

МЕТОД МАШИННОГО ОБУЧЕНИЯ НА ГРАФИЧЕСКИХ ДАНЫХ

Леонов Е.Н., Поляков В.Н.

АННОТАЦИЯ

В работе рассматривается метод машинного обучения FuzGraph, основанный на нечеткозначном представлении графических данных. Образцы повторяющихся изображений на графике $y=f(x)$ описываются в виде модели в пространстве нечетких функций принадлежности. Эти функции в свою очередь являются результатом фаззификации распределений плотности вероятности на параметрах модели выявляемого изображения. Метод машинного обучения FuzGraph удобен для обнаружения закономерностей на графических данных, имеющих стохастическую природу. В частности, тестирование и апробация метода проходила на актуальной для финансовых рынков задаче прогнозирования цен валют и акций по фигуре «флаг».

ВВЕДЕНИЕ

Исследования по теме «машинное обучение» начались в середине 50-х годов [7, 10]. Изначально машинное обучение рассматривалось как процесс создания алгоритмов, которые бы позволяли компьютерам развивать их представления и базы знаний. Впоследствии эта тематика получила свое дальнейшее развитие в рамках теории искусственного интеллекта. Существует ряд основных подходов к этому методу [10]: обучение на примерах [5], искусственные нейронные сети [11], генетические алгоритмы [2], объясняющее обучение [6], эволюционные гипотезы [2, 3, 4], вывод на прецедентах [5]. Хотя полной теории того, как люди и машины учатся, еще не предложено, в рамках перечисленных подходов сформулированы достаточно строгие модели обучения.

Наиболее популярная область исследований в машинном обучении это изучение понятий и других репрезентаций на примере. Ввод для таких обучающихся программ содержит описания примеров. Вывод состоит из разных видов репрезентаций, которые кодируются обобщенно для этих примеров. Обобщенные репрезентации могут быть понятиями, функциями, правилами или деревьями описаний, которые обеспечивают удобные средства классификации новых примеров.

В работе описан метод машинного обучения на примерах, представленных в виде графических данных. В качестве модели описания (репрезентации) выборки примеров была использована нечеткозначная формализация геометрических фигур. Выбор фигур для обучения производится человеком, экспертом в предметной области. Машинное обучение сводится к расчету параметров вероятностного распределения и их дальнейшей фаззификации. Принятие решений на основе полученных нечеткозначных описаний геометрических фигур осуществляется на основе системы правил, позволяющих фильтровать «хорошие» фигуры из текущего множества данных в процессе работы пилотной версии системы. Фактически классификация каждого нового случая осуществляется путем сопоставления текущего представления с системой нечетких функций принадлежности для требуемого класса объектов. Испытание метода на графических данных котировок валют и акций показало высокий уровень совпадения прогнозов. Метод получил название FuzGraph.

2. МОДЕЛЬ МАШИННОГО ОБУЧЕНИЯ

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

$$M_{ML} = \langle D, G, A_p, P, A_f, M, R \rangle \quad (1)$$

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 $Dg \subset D$.

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 $\{Dg\}$: $A_p(\{Dg\}, P)$, где A_p – 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 $A\phi(P, M)$.

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

3. РЕАЛИЗАЦИЯ МЕТОДА

Рассмотрим реализацию метода машинного обучения на графических данных FuzGraph на примере системы EM, предназначенной для формирования инвестиционного портфеля и выработки стратегии принятия решений на фондовом рынке. В системе для прогнозирования используется нечеткозначная модель фигуры «флаг».

В работе метод FuzGraph применялся для обучения системы на примерах, являющих собой графические образы. Обработка параметров обучающих примеров давала набор вероятностных характеристик, по которым строились функции принадлежности и, в дальнейшем, производился поиск фигур на графике текущих данных.

Фигурой в теории технического анализа называется графический образ на графике котировок валют или акций, напоминающий геометрическую фигуру. Обычно в техническом анализе различают три основных вида фигур: «флаг», «треугольник», «голова - плечи». Есть и другие фигуры, но они встречаются значительно реже [8].

Целью разработки была автоматизация известного метода предсказания котировок, основанного на фигуре «флаг». Сложность автоматизации этого метода заключается в том, что предсказание по фигурам носит визуальный, зачастую субъективный характер, и до сих пор не поддавалось компьютеризации.

Система обрабатывает двумерные графики котировок финансовых инструментов. Задача такой обработки – привести сложно формализуемые графические образы к понятным системе числовым показателям.

Опишем модель фигуры «бычий флаг». Если рассматривать среднесрочные прогнозы, флаг – это кратковременная фигура, длящаяся несколько дней (рис. 1). Обычно он возникает на динамичных рынках, где бывают резкие изменения цен. «Бычий флаг» образуется при восходящем тренде, «медвежий флаг» – при нисходящем. Свое название эта фигура получила потому, что напоминает флаг, в котором есть древко, т.е. узкий и высокий участок, и полотнище, т.е. сравнительно

протяженный участок колебаний цен в виде полосы с легким наклоном. Метод предсказания основывается на том, что обычно фигура «флаг» на графике котировок предшествует бурному росту цены (см. рис. 1).

