5. Colombo L. Tight-binding theory of native point defects in silicon // Annu. Rev. Mater. Res. – 2002. – V. 32. – P. 271 – 295.
6. Aziz M.J. Pressure and Stress Effects on Diffusion in Si // Defect and Diffusion Forum – 1998. – V.153 – 155. – P. 1 – 10.

АЛГОРИТМ ПЕРЕСЧЕТА СЕТЕЙ БАЙЕСА

А. А. Быков, А. А. Волосевич
Беларусь, г. Минск

В процессе работы интернет-магазина образуется большой массив информации в виде базы данных, содержащей статистику продаж. В таком виде эта информация недоступна для анализа человеком и неудобна для обработки программными средствами. Сети Байеса – это специализированные структуры, способные отражать в графической форме закономерности спроса и его характер посредством элементов теории вероятности [1, 2, 3].

В задаче интернет-торговли сеть отражает закономерности между характеристиками покупателей и выбранных ими товаров на основе статистики продаж. Особенностью данной задачи является следующее: как правило, поисковой системе интернет-магазина предоставляются частично заполненные формы.

В работе предлагается использовать сети Байеса для восстановления пробелов в данных поисковых форм, что существенно повышает эффективность поисковой системы, а также описываются алгоритмы пересчета сети Байеса.

Описание закономерностей спроса в сети Байеса производится посредством задания связей между характеристиками товаров и характеристиками покупателей. Характеристика – это величина, которая принимает одно из возможных дискретных значений и описывает отдельное свойство покупателя или товара. С каждой из характеристик связана таблица распределения вероятностей её значений. Таким образом, описывается некоторый усредненный товар или покупатель [2].

Связью в сети Байеса называется упорядоченная пара характеристик, при этом первая характеристика называется главной, а другая зависимой, а также таблица условных вероятностей связи. Таблица условных вероятностей – матрица, связанная с парой характеристик связи. Она имеет количество строк, равное количеству значений главной характеристики, и количество столбцов, равное количеству значений зависимой характеристики. При этом каждая строка матрицы является новым распределением зависимой характеристики, если главная характеристика принимает соответствующее значение. Поэтому сумма значений каждой строки матрицы равна единице.

Байесовские сети — это направленный ациклический граф, каждая вершина которого является характеристикой, а каждая дуга между вершинами является соответствующей связью между характеристиками. Сеть обозначается как G = (V, E), где V множество характеристик сети, E – множество связей сети [1].

Пересчет сети Байеса – итерационный процесс задания новых распределений зависимым характеристикам в соответствии со значениями главных харак-

теристик. При этом значения вероятностей некоторых характеристик сети, которые известны достоверно, остаются без изменения.

Можно сказать, что главная задача при использовании сетей Байеса – это получение новых распределений значений в незаданных характеристиках на основе неполных и иногда неточных данных в виде значений заданных характеристик сети [5].

Рассмотрим сеть изображенную на рис. 1. Допустим, нам стало известно, что покупатель 27 лет ищет телефон марки Motorolla. Требуется оценить ценовую категорию телефонов, которые стоит предложить покупателю.

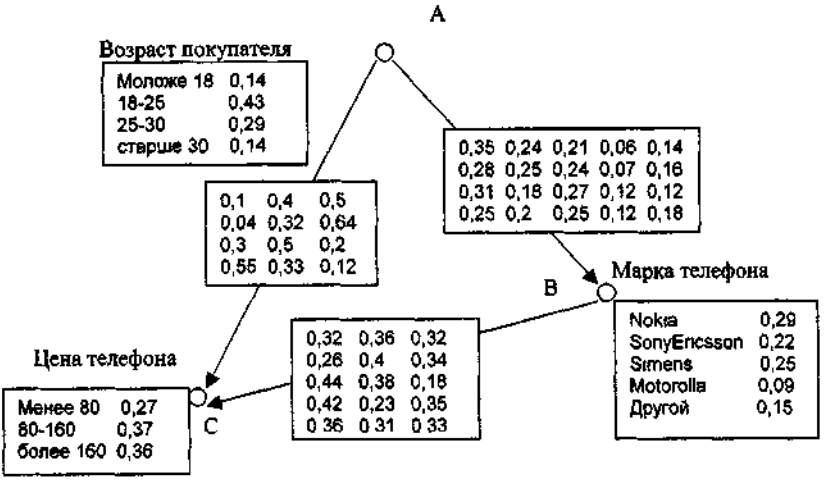

Рис. 1. Пример сети Байеса

В табл. 1 представлена матрица для связи (A, C). Из значений таблицы условных вероятностей следует, что если характеристика A («Возраст покупателя») принимает значение «18 – 25 лет», т.е. распределение её значений принимает состояние {0, 1, 0, 0}, то распределение значений характеристики C («Цена телефона») принимает состояние {0.04, 0.32, 0.64}.

