

Методы анализа больших объемов слабоструктурированной информации¹

В.Б. Бритков, А.В. Булычев

Аннотация. Лавинообразный рост возможностей получения и обработки информации приводит к ее усложнению и гетерогенности. Традиционные подходы становятся малоэффективными. Перспективным направлением повышения эффективности анализа больших объемов слабоструктурированной информации является использование технологий интеллектуального анализа данных (ИАД) - добычи знаний (Data Mining). Показано преимущество метода анализа данных с помощью экономатематических моделей аналогий.

Ключевые слова: системный анализ, анализ слабоструктурированной информации, интеллектуальный анализ данных (Data Mining), экономатематические модели.

Введение

Современные информационные технологии позволяют собирать, хранить и обрабатывать огромные массивы информации. Существует большое количество методов анализа информации - методы математической статистики, кластеризации, классификации, в том числе «комитеты» алгоритмов, когда результаты, полученные с помощью нескольких алгоритмов, усредняются для построения конечной модели. Перечисленные методы с разной степенью точности могут описывать закономерности в данных. В современных условиях ввиду большого количества накапливаемой информации происходит усложнение ее структуры.

В этих условиях для увеличения эффективности анализа больших объемов информации возникает необходимость в новых нетрадиционных методах обработки информации (таких как ИАД) посредством автоматизированных познавательных процедур с использованием баз фактов и баз знаний, автоматического порождения гипотез, процедуры объяснения исходного состояния баз

фактов с целью оправдания и принятия гипотезы [1]. На более ранней стадии развития этих технологий предпринимались попытки использования такого подхода для интерпретации большого числа расчетных данных [2].

Одной из методологий, которая помогает решать задачи различных классов поиска закономерностей и интерпретации результатов, является методология интеллектуального анализа данных Data Mining. Она применяется для нахождения и исследования закономерностей в массивах слабоструктурированной информации и построения моделей, описывающих поведение сложных систем.

Data Mining – исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых данных знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком [3, 4].

Характерной особенностью анализа данных методами Data Mining является использование различных алгоритмов нахождения закономерностей в данных [5, 6]. Расширение набора мо-

¹ Работа выполнена при поддержке РФФИ, проект № 10-07-00352-а.

делей анализа данных различной алгоритмической природы может оказаться продуктивным в том классе задач, где недостаточно точно работают классические методы: статистические, аналитические или детерминированные. Каждый этап исследования данных мы можем строить на конечном числе гипотез, которые могут подтверждаться или не подтверждаться впоследствии. Чем больше построенных моделей и описаний близки к гипотезам, тем более мы вправе предполагать достоверность результата. Безусловно, при исследовании реальных данных какие-либо выводы можно делать с конечной степенью точности. Анализ данных различными алгоритмами достаточно подробно отражен в работах [6, 7].

В настоящее время исследования в области интеллектуального анализа данных Data Mining сталкиваются с определенными трудностями. Несмотря на то, что в арсенале Data Mining существует несколько десятков различных алгоритмов анализа данных, тем не менее формально-логический вывод из полученных законов не всегда может нести в себе принципиально новые знания о системе. Так, из уравнений движений макрообъектов классической физики нельзя получить законы квантовой механики. Таким образом, существующий подход интеллектуального анализа Data Mining нуждается в дальнейшем изучении и дополнении новыми методиками анализа данных, т.к. существующих методов исследования данных не всегда достаточно для понимания новых важных закономерностей. Например, в исследовании теории катастроф и безопасности проблемы выявленные на ранней стадии, могут сэкономить большие средства, которые не придется тратить на исправление последствий катастроф. Итак, поскольку мы хотим найти в данных новые закономерности, которые не выводятся из результатов стандартных методов анализа, то часто лучше воспользоваться новыми, нестандартными подходами к анализу данных и вслед за этим эффективными вычислительными процедурами [8]. Это связано с тем, что найти аналитические решения у полученных уравнений не всегда возможно.