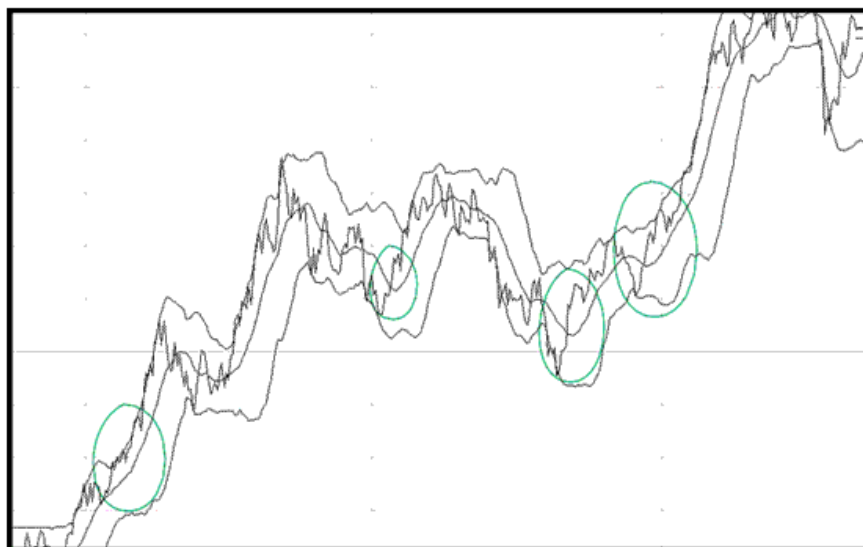


Рис. 1. Примеры образования флага на реальных графиках валют.

Параметрами флага в соответствии с рис. 2 являются:

- Длина по оси времени или количество дней, в течение которых древко или полотнище образуется. Мы будем обозначать эти параметры 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;
- Нормальная толщина канала, образованного древком и полотнищем, обозначается h_d и h_n соответственно.

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 нечетких функций принадлежности [13]. В качестве «лобового» решения можно было бы сконструировать одну многомерную функцию принадлежности. Однако это снизило бы наглядность и, как следствие, возможность промежуточного контроля результатов компьютерного моделирования. Учитывая

сравнительно большое число параметров в модели (восемь), был проведен анализ парной корреляции параметров модели, и было принято эвристическое решение строить несколько двумерных функций принадлежности. Также было сделано предложение, что распределение вероятности параметров фигуры "флаг" подчиняется нормальному распределению и описывается системой функций Гаусса для двумерного нормального распределения. Вследствие числового характера исходных данных котировок валют и акций, расчет характеристик для определенных таким образом нечетких функций принадлежности параметров фигуры «флаг» уже не составляет большого труда [9].

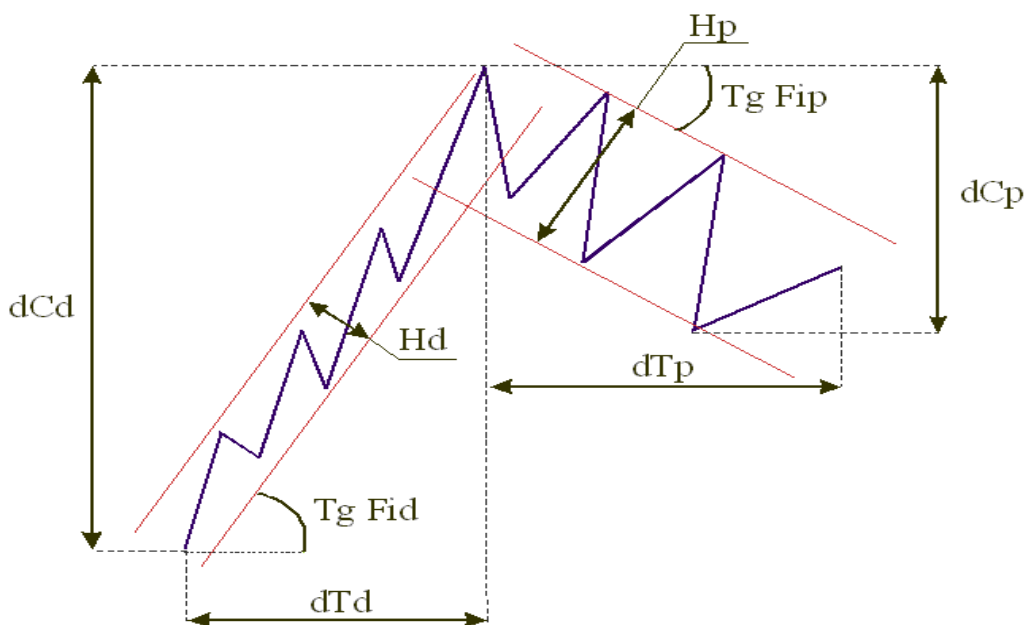


Рис. 2. Параметры модели фигуры «бычий флаг».

