

Московский Институт Стали и Сплавов
(Технологический университет)

На правах рукописи

Леонов Евгений Николаевич

**Принятие решений на основе нечёткозначных моделей и алгоритмов
обработки графических данных в технологии машинного обучения**

Специальность 05.13.01

«Системный анализ, управление и обработка информации (в производственной сфере)»

Автореферат

диссертации на соискание ученой степени кандидата технических наук

Москва 2009

Работа выполнена на кафедре АСУ Московского Государственного Института Стали и Сплавов (Технологического Университета).

Научный руководитель:

к.т.н., доцент Поляков В. Н.

Официальные оппоненты:

Заведующий кафедрой Информатики РЭА им. Плеханова,
д.т.н., профессор Попов И. И.

Заведующий кафедрой Финансового менеджмента РГТЭУ,
к.э.н., профессор Костюхин Ю. Ю.

Ведущая организация:

Государственное образовательное учреждение «Казанский государственный университет имени В.И. Ульянова-Ленина».

Защита состоится «04» марта 2009 г. в 14.00 на заседании Диссертационного совета Д.212.132.07 при Московском Государственном Институте Стали и Сплавов по адресу: 119049, Москва, Ленинский проспект, 4, ауд. К-325.

С диссертацией можно ознакомиться в библиотеке МИСИС.

Автореферат разослан « » 2009 г.

Ученый секретарь
Диссертационного Совета

к.т.н., профессор
Калашников Е.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Последние десятилетия сопряжены с увеличивающимся потоком информации.

Акции крупных металлургических предприятий участвуют в международной торговле, принося своим владельцам прибыль. Чтобы достичь этой цели, владельцы акций прибегают к специальным методикам, советам специалистов и программным комплексам, использующим особые алгоритмы.

Прибыль от торговли напрямую зависит от правильности принимаемых решений. Однако в новом информационном веке необходимо принимать решение, основываясь на огромном количестве факторов. Одной из важнейших задач для эффективного управления экономикой является создание информационных систем, которые бы обеспечивали хранение и анализ информации, а также снабжали специалистов данными, необходимыми для принятия верного решения.

Спорадически осциллирующие системы обуславливают необходимость применения новых подходов для разработки технологий поддержки принятия решений в динамической предметной области.

С ростом производительности компьютеров стало возможным решать задачи предсказания в реальном времени. Под реальным временем подразумевается время принятия решения, которое может варьироваться от минут до часов, в зависимости от вида прогноза.

Таким образом, **актуальность работы** заключается в огромном количестве доступной информации, с разной степенью влияющей на принятие решения, и отсутствием средств, которые позволяют её обработать, предложив верное решение с большой долей вероятности. Графическое представление исходной информации является специфической особенностью данной работы. На основе известных методов можно сделать вывод, что наибольшие накладные расходы происходят на этапе формирования исходной выборки. Существенно сократить их помогает машинное обучение, методика которого представлена в настоящей работе.

Цель работы заключается в исследовании поведения спорадически осциллирующих систем в динамической предметной области, представляемых в виде двумерных графиков, проблем применения технологии машинного обучения в таких средах, а также создании информационной системы поддержки принятия решений, основанной на нечёткозначных моделях и машинном обучении.

Для достижения поставленной цели были выполнены **задачи**:

- проанализированы существующие подходы к использованию графической

информации в различных сферах деятельности;

- изучены различные интеллектуальные технологии, используемые для принятия решений на основе графических данных;
- рассмотрены варианты применения технологии машинного обучения на графических данных;
- созданы методы и алгоритмы распознавания фигур на графиках в динамической предметной области;
- в целях апробации разработанного метода создан прототип системы поддержки принятия решений на финансовом рынке.

Научная новизна работы заключается в следующем:

- разработаны нечеткозначные модели фигуры «флаг» и участка ценового ряда «тренд», а также дано формальное описание общей модели произвольной ленточной фигуры;
- в работе описан метод машинного обучения на примерах, представленных в виде графических данных («FuzGraph»). В качестве модели описания (репрезентации) выборки примеров была использована нечеткозначная формализация геометрических фигур;
- разработаны метод и алгоритм распознавания фигур на графиках в динамической предметной области;
- разработано алгоритмическое решение задач прогнозирования и мониторинга на финансовых рынках.

Теоретическая значимость заключается в следующем:

- исследованы корреляционные зависимости между различными параметрами графических фигур, учитываемых в процессе принятия решений;
- описаны нечеткозначная фигура «флаг», участок ценового ряда «тренд», произвольная ленточная фигура, а также функции принадлежности для них;
- выполнена формальная постановка задачи для разработки системы принятия решений, основанной на технологии машинного обучения на графических данных («FuzGraph»), с использованием нечётких функций принадлежности;
- создан алгоритм поиска линий поддержки и сопротивления;
- построена модель машинного обучения на графических данных.

Практическая значимость заключается в:

- возможности построения информационной системы металлургического предприятия, которая позволяет эффективнее управлять его рыночными активами;
- возможности более эффективного контроля над открытыми позициями на валютных и фондовых рынках;
- использовании разработанной технологии обработки графической информации в различных областях, таких как: медицина, финансовый анализ и др.

В рамках данной исследовательской работы создан прототип системы поддержки принятия решений.

Методы исследования

При разработке программного комплекса использовались:

- методы реляционной алгебры, методы статистического анализа и теория нечётких множеств для описания моделей нечёткозначных фигур и для формирования базы правил в рамках системы принятия решений;
- парадигма машинного обучения для формирования обучающей выборки графических фигур;
- методы алгоритмического моделирования и методы объектно-ориентированного программирования для построения опытного образца системы «ЕМ»;
- методика технического анализа на финансовых рынках для апробации предложенных моделей и алгоритмов.

Результаты работы были практически реализованы в виде программного комплекса, включающего в себя набор инструментов для обучения системы экспертом, глобального анализа исторической информации, контролем над развитием тенденции в реальном времени, реляционной СУБД, хранящей все необходимые данные для работы комплекса.

Апробация работы. Результаты работы докладывались на следующих научных конференциях:

- Когнитивное моделирование в лингвистике (выставка программных продуктов) CML-2005, Варна, Болгария 2005 г.
- 8-ая международная конференция «Распознавание образов и анализ изображений» (РОАИ-8-2007), Йошкар-Ола, Российская Федерация 7-13 октября 2007 г.

Теоретические вопросы диссертации освещаются в четырёх научных публикациях, в том числе одна публикация в издании, рекомендованном ВАК.