Целью интеллектуального анализа данных Data Mining является поиск законов поведения

и функционирования исследуемой системы, которые описываются числовыми данными. Для качественного анализа данных нужно понимать их физическую природу, для чего привлекаются специалисты из предметной области. Практическая полезность полученных результатов, согласно определению Data Mining, заключается в трактовке полученных закономерностей этими специалистами. Ниже мы сосредоточимся непосредственно на анализе табличных данных и попытаемся найти закономерности среди данных и более глубоко понять законы поведения системы, ее «похожесть» на уже хорошо изученные системы. На основании этого можно глубже понять связи внутри системы, производить ее анализ, синтез новых систем и моделировать динамику развития системы.

Развитие технологий Data Mining характеризуется расширением классов методов, применяемых к анализу данных в плохоформализуемых системах. Это совпадает с общими тенденциями в науке. В качестве примеров можно привести использование академиком В.П.Масловым «Квантовой теории в экономике» [9] и большие продвижения в области «Экономифизики» [10].

1. Проблемы методологии интеллектуального анализа данных

Экономические, физические и другие системы живут по определенным законам. Вместе с тем, можно и не найти эти законы, если применять стандартные методы и алгоритмы анализа данных в силу естественной ограниченности возможности последних. Описательные и предсказательные модели с определенной конечной степенью точности могут описывать зависимости внутри данных и предсказывать будущие значения целевых переменных. Однако остается открытым вопрос об основном законе динамики развития большинства систем (для каждой системы он может быть свой). То есть, для большинства практически важных систем остается переменным вопрос об уравнениях, описывающих поведение систем. В статье предлагается метод описания поведения систем, который в некоторых случаях позволяет анализировать закономерности с большей точностью.

При анализе данных с помощью методологии Data Mining мы можем строить описательные и предсказательные модели на основе нескольких алгоритмов. Полученные результаты могут продвинуть нас в нахождении скрытых закономерностей в данных. В то же время часто после этого возникает ситуация, когда будущие значения, вычисленные по одной модели, отличаются от тех же значений, найденных по другой модели. Это означает, что по крайней мере одна из моделей дает лишь приближенные значения целевой переменной.

Еще одна проблема заключается в том, что алгоритмы Data Mining часто не принимают во внимание точки, которые являются по определенным признакам т.н. «выколотыми», или статистически не значимыми и тем самым не влияющими на результат анализа данных. Точкой мы обозначаем какое-либо состояние системы, описываемой в n -мерном пространстве измеримых и *наблюдаемых* параметров системы. Если ничего нельзя сказать о статистической устойчивости данных, то нельзя и однозначно исключать «выколотые» точки из рассмотрения. Если же точки не являются «выколотыми», то они должны вносить вклад в результат анализа наряду с другими точками. Например, в результате кластерного анализа данных в зависимости от допущения, что та или иная точка является выколотой, данные разбиваются на кластеры разными способами, что вносит изменения в результат анализа.

Проблемой является и то, что новые данные, описывающие систему и динамически поступающие на вход для исследования, часто вносят существенный вклад в результат, т.е. не всегда подтверждают уже построенные модели. На языке квантовой физики это означает, что если мы пронаблюдали какое-либо состояние системы, сам этот факт наблюдения внес изменение в волновую функцию состояния системы. И наоборот - если мы не пронаблюдали какое-либо состояние системы в результате сбора данных, это также вносит изменение в результат анализа данных.

Еще одна из системных проблем состоит в том, что при нахождении закономерностей среди данных и построении моделей мы не можем сказать, насколько точно наша модель описывает систему на всем интервале ее развития и

можем ли мы пользоваться построенными предсказательными моделями на любом участке времени.

2. Постановка задачи

Ставится задача построить предсказательную модель (в качестве прогноза выступает математическое ожидание случайных величин – параметров системы), используя предположение об определенном виде плотности вероятности случайных величин. Сравнить точность метода с точностью стандартных методов прогнозирования (регрессии, нейронных сетей, генетических алгоритмов).

Плотность $\rho = |\varphi(\vec{r})|^2$ определяется квадратом модуля решения уравнения:

$$\alpha \Delta \varphi(\vec{r}) + \beta(\vec{r})\varphi(\vec{r}) = \gamma\varphi(\vec{r}), \quad \alpha, \gamma - const,$$

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \dots + \frac{\partial^2}{\partial x_n^2}, \quad \varphi(\vec{r}) = \varphi(x_1, x_2, \dots, x_n) \quad (1),$$

x_1, x_2, \dots, x_n - величины, характеризующие систему, t - время.