Для получения вероятностных характеристик котировок валют и акций применяется технология машинного обучения на выборке фигур, составленной опытным специалистом по рынку. Поэтому основные требования, которые предъявляются к программному комплексу – это гибкость представления графических данных, удобный интуитивно понятный интерфейс. На рисунке 3 показана схема информационных потоков, обеспечивающая работу системы и эксперта на этапе машинного обучения по образцам.

Приведенная схема в программном комплексе ЕМ реализована в несколько этапов. В первую очередь необходимо ввести данные в систему. Для этого используется программа, обрабатывающая специально сформированные шаблоны MS Excel. На рисунке 4 показан пример шаблона для ввода текущих данных.



Рис. 3. 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).

Инструмент	Идентификатор	Состояние инструмента	Состояние позиции	Статистика	Результат последнего прогноза	Котировка
1	2	3	4	5	6	7
Alcoa Inc	AA	готов к прогнозу	нет данных	Закрыто по LO - 0% Закрыто по SL - 0% Закрыто принудительно - 0%	позиции не открывались	30,39
American International Group Inc	AIG	готов к прогнозу	нет данных	Закрыто по LO - 0% Закрыто по SL - 0% Закрыто принудительно - 0%	позиции не открывались	55,41
Amer Express Co	AXP	готов к прогнозу	нет данных	Закрыто по LO - 0% Закрыто по SL - 0% Закрыто принудительно - 0%	позиции не открывались	51,37
Boeing Company The	BA	готов к прогнозу	нет данных	Закрыто по LO - 0% Закрыто по SL - 0% Закрыто	позиции не открывались	58,46

Рисунок 4. Шаблон для ввода котировок (овалом выделена область ввода данных оператором).

Также можно ввести данные за весь период обучения. Для этого также существует специальный модуль, который загружает файл MS Excel, содержащий соответствующие данные (Рис.5).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	CAT														
2	2-Jan-62	38,5	38,88	38,13	38,5										
3	3-Jan-62	38,5	38,88	38,13	38,88										
4	4-Jan-62	39,75	41	39,75	39,88										
5	5-Jan-62	39,88	40,75	39,75	40,25										
6	8-Jan-62	40,25	40,88	40	40,5										
7	9-Jan-62	40,5	41,25	40,13	40,88										
8	10-Jan-62	40,88	41,13	40,25	40,38										
9	11-Jan-62	40,38	41	40	41										
10	12-Jan-62	41	41,75	40,5	41,25										
11	15-Jan-62	41,25	41,5	40,5	41,13										
12	16-Jan-62	40,75	40,75	40,25	40,5										
13	17-Jan-62	40,5	41	40,25	40,38										
14	18-Jan-62	40,38	41,25	40,38	41,25										
15	19-Jan-62	41,25	42	40,5	42										
16	22-Jan-62	42	42,5	41,88	41,88										
17	23-Jan-62	41,88	42,25	41	41										
18	24-Jan-62	41	41	40	40,75										
19	25-Jan-62	40,75	41,38	40,63	41										
20	26-Jan-62	41	41,38	40,5	41										
21	29-Jan-62	41	41,38	40,88	41,38										
22	30-Jan-62	41,25	41,25	40,63	40,75										
23	31-Jan-62	40,75	41,38	40,75	41										
24	1-Feb-62	41	42	41	41,75										
25	2-Feb-62	41,75	41,88	41,38	41,5										
26	5-Feb-62	41,5	41,63	41,38	41,38										
27	6-Feb-62	41,5	41,88	41,5	41,63										
28	7-Feb-62	41,63	41,88	41,13	41,13										
29	8-Feb-62	41,13	41,25	41,13	41,13										
30	9-Feb-62	41,13	41,25	40,63	40,63										
31	12-Feb-62	40,63	40,75	40,13	40,75										
32	13-Feb-62	40,75	40,88	40,38	40,63										

Рисунок 5. Файл с исходными данными за весь период обучения.

После того, как сформирован пул данных по инструменту, мы можем перейти к следующему пункту схемы, изображенной на рисунке 3, т.е. к заданию флагов на графике. Этот процесс показан на рисунках 6 и 7. Эксперт, руководствуясь своими знаниями и опытом, формирует выборку флагов, которая будет служить своеобразным эталоном. В соответствие с моделью фигуры «флаг», описанной выше, 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. Таким образом, происходит машинное обучение на выборке фигур.