Работа соответствует паспорту специальности 05.13.01, и выполнена в следующих областях исследования:

- Формализация и постановка задач системного анализа, оптимизации, управления, принятия решений и обработки информации.
- Методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений в технических, медицинских и социальных системах.
- Визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации.
- Методы получения, анализа и обработки экспертной информации.

Структура диссертации. Работа состоит из введения, шести глав, заключения, списка литературы и четырёх приложений.

ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ ОБРАБОТКИ ГРАФИЧЕСКИХ ДАННЫХ

В первой главе рассматриваются различные интеллектуальные технологии поддержки принятия решений на основе обработки графических данных применительно к АСУ и АСУ ТП металлургических предприятий, финансового анализа, методам медицинской диагностики, а также методика машинного обучения, как неотъемлемая часть данных технологий.

На основе изученного материала можно сделать следующие выводы:

- Развитие современных информационных систем обуславливает растущий поток информации, представляемой в графическом виде.
- Несмотря на достаточно большое количество различных интеллектуальных технологий, ни одна из них не в состоянии удовлетворительно решать задачи реального мира самостоятельно. Гибридные образования из нескольких интеллектуальных технологий более эффективны.
- Способ машинного обучения, применяемый в информационной системе, во многом определяет эффективность её работы в дальнейшем. Использование обучающей выборки, полученной в процессе машинного обучения, как основы информационной системы, позволяет устранить человеческий фактор во время принятия решения на рабочих данных.
- Для торговли на финансовых и фондовых рынках существует большое количество информационных систем, однако они наследуют недостатки используемых ими интеллектуальных технологий. Следовательно, нужны новые подходы к анализу графических данных, которые были бы ориентированы непосредственно на использование в динамической предметной области, которой и является финансовый рынок.

НЕЧЁТКОЗНАЧНЫЕ МОДЕЛИ ФИГУРЫ «ФЛАГ» И УЧАСТКА ЦЕНОВОГО РЯДА «ТРЕНД»

Во второй главе даётся подробное описание нечёткозначной модели фигуры «флаг» и участка ценового ряда «тренд».

Одной из методик технического анализа является предсказание котировок финансового инструмента по фигурам. В этом случае график представляется в виде набора фигур и переходов между ними. Каждая фигура подаёт сигнал, значимый или нет, в зависимости от конкретной ситуации. Распознавая эти сигналы и их совокупность, трейдер принимает решение.

Фигурой называется графический образ на графике котировок валют или акций, напоминающий геометрическую фигуру. Различают следующие три основных вида фигур: «флаг», «треугольник», «голова - плечи». Есть и другие фигуры, но они встречаются значительно реже.

Данное исследование направлено на автоматизацию метода предсказания котировок, основанного на фигурах. Сложность автоматизации этого метода заключается в том, что он носит визуальный, зачастую субъективный характер, и до сих пор не поддавался компьютеризации.

Среди многообразия фигур, встречающегося на финансовых рынках, и являющегося основанием для технического прогноза был выбран «бычий флаг», как фигура, встречающаяся и используемая наиболее часто. Для того чтобы описать данную фигуру параметрически, нам пришлось подобрать наиболее подходящий метод моделирования, создать модель, запрограммировать и апробировать ее на фактических данных.

Учитывая субъективный и нечеткий характер визуальной информации, которую приходится анализировать на фигурах, было принято решение использовать аппарат нечёткой логики.

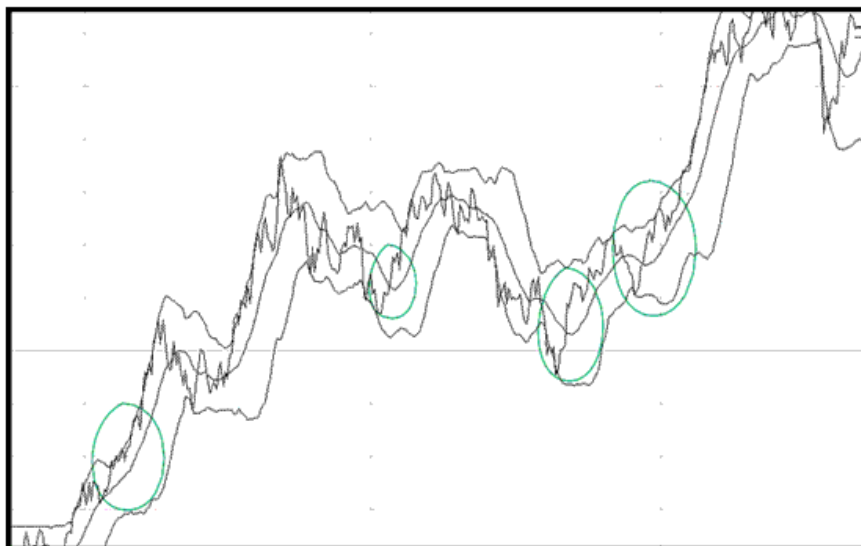


Рисунок 1 - Примеры образования флага на реальных графиках валют

Параметрами древка и полотнища являются:

- Длина по оси времени или количество дней, в течение которых древко или полотнище образуется. Эти параметры обозначены как ΔT_d и ΔT_n для древка и полотнища соответственно.
- Высота по оси котировок или количество пунктов, составляющих древко и полотнище (ΔC_d и ΔC_n).
- Угол наклона фрагмента фигуры к оси времени. Обозначается $tg(\varphi_d)$ и $tg(\varphi_n)$ для древка и полотнища соответственно. Дальнейшие исследования показали, что этот параметр чрезвычайно важен.
- Нормальная толщина канала, образованного древком и полотнищем, обозначается h_d и h_n соответственно.

Есть ряд особенностей, присущих этой фигуре:

- образуется при восходящем тренде;
- чем больше угол наклона, тем стремительнее и дальше будет отскок;
- чем меньше нормальная толщина древка, тем стремительнее и дальше будет отскок;
- если длина полотнища будет больше длины древка в 3 и более раз, то можно считать, что фигура «бычий флаг» сложилась.