В терминах предметной области переменные x_1, x_2, \dots, x_n означают производственно-экономические показатели системы. Для численного решения уравнения было выбрано 8 показателей, описывающих систему, решалась задача прогнозирования прибыли X за следующие периоды:

- x_1 - прибыль за предыдущий отчетный период,
- x_2 - объем неотгруженных товаров, на которые имеется заказ клиентов,
- x_3 - скорость производства продукции,
- x_4 - скорость транспортировки от завода до места складирования,
- x_5 - скорость таможенного оформления,
- x_6 - скорость отгрузки продукции клиентам,
- x_7 - объем ежедневных расходов на рекламу,
- x_8 - цены на товары.

плотность вероятности описывающих рассматриваемую систему величин, которая определяется с помощью решения уравнения (1), позволяет с большей точностью проводить анализ данных и строить предсказательные модели поведения системы.

Для численного решения уравнений был использован аппарат вычислительной математики и разработан программный продукт, реализующий алгоритмы численного решения и позволяющий настраивать их параметры.

3. Описание уравнения

Уравнение (1) имеет вид, аналогичный уравнению, которое в квантовой физике описывает пространственно-временное распределение с помощью представления волновой функции (уравнение Шредингера). В квантовой физике вводится представление о вероятностном описании поведения частицы путем задания волновой функции, характеризующей вероятность местонахождения частицы в той или иной области пространства. Затем с помощью гамильтониана микрообъекта выводится уравнение для этой функции, описывающей движение квантовой частицы. В этом состоит отличие от классического описания движения объектов. Пусть волновая функция $\psi(\vec{r}, t)$ задана в n -мерном пространстве. Набор координат, которые выступают в роли аргументов функции, представляет собой полный набор физических величин, которые можно измерить в системе. Тогда уравнение Шредингера записывается в виде

$$-\frac{\hbar^2}{2m} \Delta \psi(\vec{r}, t) + E_p(\vec{r}, t) \psi(\vec{r}, t) = i\hbar \frac{\partial}{\partial t} \psi(\vec{r}, t).$$

В последнем выражении \hbar - постоянная Планка; m - масса частицы, $E_p(\vec{r}, t)$ - внешняя по отношению к частице потенциальная энергия в точке $\vec{r}(x_1, x_2, \dots, x_n)$,

$\Delta = \nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \dots + \frac{\partial^2}{\partial x_n^2}$ - оператор Лапласа.

Частное решение для специального случая, когда $E_p(\vec{r})$ не является функцией времени, можно записать в виде

$$\psi(\vec{r}, t) = \varphi(\vec{r}) e^{(-i/\hbar)Et},$$

где функция $\varphi(\vec{r})$ должна удовлетворять стационарному уравнению Шредингера

$$-\frac{\hbar^2}{2m} \Delta \varphi(\vec{r}) + E_p(\vec{r}) \varphi(\vec{r}) = E \varphi(\vec{r}).$$

Последнее уравнение может быть решено аналитически для небольшого числа частных

типов функции $E_p(\vec{r})$. E является полной энергией частицы при движении, описываемом функцией $\psi(\vec{r}, t)$. Если система может пребывать в состояниях, описываемых функциями ψ_1 и ψ_2 , то она может пребывать и в состоянии, описываемом функцией $\psi_\Sigma = c_1 \psi_1 + c_2 \psi_2$ при любых комплексных c_1 и c_2 .

Квадрат модуля функции $|\psi(\vec{r}, t)|^2$ представляет собой плотность вероятности ω обнаружить систему в положении, описываемом значениями $\vec{r} = \vec{r}_0$ в момент времени t .