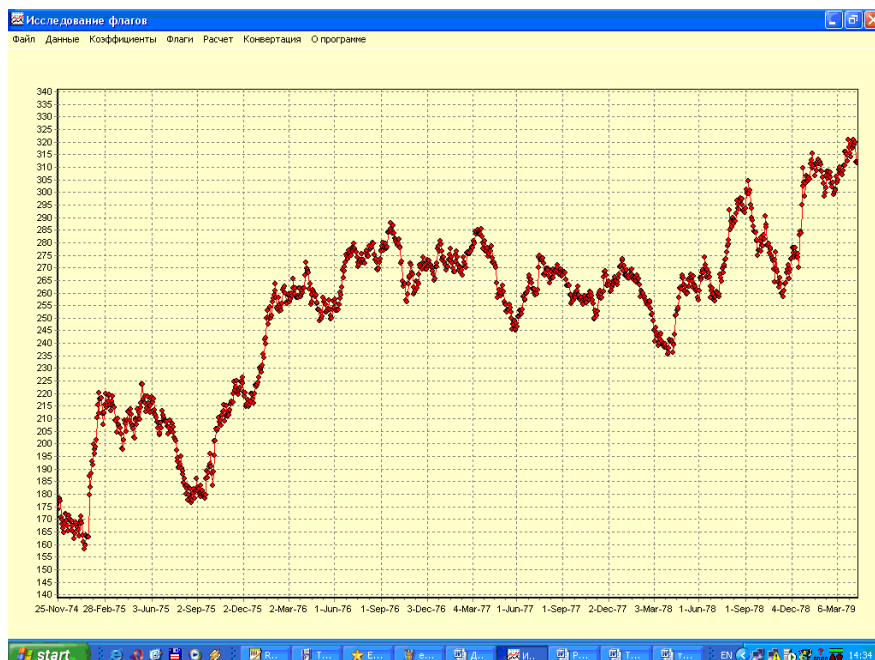


Рисунок 6. Исторический график изменения котировок (акции IBM) за период 1962-2004 гг.

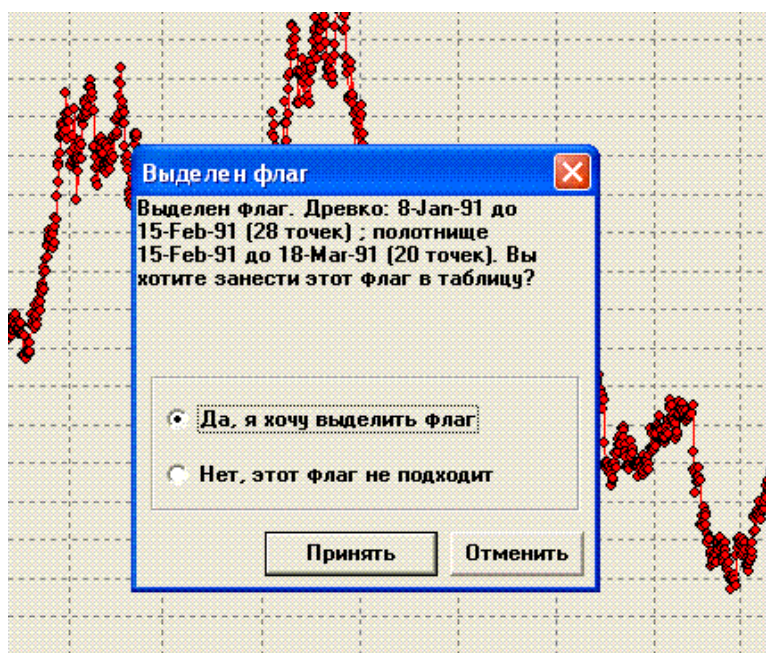


Рисунок 7. Выделение флага.

Результаты расчета параметров вероятностного распределения представлены на рис. 8.

Статистические данные								
Файл								
	dTd	dCd	Tgd	Hd	dTp	dCp	Tgp	Hp
MAX	2400	3,063E5	1,75	992,62	1900	2,037E5	-1	827,74
MIN	400	43700	0,72833	215,76	400	40000	-1,2716	224,02
SIG	1,2133E6	2,0915E1	0,26141	1,9325E5	6,6333E5	7,3911E5	0,019802	1,053E5
M	1133,3	1,4E5	1,2515	485,34	966,67	1,0667E5	-1,1146	457,13
k(dTd-dC)	0,98349							
k(dTp-dC)	0,99599							
k(dTd-dT)	-0,30467							
k(dCd-dC)	-0,39121							
k(Tgd-Hd)	0,078063							
k(Tgp-Hp)	0,11636							

Рис. 8. Пример расчета вероятностных характеристик обучающей выборки флагов (по данным котировок акций компании IBM).