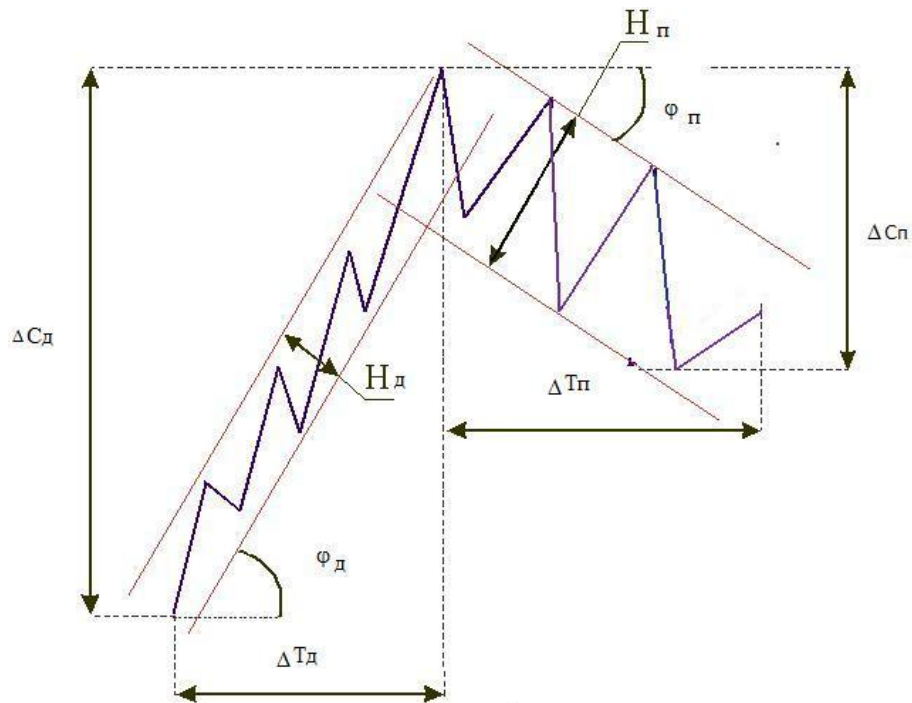


Рисунок 2 - Параметры модели фигуры технического анализа «бычий флаг»

ΔC – высота полотнища (древка) (диапазон изменения котировок валют). ΔC рассчитывается как разность между максимальным и минимальным значением цены валюты:

$$\Delta C = C_{\text{макс}} - C_{\text{мин}} , \quad (1)$$

ΔT – диапазон изменения времени (расстояние в днях по горизонтальной оси от первой до последней точки фигуры «флаг»). ΔT рассчитывается как разность между максимальным и минимальным значением времени:

$$\Delta T = T_{\text{макс}} - T_{\text{мин}} , \quad (2)$$

$\text{tg}(\varphi)$ - тангенс угла наклона полотнища (древка). $\text{tg}(\varphi)$ определяется как угол наклона линии поддержки (сопротивления).

$\Delta h_{\text{в}}$ – толщина полотнища (древка) по вертикали. $\Delta h_{\text{в}}$ рассчитывается как разница между котировками ($b_{\text{макс}}$, $b_{\text{мин}}$), полученными на точках пересечения вертикальной прямой с линиями поддержки и сопротивления по формуле:

$$\Delta h_{\text{в}} = b_{\text{макс}} - b_{\text{мин}} , \quad (3)$$

h – «нормальная» толщина полотнища (древка). h рассчитывается как минимальное расстояние между линиями поддержки и сопротивления. Для расчета h используется формула:

$$h = \frac{\Delta h_6}{\sqrt{1 + \operatorname{tg}^2 \varphi}}, \quad (4)$$

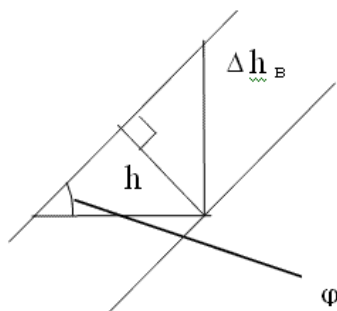


Рисунок 3 - Нормальная толщина полотна (древка)

Задача идентификации флага сводится к решению трех подзадач:

- идентификация древка;
- идентификация полотна;
- формирование сигналов об открытии позиции.

Постановку задачи в логико-математической форме можно представить в следующем виде:

1. Имеется упорядоченное множество C значений функций $c_i(t_i)$, где c_i – котировка валюты $c_i \in C$, t_i – момент времени $t_i \in T$.

2. Найти интервал значений $T' \subset T$, для которого справедливо:

$$\left. \begin{array}{l} \Delta C_{\min} < c'_n - c'_0 < \Delta C_{\max} \\ t'_0 \leq t'_j \leq t'_n, \quad \forall t'_j \in T' \end{array} \right\} \text{Ограничение по высоте полотна (древка),} \quad (5)$$

$$\left. \begin{array}{l} \Delta T_{\min} < t_n - t_0 < \Delta T_{\max} \\ t'_0 \leq t'_j \leq t'_n, \quad \forall t'_j \in T' \end{array} \right\} \text{Ограничение по длительности полотна (древка),} \quad (6)$$

$$\left. \begin{array}{l} \varphi_{\min} \leq \frac{c'_n - c'_0}{t'_n - t'_0} \leq \varphi_{\max} \\ t'_0 \leq t'_j \leq t'_n, \quad \forall t'_j \in T' \end{array} \right\} \text{Ограничение по углу наклона полотна (древка),} \quad (7)$$

$$\left. \begin{array}{l} \Delta t_{\min} \leq t'_n - t'_0 \leq \Delta t_{\max} \\ \forall c'_i \in C \end{array} \right\} \text{Ограничение по длительности полотна (древка),} \quad (8)$$

Решение задачи выявления (идентификации) флага включает, как отмечалось, две подзадачи:

- выявление древка;
- выявление полотнища.

Обе эти подзадачи в свою очередь сводятся к задаче построения линий поддержки и сопротивления для семейства точек $S=\{c_i(t_i), i=1 \dots n\}$ и сравнения параметров этих прямых с типовыми параметрами флага.

Будем искать решение задачи построения линий поддержки (сопротивления) как выбор из множества всех прямых соединяющих пары точек с координатами (c_i, t_i) (c_j, t_j) для всех пар $i, j \in I$.

На рисунке 4 представлена блок-схема решения задачи построения линии поддержки (сопротивления).

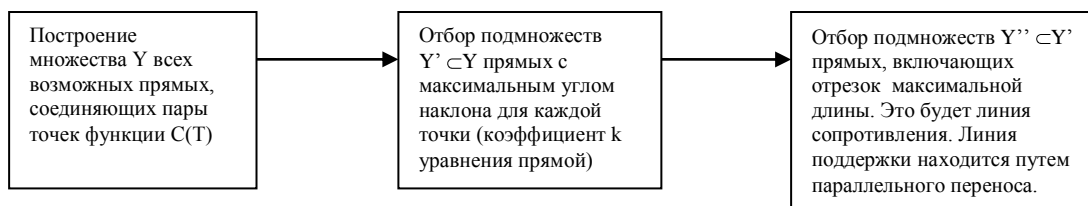


Рисунок 4 - Блок-схема решения задачи построения линии поддержки (сопротивления).