В заданном состоянии системы, описываемом функцией $\psi(\vec{r}, t)$, вероятность P того, что система будет обнаружена в области конечного объема V пространства состояний, равна

$$P = \int_V dP = \int_V \omega dV = \int_V \psi^* \psi dV.$$

Вероятностный смысл волновой функции накладывает определенные ограничения на ее вид:

- условие конечности - это условие требует, чтобы волновая функция была квадратично интегрируемой;

- условие однозначности; функция должна быть однозначной функцией состояний и времени, так как плотность вероятности обнаружения в заданном состоянии должна определяться в каждой задаче однозначно;

- условие непрерывности - в любой момент времени функция должна быть непрерывной функцией пространственных координат; кроме того, непрерывными должны быть также частные производные функции $\frac{\partial \psi}{\partial x_1}, \frac{\partial \psi}{\partial x_2}, \dots, \frac{\partial \psi}{\partial x_n}$.

Найдем решение стационарного уравнения Шредингера в одномерном случае:

$$-\frac{\hbar^2}{2m} \frac{d^2 \varphi(x)}{dx^2} + U(x) \varphi(x) = E \varphi(x). \quad (2)$$

Для постановки задачи о нахождении решения (2) задаются граничные условия для интервала $[a, b]$:

- $\alpha_1 \varphi(a) + \beta_1 \frac{d\varphi(a)}{dx} = \gamma_1$
 - $\alpha_2 \varphi(b) + \beta_2 \frac{d\varphi(b)}{dx} = \gamma_2,$
- $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2 - const.$

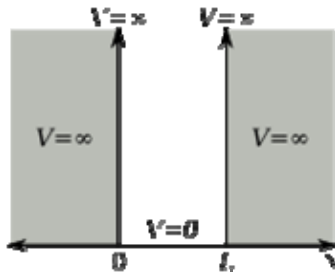


Рис.1.

Рассмотрим случай потенциальной энергии, как показано на Рис.1:

$$-\frac{\hbar^2}{2m} \frac{d^2 \varphi(x)}{dx^2} = E \varphi(x).$$

Решения этого уравнения

$$\varphi_n(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right), \quad E_n = \frac{n^2 \hbar^2 \pi^2}{2mL^2},$$

n - натуральное число.

На Рис. 2 показан вид решений для различных значений n .

Сложный потенциал в уравнении (2) затрудняет поиск решения в пространстве элементарных функций, поэтому необходимо привлекать численные методы. Например, метод конечных разностей на выбранной сетке (здесь h - шаг дискретизации, n - номер узла сетки, U_n - значение потенциальной энергии $U(x)$ на узлах сетки):

$$\begin{aligned} \frac{d^2 y(x)}{dx^2} &= \frac{y_{n-1} - 2y_n + y_{n+1}}{h^2}, \\ -\frac{\hbar^2}{2m} \frac{y_{n-1} - 2y_n + y_{n+1}}{h^2} + U_n y_n &= E y_n, \\ -y_{n-1} + \left(2 + h^2 \frac{2mU_n}{\hbar^2}\right) y_n - y_{n+1} &= h^2 \frac{2mE}{\hbar^2} y_n. \end{aligned}$$

Обозначая $\mathcal{G}_n = h^2 \frac{2mU_n}{\hbar^2}$ и $e = h^2 \frac{2mE}{\hbar^2}$, получаем $-y_{n-1} + (2 + \mathcal{G}_n - e)y_n - y_{n+1} = 0$.

Под последним выражением понимается система уравнений для всех возможных n .

Для нахождения коэффициентов α, β, γ в уравнении (1) используется метод конечных разностей на множестве исторических данных. Вторым шагом оценивается точность построенной модели на нескольких тестовых выборках. В случае удовлетворительного результата модель используется для прогнозирования величин. В

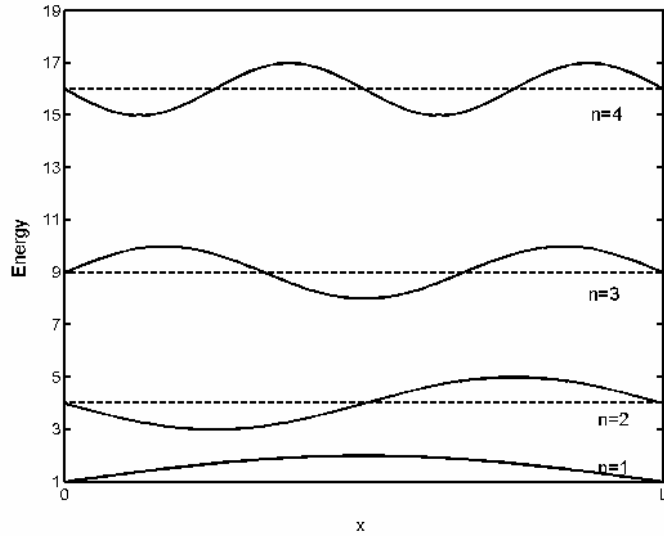


Рис.2.