Функции принадлежности для фигуры флаг могут быть визуализированы в трехмерном пространстве. На рисунках 9-14 в качестве иллюстрации приведены графики функций принадлежности для швейцарского франка, построенные на основании статистического анализа котировок валюты за десятилетний период (1990-1999 гг.). Для большей наглядности графики приведены в виде линий уровня (слева) и в изометрической проекции (справа).

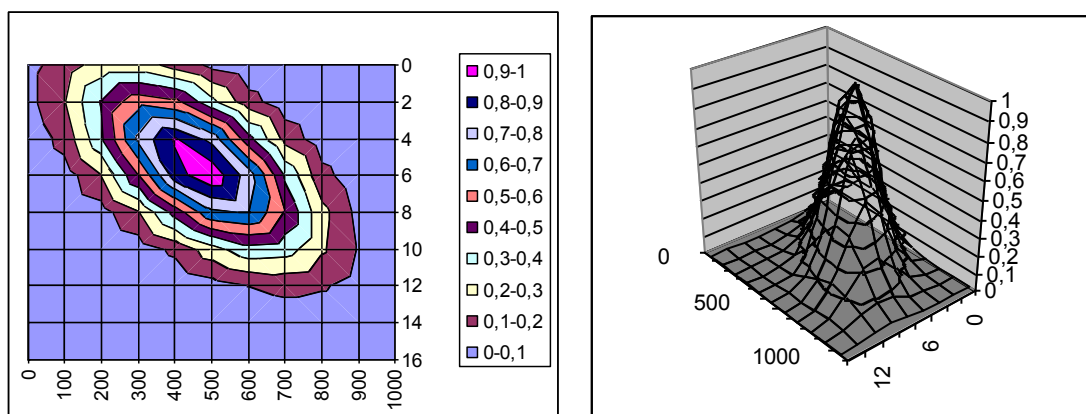


Рис. 9. Функция принадлежности m_1 (ΔC_d , ΔT_d). По осям ординат отмечены изменение котировок (в пунктах) и длительность (в днях) для древка.

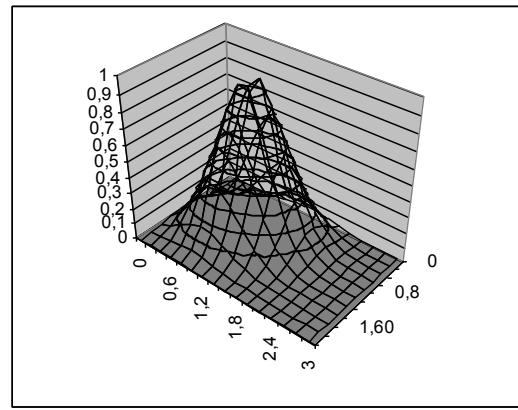
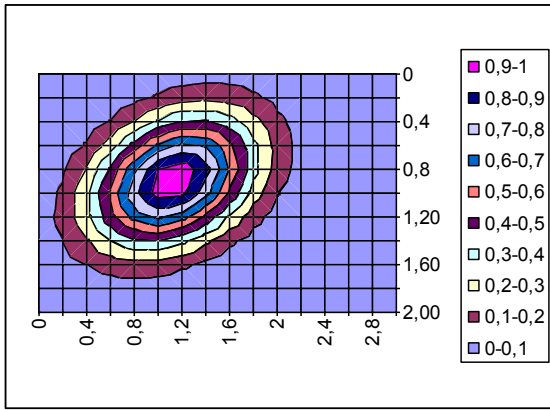


Рис. 10. Функция принадлежности $m_2(\text{tg}(\varphi_d), h_d)$. Данные по оси ординат в безразмерных единицах.

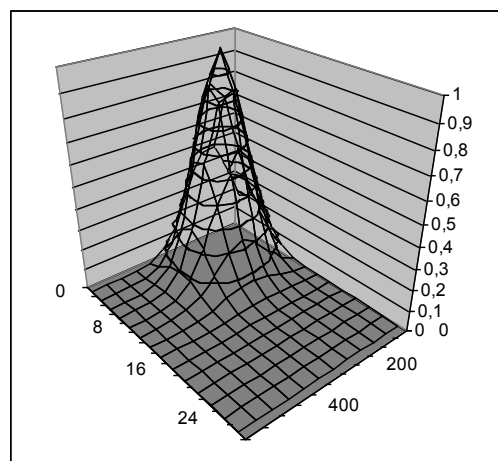
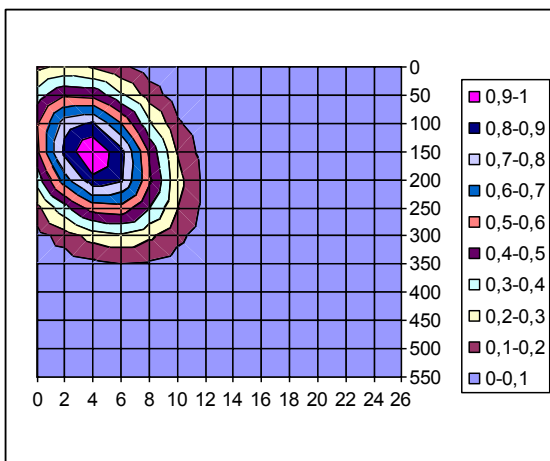


Рис. 11. Функция принадлежности $m_3(\Delta C_n, \Delta T_n)$. По осям ординат отмечены изменение котировок (в пунктах) и длительность (в днях) для полотнища.