Графические иллюстрации к решению задачи построения линий поддержки и сопротивления представлены на рисунке 5.

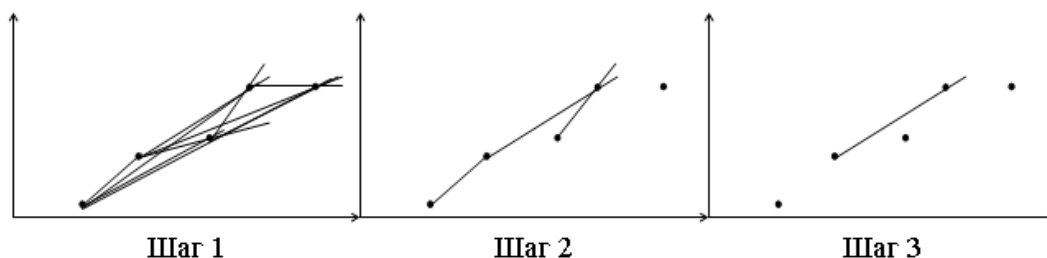


Рисунок 5 – Иллюстрация к решению задачи построения линий поддержки и сопротивления

Рассмотрим предложенное решение задачи более подробно.

Шаг 1. Найти множество прямых $Y = \{y_{12}, y_{13}, \dots, y_{ij}, \dots, y_{n-2, n}, y_{n-1, n}\}$ для каждой точки на графике $S(T)$.

Уравнение прямой имеет вид

$$y=kx+b, \quad (9)$$

То есть фактически нам необходимо найти пары коэффициентов $\{(k_{ij}, b_{ij}), i=1 \dots n-1; j=2 \dots n\}$. Каждая пара и определяет соответствующую прямую.

Шаг 2. Для каждой точки i из множества пересекающих ее прямых выбрать одну прямую, имеющую максимальное значение коэффициента. Таким образом, мы сформируем новое множество прямых $Y' \subset Y, Y' = \{Y': \max_i(k_{ij}), i=1 \dots n-1\}$.

Шаг 3. Из полученного на втором шаге множества прямых Y' выбрать прямую, включающую отрезок максимальной длины. Это будет линия сопротивления. Линия поддержки находится путем параллельного переноса.

После формального описания фигуры «флаг» можно перейти к описанию фигуры «бычий флаг». Если рассматривать среднесрочные прогнозы, флаг – это кратковременная фигура, длящаяся несколько дней. Обычно он возникает на динамичных рынках, где бывают резкие изменения цен. Бычий флаг образуется при восходящем тренде, медвежий флаг – при нисходящем.

Самая главная особенность флага, которая служит основой для прогноза, это то, что после него идет прорыв цен («отскок»), чаще всего равный величине древка флага.

Процесс принятия решений по фигурам носит субъективный характер и во многом определяется наличием опыта у эксперта. В связи с этим возникло предположение, что эксперт руководствуется некоторым обобщенным образом фигуры флаг, который можно рассматривать как функционал от одной или совокупности нечётких функций принадлежности. Учитывая сравнительно большое число параметров в модели (восемь), можно было бы сконструировать одну многомерную функцию принадлежности. Однако это снизило бы наглядность и как следствие возможность промежуточного контроля результатов компьютерного моделирования. В связи с этим был проведен анализ парной корреляции параметров модели, и было принято эвристическое решение строить шесть двумерных функций принадлежности.

Среди таких функций были выделены следующие:

$$m_1(\Delta C_d, \Delta T_d), \quad (10)$$

$$m_2(\text{tg}(\varphi_d), h_d), \quad (11)$$

$$m_3(\Delta C_n, \Delta T_n), \quad (12)$$

$$m_4(\text{tg}(\varphi_n), h_n), \quad (13)$$

$$m_5(\Delta C_d, \Delta C_n), \quad (14)$$

$$m_6(\Delta T_d, \Delta T_n), \quad (15)$$

Считаем, что распределение вероятности параметров фигуры «флаг» подчиняется нормальному распределению и описывается системой функций Гаусса для двумерного нормального распределения.

$$p(x, y) = C \cdot e^{-Q(x-a_1, y-a_2)}, \quad (16)$$

где

$$C = \frac{p(x, y)}{[2\pi\sigma_1\sigma_2\sqrt{1-R^2}]^{-1} \pi}$$

функция плотности распределения;
коэффициент нормировки;

число пи;

дисперсии одномерных распределений для
переменных x и y соответственно;

математические ожидания одномерных
распределений переменных x и y соответственно;

R - коэффициент корреляции, имеющий физический смысл условной вероятности событий $(x-a_1)$ и $(x-a_2)$;

$$R = \frac{\sum_{i=1}^n [(x_i - a_1) \cdot (y_i - a_2)]}{(n-1) \cdot \sigma_1 \cdot \sigma_2}, \quad (17)$$

$$Q(x - a_1, y - a_2) = \frac{1}{2(1-R^2)} \left[\frac{(x - a_1)^2}{\sigma_1^2} + \frac{(y - a_2)^2}{\sigma_2^2} - \frac{2R(x - a_1)(y - a_2)}{\sigma_1\sigma_2} \right], \quad (18)$$

Построение функций (10-15) осуществлялось прямым расчетом двумерных функций распределения Гаусса (16) с их последующей фазификацией путем нормировки максимума функции (16) на единицу.

Поскольку любая фигура «бычий флаг» является частью графика котировок финансового инструмента, её нельзя не рассматривать в контексте тренда, на котором она находится. Несмотря на то, что для трейдера более наглядным является графическое представление данных, для построения модели необходимо параметризовать тренд подобно тому, как была параметризована фигура «флаг». Для сравнения различных трендов вводим параметры: ΔT , ΔC , $\text{tg } \varphi$, h .

ΔC – высота тренда (диапазон изменения котировок валют). ΔC рассчитывается как разность между максимальным и минимальным значением цены валюты:

$$\Delta C = C_{\text{макс}} - C_{\text{мин}}, \quad (19)$$

ΔT – диапазон изменения времени (расстояние в днях по горизонтальной оси от первой до последней точки тренда). ΔT рассчитывается как разность между максимальным и минимальным значением времени:

$$\Delta T = T_{\text{макс}} - T_{\text{мин}}, \quad (20)$$

$\text{tg}(\varphi)$ - тангенс угла наклона тренда. $\text{tg}(\varphi)$ определяется как угол наклона линии поддержки (сопротивления).