качестве прогноза выступает численная оценка интеграла (интегрирование ведется по всему возможному множеству значений переменных (x_1, x_2, \dots, x_n)): $M[X] = \int \dots \int X(x_1, x_2, \dots, x_n) |\varphi(x_1, x_2, \dots, x_n)|^2 dx_1 dx_2 \dots dx_n$.

4. Описание входных данных

Мы рассматриваем класс входных данных, традиционно представленных в виде прямоугольной таблицы, в которой значения в столбцах представляют собой динамику развития какого-либо параметра системы. Каждая строка представляет собой значения всех параметров системы в определенный момент времени. Для анализа данных будем применять алгоритмы и методы интеллектуального анализа данных Data Mining.

Пусть имеется прямоугольная таблица данных, описывающая работу производственного предприятия. В первоначальной таблице 20 столбцов (полей) и порядка 100000 записей. В каждом поле представлена временная определенное производственно-экономического показателя предприятия в числовом виде. Каждая запись, в свою очередь, соответствует значениям показателей в определенный момент времени. Ниже приведен фрагмент таблицы (выбраны 7 столбцов). В первом столбце таблицы помещаются числовые названия артикулов продуктов из товарной номенклатуры компании. Во втором столбце находятся данные, которые описывают

Артикул продукта	Склад	Изменение статуса	Финансовый документ	Количество	Прибыль от операции, руб.	Дата документа
0512040	Фм-Чехов	Выход из обработки	TRF73248	396	19198,08	25.01.2008
1400310	Импорт	Вход на обработку	140642	20	412	23.01.2008
2025430	Импорт	Вход на обработку	00056F	12	2643,6	22.01.2008
2025430	Клиент	Отгрузка клиентам	RUB069489	12	5287,2	22.01.2008
2045200	Фм-Чехов	Выход из обработки	TRF69066	1296	98816,96	22.01.2008
2045220	Фм-Чехов	Выход из обработки	TRF72988	288	216800,2	25.01.2008
2045244	Фм-Чехов	Выход из обработки	TRF72774	96	86400,6	24.01.2008
2058860	Фм-Чехов	Выход из обработки	TRF72455	904	128624,4	24.01.2008
2058870	Фм-Чехов	Выход из обработки	TRF69749	1430	175000,8	22.01.2008
2058880	Фм-Чехов	Выход из обработки	TRF67746	480	49104,4	21.01.2008
2058880	Клиент	Отгрузка клиентам	RUB067747	480	73656,6	21.01.2008
2076020	Импорт	Вход на обработку	00057F	6	129,76	22.01.2008
2089850	Фм-Чехов	Выход из обработки	TRF71297	936	106141,2	24.01.2008
2089890	Фм-Чехов	Выход из обработки	TRF71297/1	48	36057,6	24.01.2008
2102352	Фм-Чехов	Выход из обработки	TRF68593	1584	116424,6	22.01.2008
2102352	Клиент	Отгрузка клиентам 1	RUB069476	1326	146192,25	22.01.2008
2115492	Фм-Чехов	Выход из обработки	TRF71563	1	1126,4	24.01.2008

место, где вычисляются показатели прибыли от операции согласно процедуре управленческого учета (ФМ-Чехов – склад предпродажной подготовки, импорт – склад таможи, клиент – склад клиента, т.е. конечный пункт, куда отгружаются продукты). Столбец (поле) «Статус перехода» – числовое обозначение смены статуса продукта, например, отгрузка со склада предпродажной подготовки ФМ-Чехов на склад клиента. Столбец «Финансовый документ» согласно процедуре управленческого учета обозначает номер внутреннего финансового документа компании, обосновывающего смену статуса продукта. Остальные 3 столбца описывают количество, прибыль и дату получения прибыли для соответствующего артикула.