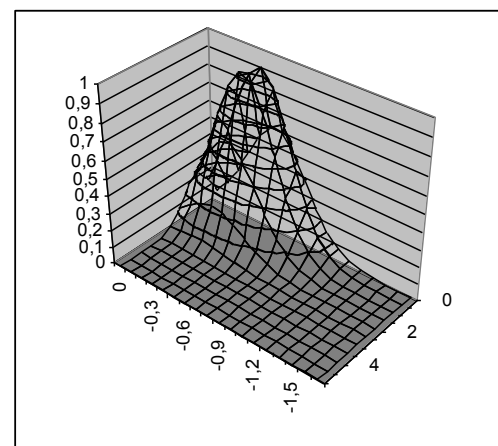
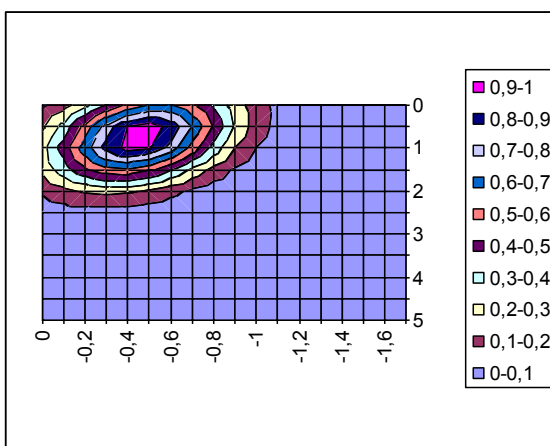


Рис. 12. Функция принадлежности $m_4(\text{tg}(\varphi_n), h_n)$. Данные по оси ординат в безразмерных единицах.

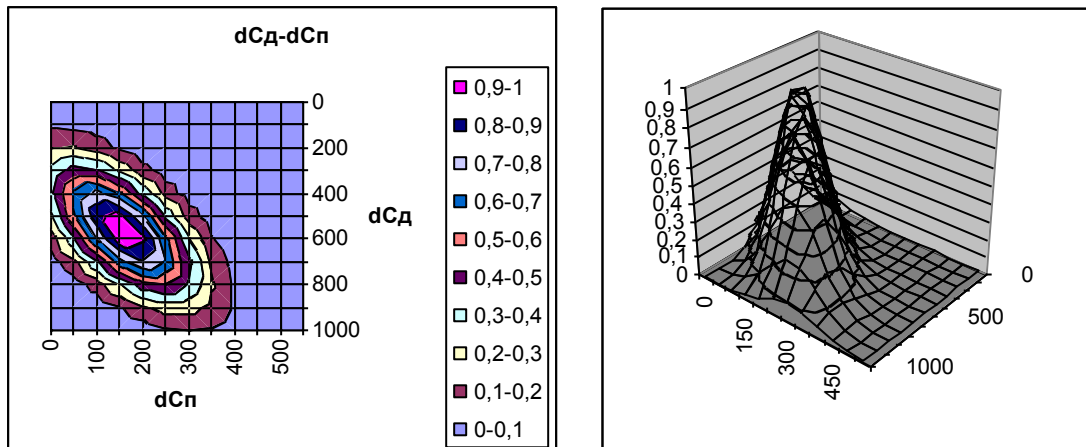


Рис. 13. Функция принадлежности $m_5(\Delta C_d, \Delta C_n)$. По осям ординат отмечены изменение котировок (в пунктах) для деревка и для полотнища.