Δh_b - толщина тренда по вертикали. Δh_b рассчитывается как разница между котировками (b_{max} , b_{min}), полученными на точках пересечения вертикальной прямой с линией поддержки и линией сопротивления по формуле:

$$\Delta h_b = b_{\text{max}} - b_{\text{min}}, \quad (21)$$

h – нормальная толщина тренда. h рассчитывается как минимальное расстояние между линией поддержки и сопротивления. Для расчета h используется формула:

$$h = \frac{\Delta h_b}{\sqrt{1 + \text{tg}^2 \varphi}}, \quad (22)$$

При сравнении трендов возникают субъективные компоненты, выраженные нечеткими оценками типа «восходящий» тренд, «нисходящий» тренд, «консолидация». Нечеткая логика предлагает решение для подобной ситуации. Вводится функция принадлежности – связь количественного значения с его качественным лингвистическим описанием: $m(x)$. Можно построить следующие функции принадлежности:

- $m(\Delta T)$ – функция принадлежности от диапазона времени тренда,
- $m(\Delta C)$ - функция принадлежности от диапазона цены тренда,
- $m(\text{tg}(\varphi))$ - функция принадлежности от тангенса угла наклона тренда,
- $m(h)$ - функция принадлежности от толщины тренда.

При введении такого количества функций принадлежности размерность задачи сильно увеличивается. Эмпирически известно, что между несколькими параметрами, характеризующими тренд, существует корреляционная связь. Были выявлены следующие корреляционные пары параметров:

- $\Delta C - \Delta T$

- $\text{tg}(\varphi) - h$

Чтобы уменьшить размерность задачи, переходим к построению нечетких функций принадлежности от двух переменных:

- $m(\Delta C, \Delta T)$
- $m(\text{tg}(\varphi), h)$

Функция принадлежности от двух переменных имеет вид трехмерной поверхности, что визуализирует процесс принятия решений.

Нечеткозначные модели фигуры «флаг» и участка ценового ряда «тренд» дают возможность создания общей модели для произвольной ленточной фигуры, которую можно построить на некотором графике функции $F(x)$. Описание этой модели представлено во второй главе диссертации.

ТЕХНОЛОГИЯ МАШИННОГО ОБУЧЕНИЯ В УСЛОВИЯХ ИСПОЛЬЗОВАНИЯ НЕЧЕТКОЙ ИНФОРМАЦИИ

Обучающая выборка флагов определяется человеком, экспертом, который может отделить флаги, дающие верный сигнал, от тех, что дают ложный сигнал. Машинное обучение сводится к расчету параметров фигуры «флаг» и их дальнейшей фазификации.

Модель машинного обучения M_{ML} на произвольной совокупности графических данных представим как n -ку:

$$M_{ML} = \langle D, G, A_p, P, A_f, M, R \rangle, \quad (23)$$

D - множество численных данных графических зависимостей описываемого процесса. Множество D представляет собой набор функциональных отношений (x, y) на области определения $x \in X$. Рассматривается простейший вариант функции от одной переменной $y=f(x)$. Зависимость носит принципиально не аналитический, а стохастический характер.

G - обучающее множество графов-образцов, отобранных экспертом. Каждый граф g из множества G строится на подмножестве численных данных $D_g \subset D$.

P - множество функций распределения, описывающих вероятностные параметры прототипической графической фигуры. Содержание этого множества (его мощность и типы элементов) задается разработчиком модели в процессе параметрического описания фигуры. Создание такого множества представляет собой во многом эвристический

процесс. Множество P получается путем алгоритмического преобразования обучающей выборки $\{Dg\}$: $Ap(\{Dg\}, P)$, где Ap – алгоритмическая процедура расчета функций распределения вероятности.

M – множество функций принадлежности, описывающих нечеткозначные параметры прототипической графической фигуры. Функции M получаются путем фаззификации функций вероятностного распределения с помощью соответствующей алгоритмической процедуры $Aф(P, M)$.

R – множество правил принятия решения. Обычно используется базовое правило в виде логического условия, аргументами которого являются значения функций принадлежности из M .

После того, как сформирован пул исторических данных по инструменту, становится возможным отметить флаги на графике. Эксперт, руководствуясь своими знаниями и опытом, формирует выборку флагов, которая будет служить своеобразным эталоном. В соответствии с моделью фигуры «бычий флаг», производится расчет для каждого элемента выборки. Результатом служит набор значений, характеризующих экспертную выборку. В свою очередь при помощи специальных функций формируются вероятностные данные для инструмента. Они сохраняются в базе данных и используются в дальнейшем как образец для поиска схожих фигур на двумерном графике, содержащем исходные данные.

На рисунке 6 показана схема информационных потоков, обеспечивающая работу системы и эксперта на этапе машинного обучения по образцам.

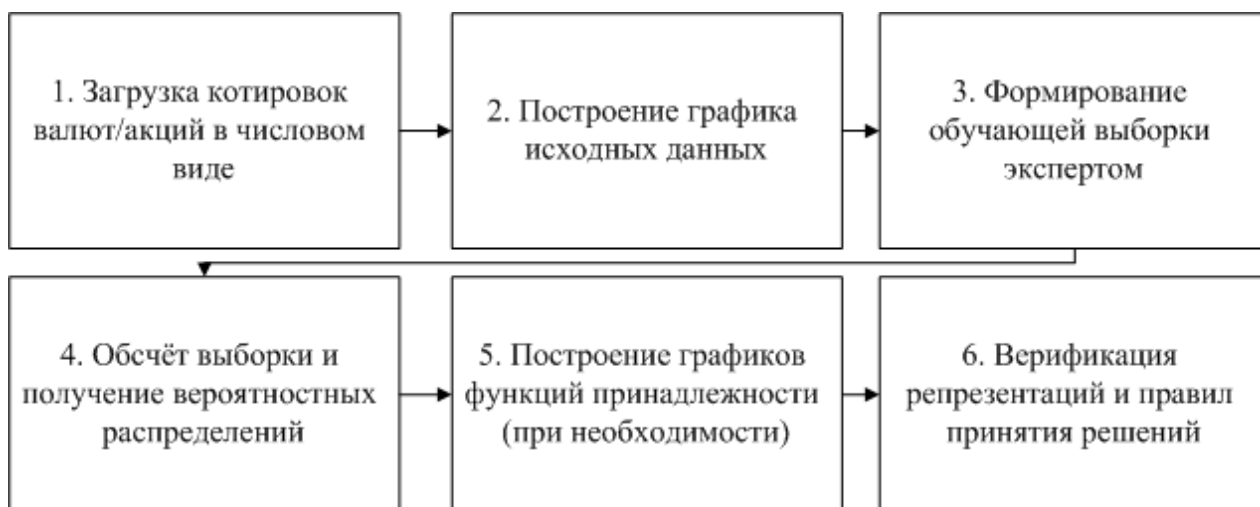


Рисунок 6 - Схема информационных потоков процесса анализа исторических данных в режиме машинного обучения по образцам (пункты 1,2,4,5 выполняются программой, пункт 3 выполняется экспертом, пункт 6 – инженером по знаниям)

Исходя из того, что флаги существуют в контексте трендов, на которых они расположены, то имеет большое значение формирование обучающей выборки для трендов. Она производится образом, схожим с формированием выборки для флагов.

