambda_1)$ не принадлежат друг другу:

$$v_1 \notin \mathfrak{m}_2(\Lambda_2)\ v_2 \notin \mathfrak{m}_1(\Lambda_1).$$

Теорема 11. Две трансформации графов $\Gamma \overset{p_1,m_1}{\to} \Omega_1$ и $\Gamma \overset{p_2,m_2}{\to} \Omega_2$ являются параллельно независимыми, если существуют морфизмы $i : \Lambda_1 \to \Delta_2$ и $j : \Lambda_2 \to \Delta_1$, такие что $f_2 \circ i = \mathfrak{m}_1$ и $f_1 \circ j = \mathfrak{m}_2$:



Прикладной алгоритм упрощения моделей строится на основе вышеобозначенной операции.

Определение 19. Шаблон θ — гиперсхема, обладающая наименьшей сложностью среди всех гиперсхем, таких что при их взаимном замещении получаемые модели оказываются эквивалентными. Сложность гиперсхемы определяется как сложность суперпозиции при замещении всех символов $\{=\}$ и $\{\#\}$, означающих соответственно произвольную независимую переменную и произвольное поддерево, на константы.

Экспертно выбирается некоторый набор шаблонов Θ. алгоритм упрощения состоит из двух шагов:

1. Все поддеревья Γ_j в выбранном дереве Γ проверяются на эквивалентность шаблонам из Θ согласно заданным правилам.

2. Если какое-либо поддерево Γ_j в дереве эквивалентно дереву из Θ, данное поддерево заменяется соответствующим элементом из Θ.

Процедура повторяется до тех пор, пока после вышеперечисленных итераций дерево Γ не останется неизменным. При наличии в множестве порождающих функций коммутативных

функций вводится алфавитное упорядочение для ветвей, выходящих из вершины γ_i дерева Γ, соответствующей коммутативной порождающей функции g_i.

Четвертая глава

В данной главе приведено описание вычислительного эксперимента на реальных данных финансового рынка. В экспериментах построенные ранее экспертные модели сравниваются с порожденными алгоритмом $\mathfrak{G}$. При этом рассматриваются теоретические представления о характере нелинейной зависимости между зависимой и свободными переменными. Порожденные модели показывают более высокое качество работы, чем ранее использованные, при этом их сложность C не превосходит сложность экспертных моделей. Также характер зависимости в порожденных моделях подтверждает предположения экономистов.

Заключение

Основные результаты диссертационной работы.

1. Разработан алгоритм направленного порождения моделей. Разработаны новые алгоритмы вычисления структурной сложности порождаемых суперпозиций и алгоритмы вычисления расстояния между порождаемыми суперпозициями.

2. Разработан метод последовательного направленного порождения суперпозиций, исследованы свойства порождаемых суперпозиций.

3. Введено понятие изоморфных суперпозиций, разработан метод их обнаружения. Разработан алгоритм поиска изоморфных подграфов, соответствующих порожденным суперпозициям.

4. Разработан новый метод порождения экспертно-интерпретируемых моделей. Создана базовая библиотека правил порождения экспертно-интерпретируемых моделей.

Публикации автора по теме диссертации

1. **Сологуб Р.А. Алгоритмы порождения нелинейных регрессионных моделей // Информационные технологии, 2013. No 5. С. 8 – 12**

2. **Сологуб Р.А. Порождение регрессионных моделей поверхности волатильности биржевых опционов // Информационные технологии, 2012. No 8. С. 47 – 52**

3. **Стрижов В.В., Сологуб Р.А. Индуктивное порождение поверхности волатильности опционных торгов // Вычислительные технологии, 2009. No 5. С. 102—113.**

4. Sologub R., Strijov V. The inductive generation of the volatility smile models // SIAM Financial Modeling 08 conference proceedings. P. 21.

5. Sologub R. Inductive generation of foreign exchange forecast models // 23rd European Conference On Operational Research proceedings. P. 162.

6. Sologub R. Model generation for equity-futures spread forecasting // 24th European Conference On Operational Research proceedings. P. 168.

7. Стрижов В.В., Сологуб Р.А. Индуктивное построение регрессионных моделей волатильности // сборник трудов конференции МКО-2009. С. 58.

8. Стрижов В.В., Сологуб Р.А. Индуктивное порождение регрессионных моделей волатильности // Сборник трудов конференции ИОИ-2008. С. 215-216.