Значения полей в моменты измерения определены однозначно с доступной нам точностью. Вместе с тем, в этот же момент времени эти показатели могли бы иметь другие значения, зависящие от деятельности и решений руководства компании в прошлом, от случайных факторов внутри компании, от текущей ситуации на рынке. Таким образом, по аналогии с вероятностной природой в описании физических объектов (микрообъектов) мы можем присвоить вероятности полученным состояниям системы, которые выражаются набором производственно-экономических показателей, т.е. мы строим модель, где состояние нашей экономической системы имеет вероятностную природу.

В результате проведенных наблюдений за 2007-2009 г.г. было обнаружено, что в феврале, марте, апреле, ноябре и декабре вероятность получения прибыли из диапазона вблизи указанных на графике по оси абсцисс Р сумм в млн. руб. имеет зависимость, показанную на Рис. 3:

На графике видны флуктуации (шум), вносимые случайными факторами, влияющими на прибыль. Далее, в январе, июле, августе, ситуация изменилась и соответствующий график выглядит так как это показано на Рис. 4:

В мае, июне, сентябре и октябре график имеет иной вид как на Рис.5.

5. Анализ полученных закономерностей

Система находится в разных состояниях в каждой из 3-х вышеперечисленных групп месяцев. Данные за несколько лет подтверждают, что, несмотря на то, что прибыль росла каждый год, она принимала значения вблизи максимумов функций (с поправкой на среднегодовой рост), изображенных на рисунках. Возможно, что существуют другие локальные экстремумы в пределах указанных графиков. Были обнаружены экстремумы и у других показателей деятельности фирмы. В нашем примере в качестве целевой переменной мы рассматриваем прибыль компании.



Рис.3.

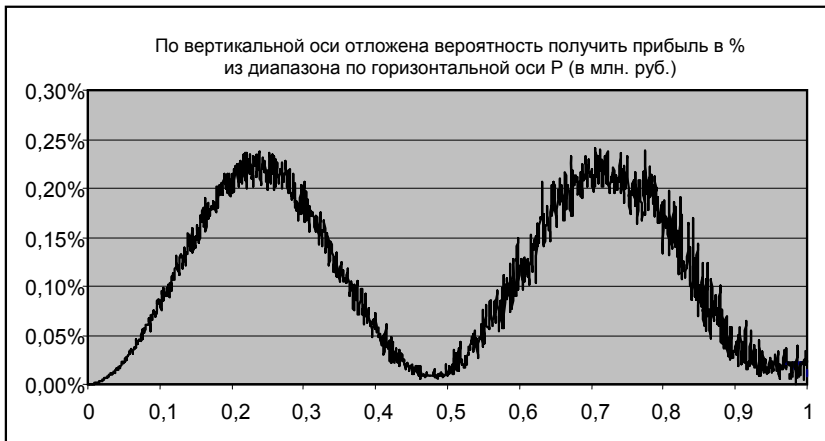


Рис.4.

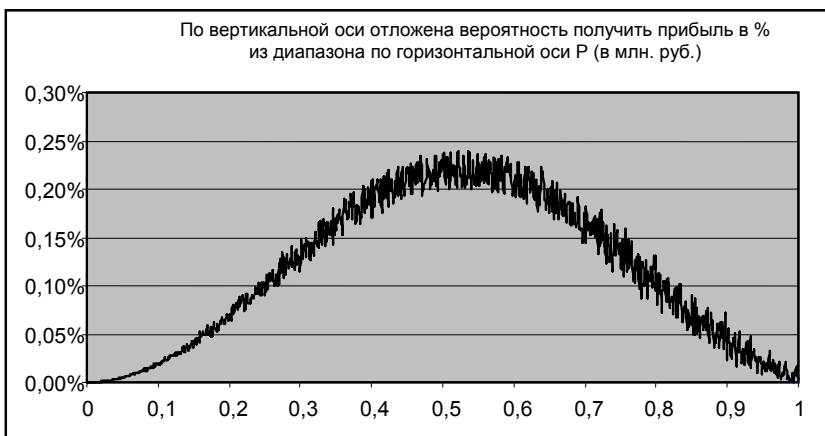


Рис.5.