ПОСТАНОВКА ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ НА НЕЧЕТКИХ ДАННЫХ В ГРАФО-АНАЛИТИЧЕСКОЙ ФОРМЕ В РАМКАХ ПОСТРОЕНИЯ СИСТЕМЫ ПРИНЯТИЯ РЕШЕНИЙ

В этой главе представлена постановка задачи оптимизации качества прогнозирования обучаемой системы в графоаналитической форме.

Исходными данными для системы являются исторические данные H , которые представляют собой множество точек на отношении «дата (d_i)-значение(c_i)»:

$$H = \{(d_1, c_1), (d_2, c_1), \dots (d_n, c_n)\} \quad (24)$$

На исторических данных выбираются обучающий и проверочный участки, $H_{об}$ и $H_{пр}$ соответственно. Причём $H_{об} \subset H$, $H_{пр} \subset H$, $H_{пр} \cap H_{об} = \emptyset$.

Экспертом формируется исходная обучающая выборка, представляющая собой набор примеров, впредь используемых как образец. Каждый элемент обучающей выборки представляет собой фигуру на графике, которую также можно представить как множество точек «дата-значение». Конкретный вид фигуры (или нескольких фигур) напрямую зависит от выбранной области применения. Например, для проверки данной модели на финансовом рынке были введены нечёткозначные модели фигуры «флаг» и участка ценового ряда «тренд». Таким образом, обучающую выборку, созданную экспертом, можно выразить как $S = \{S_i : S_i \subset H_{об}\}$.

Следующим шагом будет вычисление параметров модели. Для этого рассчитываются параметры каждой фигуры, а также параметры всей выборки в целом.

Третьим этапом является получение прогнозов. Прогноз, в данном контексте, это такое событие, возникшее на каком-то участке функции $H_{пр}$, когда выявленная фигура на этом участке попала в диапазон допустимых значений обучающей выборки S . Суть прогноза подразумевает, что он либо сбудется (будет успешным), либо нет (будет ложным). Соответственно, сбывшиеся прогнозы обозначим как N_+ , а не сбывшиеся N_- . Также существуют нейтральные прогнозы N_0 , которые в контексте финансового рынка означают неопределённое (или консолидированное) движение рынка после точки

выявления этого прогноза. Таким образом, $N_+ + N_- + N_0 = N_T$, где N_T означает полное количество прогнозов на заданном участке из множества H .

На следующем этапе нужно ввести понятие качества прогноза Z . Достижение наилучшего качества прогнозов можно получить, только углубившись в конкретную предметную область, чтобы учесть её приоритеты. Так, например, на финансовом рынке огромную роль играет получение прибыли путём своевременного закрытия позиций, ибо даже одна передержанная позиция, не закрытая вовремя, может принести своему владельцу столь высокие убытки, что они уничтожат всю полученную прибыль по другим позициям. Следовательно, нужно сделать акцент на уменьшении убытков. В данном случае достижение максимального качества прогноза выглядит так:

$$\max_{n \subset N_-} Z(n), N_- \rightarrow 0, N_- = N_t : N_- \in N_t, N_t > 0, \quad (25)$$

В других предметных областях, в частности, медицине, иные приоритеты, что неизбежно влечёт изменение критериев оценки качества прогноза. Выше немало внимания уделялось важности методов электрокардиографии и реографии для раннего выявления инфарктов и инсультов. Используя модель прогнозирования, предложенную в данной работе, становится возможным построить систему, которая могла бы по виду электрокардиограммы и реограммы предсказывать появление этих крайне опасных состояний. Приоритеты данной предметной области таковы, что лучше допустить возникновение ложного прогноза (в медицинских терминах это означает «поставить неверный диагноз»), чем пропустить истинный, ведь это может повлечь смерть человека. Следовательно, достигнуть максимального качества прогнозов в этом случае можно при соблюдении следующих условий:

$$\max_{n \subset N_+} Z(n), \max N_+, N_+ = N_t : N_+ \in N_t, N_t > 0, \quad (26)$$

Другими словами, нужно пытаться выявить максимальное количество положительных прогнозов, уделяя гораздо меньше внимания ложным.

Общее выражение для Z можно записать так:

$$Z = \frac{P_r \cdot R_e}{\alpha \cdot P_r + \beta \cdot R_e}, \quad (27)$$

, где: α и β – нормировочные коэффициенты; P_r – показатель точности; R_e – показатель отказоустойчивости. В контексте двух примеров, рассмотренных выше, можно

сделать вывод, что в случае прогноза на финансовом рынке нужно установить нормировочный коэффициент $\beta = 0$, чтобы качество Z было прямо пропорционально отказоустойчивости R_e . В случае прогноза по данным одного из методов медицинской диагностики, нужно установить нормировочный коэффициент $\alpha = 0$, чтобы качество Z было прямо пропорционально точности P_r .

$$P_r = \frac{TP}{TP + FP}, \quad (28)$$

$$R_e = \frac{TP}{TP + FN}, \quad (29)$$

, где: TP – количество сбывшихся положительных прогнозов, FP – количество несбывшихся положительных прогнозов, FN – количество несбывшихся отрицательных прогнозов.

На рисунке 7 представлена блок-схема решения задачи машинного обучения на нечётких данных в графоаналитической форме в рамках построения системы принятия решений.