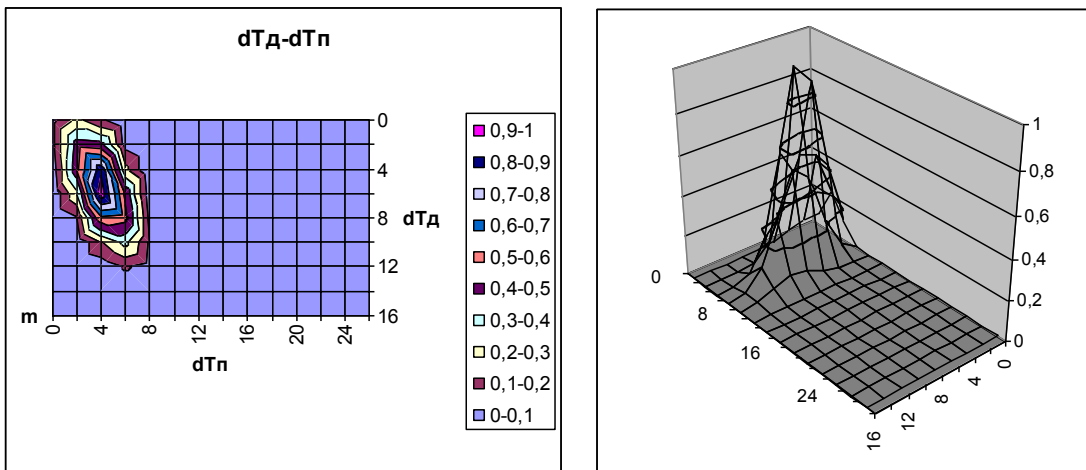


Рис. 14. Функция принадлежности $m_6(\Delta T_d, \Delta T_n)$. По осям ординат отмечены длительность (в днях) для деревка и для полотнища.

Из семерки (1) следует, что для принятия эффективных решений на основе заданной модели обучения необходимо выработать соответствующие правила. Для задачи прогнозирования рынков по фигуре «флаг» решением является приказ об открытии новой позиции на основании прогноза, полученного путем анализа данных о текущих котировках, поступающих с рынка. При анализе выявляется наличие деревка 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 $A = \{ A_1, \dots, A_i, \dots, A_6 \}$:

$$A_i = m_i \geq 0,5, \quad (2)$$

где m_i – значение функции принадлежности на текущую дату.

Затем были сформулированы непосредственно правила, которые составили следующее множество:

$$R = \{R_{\otimes}, R_{\oplus}, R_{c1} \dots R_{cn}\}, \quad (3)$$

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

$R_{\oplus} = A_1 \vee A_2 \dots \vee A_i \dots \vee A_n$ – 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.

Системы EM реализована в среде программирования Delphi 5.0 с использованием СУБД MS SQL для хранения данных.

РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ

Метод FuzGraph проходил апробацию на данных котировок валют на рынке Forex и данных котировок акций американских компаний, входящих в индекс Доу-Джонс.

Первоначально эффективность метода машинного обучения оценивалась на швейцарском франке на рынке Forex. При этом обучающие данные выбирались на историческом периоде котировок CHF/USD за 1990-1999 гг. Метод тестировался на периоде 2000-2002 г.г., на котором также экспертом были отобраны эталонные и «ложные» флаги. Для жесткого правила эффективность прогнозов «на отказ» на тестовой выборке данных за 2000-2002 г. составляла 100 %, то есть ни одного «ложного» флага программой выбрано не было. Вместе с тем эффективность прогнозов «на точность» была сравнительно низкой: программа предсказала только 12,5 % случаев, отмеченных вручную экспертом. В то же время мягкое правило срабатывает гораздо чаще: эффективность прогнозов «на точность» составила 100 %, но платой за это явилась большее количество ошибочных прогнозов. В случае применения «мягкого» правила система пропустила 22,2 % «ложных» флагов. Задачи прогнозирования рынков очень чувствительны к ошибкам классификации «на отказ», так как потери от неудачных позиций могут наносить

ощутимый ущерб рыночным игрокам. Поэтому в реальных системах предпочтительным является использование более жестких правил.

При тестировании системы EM на компаниях индекса DJ эффективность оценивалась по другой методике. Система также обучалась на исторической выборке, но эталонная выборка экспертом для тестирования не формировалась. Все «флаги», отобранные программой, делились по факту на «хорошие» и «плохие» и считалось итоговое соотношение эффективности прогноза для лучшей обучающей выборки. «Хорошим» считался «флаг», для которого прогноз сбывался, «плохим» – для которого прогноз оказывался неудачным. В случае неудачного результата ($N_x/N < 0,5$) проводилось повторное обучение.

В таблице 1 приведены результаты тестирования метода в режиме *in vitro* на реальных данных. Эти результаты получены с применением одного из полужестких правил.

Таблица 1. Результаты тестирования для компаний из индекса Доу-Джонс.