Вид графиков показывает, что на первом этапе анализа (позже мы это исследуем подробнее) можно найти «похожесть» по определенным параметрам между состоянием экономической системы и состоянием, описываемым

волновой функцией квантовой нерелятивистской частицы в потенциальном поле [13].

Во-первых, функция вероятности прибыли имеет вид похожий на функцию квадрата модуля волновой функции, являющейся решением

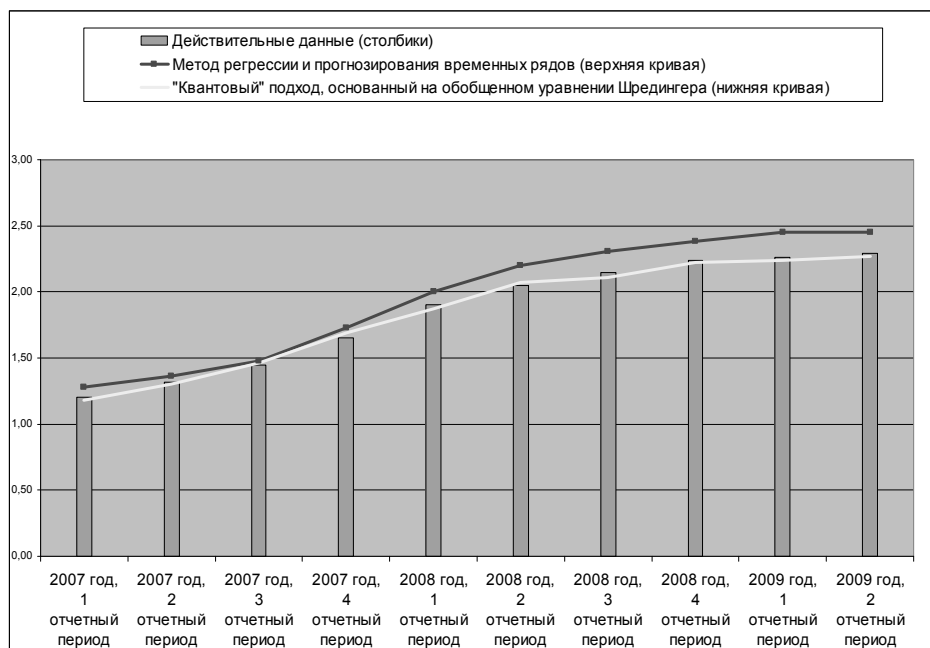


Рис. 6.

частного стационарного вида уравнения Шредингера [14] вид (нормированной по всему пространству на 1).

Во-вторых, это описание одной и той же системы, которая в разные периоды времени находилась в разных состояниях (на разные состояния указывает то, что снова выполняется условие нормирования вероятности на 1 и что вид функции вероятности изменился). Полученные аналогии в первом приближении можно объяснить так. На уровне макроэкономики страны фирма похожа на микрообъект. Фирма действует в определенном поле экономических и юридических законов, имеет потенциал и «вес» на рынке.

6. Сравнение со стандартными подходами Data Mining

Описанный подход является новым в методах Data Mining. Он используется в совокупности со стандартными алгоритмами:

- алгоритмы поиска ассоциативных правил;
- алгоритмы кластеризация (методом агломерации, деления);
- алгоритмы задач классификации и регрессии (в т.ч. прогнозирование временных рядов);
- алгоритмы построения нейросетей и генетические алгоритмы.

В задаче регрессии при известной плотности вероятности, увеличивается точность прогноза, а алгоритмы кластеризации дают более точное разбиение на группы схожих элементов. Получена оценка, что в нашей задаче использование описанного подхода помогает более точно (на 15-20%) по сравнению со стандартными методами (построения нейросетей, генетическими алгоритмами, регрессии) планировать прибыль и бюджет компаний на будущий период (Рис. 6, по оси ординат - млн. руб.). Значение плотности вероятности в пространстве измеримых параметров для каждой исследуемой системы описывается разными законами. Данный подход рассматривается как дополнительный к существующим методам и позволяющий строить более точные прогнозы.