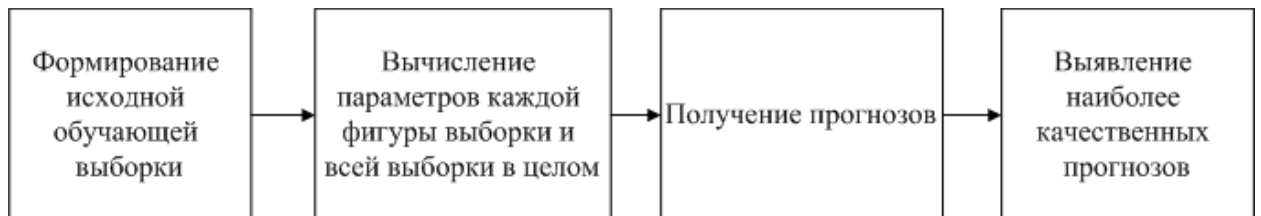


Рисунок 7 - Блок-схема решения задачи машинного обучения.

МОДЕЛИ И АЛГОРИТМЫ РАСПОЗНАВАНИЯ ФИГУР НА ГРАФИКАХ В ДИНАМИЧЕСКОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

В данной главе описаны модели и алгоритмы распознавания фигур на графиках в динамической предметной области.

Для принятия эффективных решений на основе заданной модели обучения необходимо выработать соответствующие правила. В задаче прогнозирования рынков по фигуре «флаг» решением является приказ об открытии новой позиции на основании прогноза, полученного путем анализа текущих котировок на рынке. При анализе выявляется наличие древка и полотнища, оцениваются их качественные характеристики путем сравнения с эталоном, заданным совокупностью функций принадлежности, и в случае положительных характеристик, отдается приказ (ордер) об открытии позиции.

В качестве элементов, составляющих правила, сначала были введены бинарные условия сравнения $A = \{ A_1, \dots A_i, \dots A_6 \}$:

$$A_i = m_i \geq 0,5, \quad (30)$$

где m_i – значение функции принадлежности на текущую дату.

Затем были сформулированы непосредственно правила, которые составили следующее множество:

$$R = \{R_{\otimes}, R_{\oplus}, R_{c1} \dots R_{cn}\}, \quad (31)$$

где $R_{\otimes} = A_1 \wedge A_2 \dots \wedge A_i \dots \wedge A_6$ – конъюнктивная форма, сконструированная на множестве A (жесткое правило);

$R_{\oplus} = A_1 \vee A_2 \dots \vee A_i \dots \vee A_6$ – дизъюнктивная форма, сконструированная на множестве A (мягкое правило);

R_{c_j} – произвольные дизъюнктивно-конъюнктивные формы, сконструированные на множестве A (полужесткие правила);

Лучшие правила выбираются из множества R эмпирическим путем в процессе тестирования модели и анализа результатов прогнозирования.

Алгоритм принятия решения выглядит следующим образом:

- выявление древка;
- выявление полотнища;
- выбор правила принятия решения.



Рисунок 8 – Методика прогнозирования на основе нечеткозначной модели «флаг» и метода «FuzGraph».

Схема комплексного анализа, учитывающего исторические выборки флагов и трендов, может выглядеть следующим образом:

- выявление «бычьего флага»;
- выявление «UP-тренда»;
- выбор комплексного правила принятия решения.

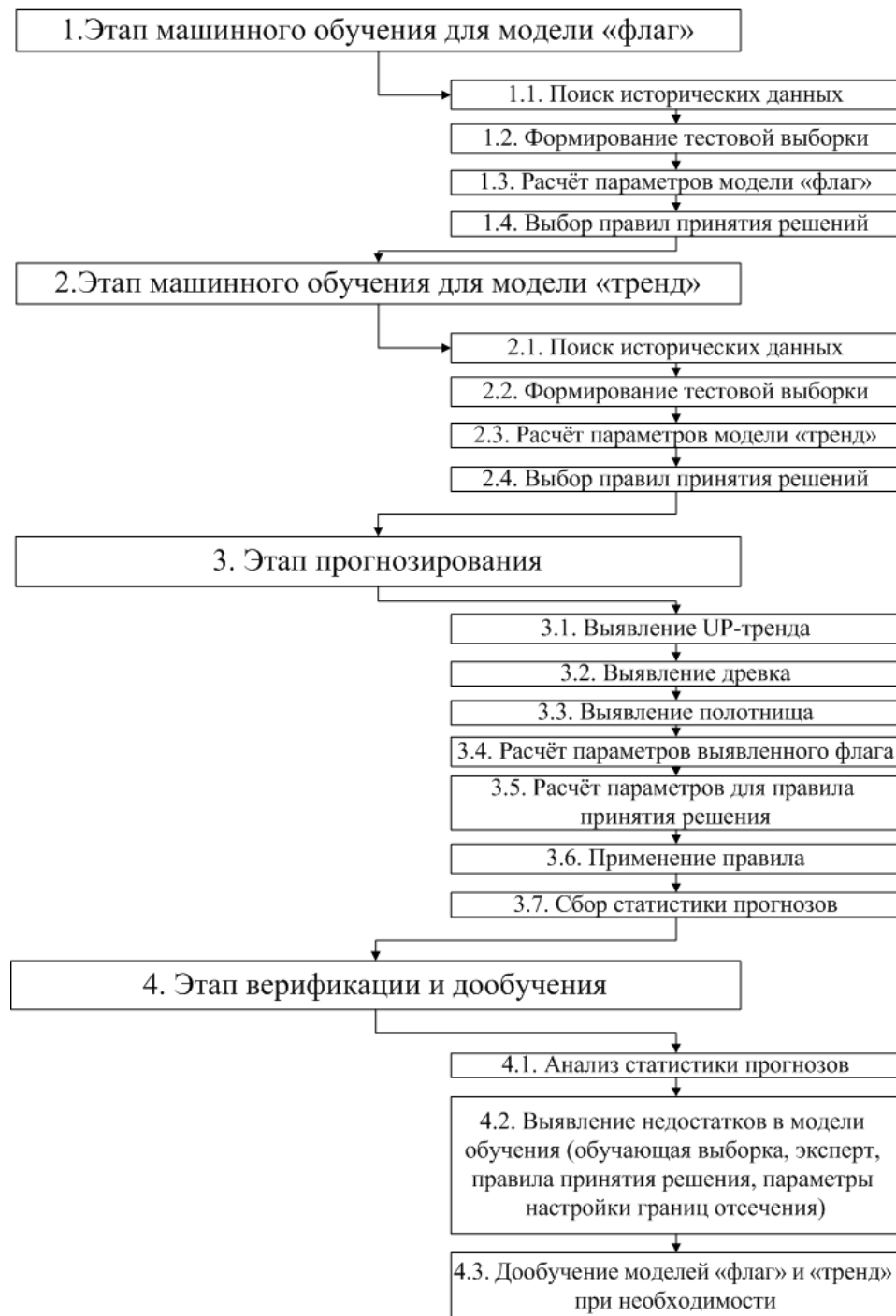


Рисунок 9 - Методика прогнозирования на основе комплексной работы с моделями «бычий флаг» и «UP-тренд»

Опыт существующих механических торговых систем показывает с одной стороны эффективность различных индикаторов, а с другой стороны – преимущество именно

комплексного подхода. Отсюда следует, что, используя комплексный подход, можно добиться выявления таких сигналов, посылаемых рынком, которые бы с большой долей вероятности гарантировали успех.