Компания	Обучающая выборка за период, г.	Число флагов в обучающей выборке	Число итераций обучения	Период тестирования, г.	Число обнаруженных флагов N	Число «хороших» флагов N_x	Число «плохих» флагов, N_p	Соотношение N_x/N , %	Число флагов на год
AA	1962-1980	40	5	1982-2004	40	32	8	75 %	1,7
AIG	1984-1993	30	1	1994-2004	9	8	1	89 %	0,8
BA	1964-1978	27	3	1983-2004	64	37	27	58 %	3
C	1981-1991	14	7	1992-2004	43	28	15	65 %	3,3
CAT	1962-1974	47	2	1977-2004	10	7	3	70 %	0,3
DIS	1977-1980	24	1	1986-1998	37	25	12	68 %	2,8
GE	1982-1987	10	3	1988-2004	64	41	23	64 %	3,8
GM	1982-1994	66	2	1995-2004	13	8	5	61,5 %	1,3
HD	1984-1996	43	1	1997-2004	5	4	1	85 %	0,6
HON	1970-1983	35	1	1994-2004	23	16	7	69,6 %	2,1
HPQ	1966-1979	39	1	1980-2004	10	6	4	60 %	0,4

Для всех компаний, вошедших в тестовый пул, с помощью метода FuzGraph удалось получить позитивные тестовые результаты. Отметим, что самый худший результат (58 %) получен для компании Boeing (BA) после третьей итерации. Первые две итерации давали совсем негативный результат. Самый лучший результат (89 %) получен для компании American International Group Inc (AIG) с первой попытки обучения.

С практической точки зрения важной особенностью метода является то, что после обучения система начинает фильтровать «флаги» очень жестко. В результате предъявления таких строгих требований отбираются только лучшие прогнозы. Следствием этой стратегии является относительно редкая встречаемость флагов в тестовой выборке. Среднее количество выявленных программой флагов за год колеблется для тестовой выборки компаний в пределах от 0,3 (Caterpillar/CAT) до 3,8 (General Electric/GE). Такая редкая встречаемость также является следствием выбора категории данных для проведения тестирования, так как метод тестировался на среднесрочных данных (Daily).

ЗАКЛЮЧЕНИЕ

Предложенный метод машинного обучения на графических данных FuzGraph имеет ряд преимуществ:

- Низкая простота настройки и перенастройки. Т.е. при изменении входной информации перерасчет статистических данных и повторное включение в работу системы требует минимум затрат.
- Параметрический характер модели. Метод позволяет учитывать факторы, недоступные эксперту вследствие субъективности его оценки.
- Гибкость и вариативность. Используемый подход позволяет унифицировано работать с множеством разнообразных инструментов.
- Компактное описание характеристик. Увеличивает наглядность модели и позволяет контролировать промежуточные результаты компьютерного моделирования.

Предварительная апробация метода на фактических данных в режиме *in vitro* показала его эффективность. Очевидно, что метод FuzGraph может быть также использован и для других приложений, в которых необходимо распознавать типичные графические образы на больших выборках стохастических данных: при анализе радиосигналов, в задачах анализа речи и других акустических сигналов, в системах контроля и мониторинга состояния сложных промышленных объектов, в медицине.

ЛИТЕРАТУРА

1. Carbonell J. "Learning by Analogy" in Machine Learning: An Artificial Intelligence Approach. Michalski R., Carbonell J., and Mitchell T. (eds.), San Francisco: Morgan Kaufmann, 1983.

2. Holland J., *Adaptation in Natural and Artificial Systems*, Ann Arbor: The University of Michigan Press, 1975.
3. Koza J., *Genetic Programming: On the programming of Computers by Means of Natural Selection*. Cambridge. MA: MIT Press.1992.
4. Koza J., *Genetic Programming II: Automatic Discovery of Reusable programs*. Cambridge. MA: MIT Press.1994.
5. Kolodner J., *Case-Based Reasoning*, San Francisco: Morgan Kaufmann, 1993.
6. DeJong G., Mooney R., "Explanation-Based Learning: An Alternative View." *Machine Learning*. 1:145-176, 1986. Reprinted in Shavlik. J. and Dietterich. T. *Readings in Machine Learning*. San Francisco: Morgan Kaufmann, 1990. pp 452-467.
7. Samuel A. "Some Studies in Machine Learning Using the Game of Checkers." *IDM Journal of Research and Development*. 3:211-229, July 1959.
8. Элдер. А. Как играть и выигрывать на бирже. - М.: КРОН-ПРЕСС, 1996.
9. Поляков В., Шевченко А. Технология принятия решений на валютных и фондовых рынках с использованием нечеткозначной модели фигуры «флаг» Труды восьмой национальной конференции по искусственному интеллекту. КИИ-2002. 7-12 октября. Коломна Россия. 2002. т.1.с. 343-352.
10. Nils. J. Nilsson. "Introduction to Machine Learning". Stanford, CA 94305.
11. McCulloch W.S. and Pitts W.H. "A Logical Calculus of the Ideas Immanent in Nervous Activity". *Bulletin of Mathematical Biophysics*, Vol. 5. pp 115-133. Chicago: University of Chicago Press. 1943.
12. А.Н.Аверкин, И.З. Батыршин, А.Ф.Блишун, В.Б.Силов, В.Б.Тарасов *Нечеткие множества в моделях управления и искусственного интеллекта. /Под.ред. Д.А.Поспелова. -М.: Наука. Гл.ред.физ.-мат.лит., 1986. -312 с.*