Заключение

В статье продемонстрирован новый подход к интеллектуальному анализу данных. Показано, что в качестве уравнения, описывающего состояние и динамику развития системы, можно воспользоваться аналогом уравнения Шредингера с обобщенными параметрами, применимыми к исследуемой системе. В результате применения метода удастся с большей точностью предсказывать экономические показатели

деятельности компании. При анализе данных принимаются во внимание все наблюдаемые состояния системы (т.е. отсутствуют «выколотые» точки, которые в случае исследования статистическими методами могли быть проигнорированы). В разложении обобщенной функции состояния системы присутствуют периодические функции. Это означает, что при анализе мы можем учитывать цикличность поведения системы. Учитываются наблюдаемые состояния системы, и в случае необходимости закон поведения системы может быть скорректирован с учетом ее текущего состояния.

Границы применимости описанного подхода зависят от задачи. Вывод о применимости можно сделать, сопоставив расчетные показатели и реальные показатели деятельности.

Литература

1. Ю.М. Арский, В.К. Финн. Принципы конструирования интеллектуальных систем. Информационные технологии и вычислительные системы. № 4. М.: 2008. стр. 4-37.
2. Геловани В.А., Болоткин С.И., Бритков В.Б. и др. Программное обеспечение системы моделирования. Препринт – ВНИИСИ, М., 1978г. 26 с.
3. Бритков В.Б., Маленкова Н.А. Интеллектуальные методы в решении задач использования океанологической информации. // Информационные технологии Государственного фонда данных о состоянии окружающей природной среды Труды ВНИИГМИ-МЦД. Вып.170. Санкт-Петербург: Гидрометеоздат – Обнинск: ВНИИГМИ-МЦД, 2002. сс. 252-277.
4. Berson A, Smith S. J. Data Warehousing, Data Mining & OLAP. McGrawHill, 1997.
5. Барсебян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. БХВ-Петербург, 2008.
6. Булычев А.В., Бритков В.Б. Технологии интеллектуального анализа Data Mining и их использование при решении задач логистической оптимизации. Труды 51-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук»: Часть VII. Управление и прикладная математика. Том 3. — М.: МФТИ, 2008. — 138 с. ISBN 978-5-7417-0271-0
7. Бритков В.Б., Булычев А.В. Интеллектуальный анализ социально-экономической информации регионов России. Труды II международной научной конференции ИСА РАН «Системный анализ и информационные технологии». Том II. Обнинск, 10-14 сентября, 2007г., с 100-104.
8. Калиткин. Н. Н. Численные методы. М., Наука, 1978.
9. В.П. Маслов. Квантовая экономика, Наука, М., 2005
10. Материалы первого всероссийского конгресса по экономифизике «Экономифизика, финансовые рынки, экономический рост». М. Изд-во Финансовой академии при Правительстве Российской Федерации, 2009.
11. Булычев А.В. Использование вероятностного моделирования и уравнений квантовой физики в интеллектуальном анализе данных – Data Mining. Третья Международная конференция «Системный анализ и информационные технологии» САИТ – 2009 (14-18 сентября 2009 г., Звенигород, Россия): Труды конференции. М., 2009. с. 270-277.
12. Боум А. Квантовая механика: основы и приложения. М. Мир, 1990. – 720с. ISBN 5-03-001311-3
13. Тарасов Л.В. Основы квантовой механики, Москва, издательство «Высшая школа», 1978
14. Березин Ф. А., Шубин М. А. Уравнение Шредингера. Изд-во МГУ, 1983.

Владимир Борисович Бритков. Заведующий лабораторией Института системного анализа Российской академии наук. Окончил Московский физико-технический институт (Государственный университет) в 1968 году. Кандидат физико-математических наук, доцент. Автор более 200 печатных работ, 4-х монографий. Область научных интересов: интеллектуальный анализ данных, интеллектуальные информационные системы; моделирование и анализ слабоформализуемых систем, социальные системы и модели стратегического равновесия; информационная поддержка принятия решений при чрезвычайных ситуациях.

Александр Викторович Булычев. Аспирант Института системного анализа Российской академии наук. Окончил Московский физико-технический институт (Государственный университет) в 2006 году. Автор 6 печатных работ. Область научных интересов: интеллектуальный анализ данных Data Mining; экономифизические модели; вероятностное моделирование производственных процессов; оптимизация производственных операций. E-mail: alexandresun@rambler.ru.