ПРОГРАММНЫЙ КОМПЛЕКС ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЯ «ЕМ»

Программный комплекс «ЕМ» использует нечёткозначные модели, описанные в данной работе, а также алгоритмы поиска фигур на графиках котировок финансовых инструментов.

Работа с финансовым инструментом начинается с формирования обучающей выборки. Эксперт, просматривая исторические данные этого инструмента, отмечает бычьи флаги и тренды. По этим данным рассчитываются наборы статистических данных, сохраняемые в базе данных.

Блок анализа «ЕМ» работает в двух режимах: прогноз за период и анализ на текущую дату. Прогноз за период даёт возможность проанализировать участок исторических данных и выявить на нём флаги. Эту функцию можно использовать в качестве проверки эффективности обучающей выборки для данного финансового инструмента. Анализ на текущую дату использует текущую дату как отправную точку для прогнозирования, что даёт возможность получать прогнозы в режиме реального времени. Вне зависимости от выбранного режима, блок анализа позволяет наложить на выборку найденных фигур набор правил. Используя «жёсткое» правило можно оставить только те флаги, которые почти наверняка принесут прибыль. Использование «мягкого» правила оставляет в выборке максимальное количество флагов, однако многие из них могут быть ложными. Другие правила являются различными соединениями функций принадлежности. Комбинирование правил позволяет гибко настроить систему под особенности конкретного инструмента.

«ЕМ» обладает способностью объединять инструменты в группы и проводить групповой анализ. Чтобы получить максимальную прибыль при минимуме потерь достаточно выставить для каждого инструмента «жёсткое» правило, и провести анализ для группы таких инструментов. Небольшая, но гарантированная прибыль по каждому инструменту в совокупности с большим количеством инструментов позволяет достичь оптимального результата.

ЗАКЛЮЧЕНИЕ

В ходе данной работы было проведено исследование поведения спорадически осциллирующих систем в динамической предметной области, представляемых в виде двумерных графиков, а также проблем применения технологии машинного обучения в таких средах. Была создана информационная система поддержки принятия решений «ЕМ», основанная на нечёткозначных моделях и машинном обучении.

В ходе данной работы выяснилось, что в настоящее время множество решаемых задач по созданию информационных систем поддержки принятия решений, применяемых в различных областях человеческой деятельности, сопряжены с обработкой графических данных в качестве исходной информации. В свою очередь, данная работа была направлена на создание моделей и алгоритмов, которые позволяли бы выявлять фигуры на двумерных графиках как в реальном времени, так и в ретроспективе на исторических данных. Областью применения данных разработок являются компоненты АСУ и АСУ ТП металлургических предприятий, медицинские приборы, системы финансового анализа. В ходе апробации моделей на финансовом рынке были представлены нечёткозначные модели фигуры «флаг» и участка ценового ряда «тренд», что позволило в свою очередь предложить методику машинного обучения «FuzGraph» и алгоритм выявления этих фигур на реальных данных. Также была предложена общая модель произвольной ленточной фигуры, частными случаями которой являются фигура «флаг» и участок ценового ряда «тренд». Наличие такой модели даёт возможность выявлять на графиках не только «флаг» и «тренд», но также произвольные фигуры, которые могут быть представлены как набор последовательных отрезков.

Созданные в ходе данной работы модели, а также опытный прототип программы «ЕМ», позволили наглядно убедиться в применимости и эффективности машинного обучения и нечёткозначных систем принятия решений не только для металлургии, но и для других областей науки и бизнеса.

Проделанная работа привела к следующим результатам и выводам:

1. Выполнена формальная постановка задачи принятия решения на основе нечетких данных в графоаналитической форме. Формализация учитывает влияние на качество системы таких факторов как точность и отказоустойчивость, что, вместе с механизмом управляющих правил, позволяет гибко приспосабливать модель под нужды конкретной предметной области.
2. Разработаны и описаны нечеткозначные модели фигуры «флаг» и участка

ценового ряда «тренд», а также дано формальное описание общей нечёткозначной модели произвольной ленточной фигуры.

3. Произведена формальная постановка задачи машинного обучения на нечётких данных. В качестве модели описания (репрезентации) выборки примеров была использована нечеткозначная формализация геометрических фигур.
4. Предложен метод и алгоритм распознавания фигур на графиках в динамической предметной области, а также алгоритмическое решение задач прогнозирования на финансовых рынках.
5. Исследованы корреляционные зависимости между различными параметрами графических фигур, учитываемых в процессе принятия решений.
6. Создан алгоритм поиска линий поддержки и сопротивления.
7. На основе моделей и алгоритмов, предложенных в данной работе, создан и апробирован опытный образец системы поддержки принятия решений «ЕМ».

Теоретические вопросы диссертации освещаются в следующих научных публикациях, в том числе одна публикация в издании, рекомендованном ВАК:

- В.Н. Поляков, Е.Н. Леонов. Метод машинного обучения на графических данных (Тезисы) - Труды международной конференции «Когнитивное моделирование в лингвистике 2005». Изд. Учеба, М.-Варна, 2005. с.282.
- В.Н. Поляков, Е.Н. Леонов, И.В. Поляков. Формирование инвестиционного портфеля и выработка стратегии принятия решения на фондовом рынке, основанной на нечёткозначной модели фигуры «флаг». Деп. ВИНТИ – Москва, 2006 – 63 с.
- В.Н. Поляков, Е.Н. Леонов, И.В. Поляков. Построение комплексной модели принятия решения на финансовых и фондовых рынках на основе взаимосвязи технических факторов – Труды 8-ой международной конференции «Распознавание образов и анализ изображений» (РОАИ-8-2007), 2007 г., Том 3. с. 100.
- В.Н. Поляков, Е.Н. Леонов. Метод машинного обучения на графических данных. Учёные записки Казанского государственного университета. Серия Физико-математические науки. 2007. Том 149, книга 2, с. 92-104.