Таблица 1. Условные вероятности связи (A, C).

$P(C \mid A)$	Менее 80	80-160	Более 160
Моложе 18 лет	0,1	0,4	0,5
18-25 лет	0,04	0,32	0,64
25-30 лет	0,3	0,5	0,2
Старше 30 лет	0,55	0,33	0,12

Если установлены значения характеристик узлов A и B, например $P(a_3)=1$, $P(b_4)=1$, то распределения значений характеристики C равно $P'_A(C)=(0.3,0.5,0.2)$ и $P'_B(C)=(0.42,0.23,0.35)$.

Окончательно распределения значений в узле C предполагается рассчитывать как среднее арифметическое:

$$P'(C) = (\frac{0.3+0.42}{2}, \frac{0.5+0.23}{2}, \frac{0.2+0.35}{2}) = (0.36;0.37;0.28)$$

В данном примере произведена одна итерация алгоритма пересчета. В реальных сетях итераций может быть значительно больше.

Алгоритм пересчета должен произвести пересчет всех незаданных характеристик. Часто для решения данной задачи необходим пересчет связей сети в обратном направлении. Для этих целей алгоритм пересчета сетей использует формулу Байеса:

$$P(b|a) = P(a|b) * \frac{P(b)}{P(a)}$$

Особенностью сети Байеса является то, что она состоит из факторов проблемной области (характеристик и связей), т. о. сеть используется как модель. Такая универсальность сетей позволяет получать решения при любых комбинациях известных факторов, а также порождает три тесно связанные проблемы:

- Уменьшение точности пересчета.
- Увеличение количества связей.
- Увеличение количества итераций пересчета.

В первых работах по сетям Байеса стала очевидна проблема точного вывода – составление формул, по которым можно было бы сразу вычислить распределения значений в незаданных характеристиках [2, 9]. Итерационный алгоритм пересчета, который используется сейчас для пересчета сетей Байеса, существенно упростил их использование, однако поставил вопрос о циклах в структуре сети. Поэтому на сегодняшний день сети Байеса – это ациклический граф [1, 2, 3].

Для решения проблемы зацикливания в данной работе предлагается механизм затухания пересчета. Суть его в следующем. Отказ от дальнейшего пересчета происходит потому, что значение, полученное на предыдущей итерации, обладает слишком большой ошибкой, или потому, что изменение значений, полученных на предыдущей итерации, оказалось незначительным. Ошибка, получаемая при пересчете характеристик в сети Байеса, обусловлена тем, что:

- При построении сети Байеса, как и любой другой экспертной системы приходится ограничивать количество факторов, влияющих на решение.
- Рассматривать взаимное влияние факторов независимо. Так, как в предыдущем примере мы рассматривали влияние характеристик «Возраст покупателя» и «Марка телефона» независимо.

Также в сети необходимо устранить противоречивость связей друг другу. В данной работе предлагается добавить в алгоритм построения сети Байеса алгоритмы тестирования совместной работы связей сети. Для изменения параметров связи можно использовать EM-алгоритм [6] и алгоритм обратного распределения ошибок.

EM-алгоритм используется в задачах восстановления пропусков в данных, обработки изображений, распознавании речи. Алгоритм обратного распределе-

ния ошибок используется для обучения нейронных сетей. В работе предлагается адаптация связки этих алгоритмов для тестирования и корректирования сетей Байеса. Эталонными значениями для работы алгоритма являются совместные вероятности событий. Данные величины можно получить на основе статистики и, из-за их большого количества, хранить в базе данных.

ЕМ-алгоритм основан на использовании функции максимального правдоподобия и состоит из двух шагов Е-шаг и М-шаг. На Е-шаге вычисляется ожидаемое значение некоторой характеристики на основе существующих связей. На М-шаге находится разность полученных значений от эталонных и решается задача поправки коэффициентов в участвовавших в пересчете связях.

Основная идея алгоритма обратного распространения ошибки состоит в анализе сигналов ошибки в направлении обратном прямому распространению сигналов и корректировке коэффициентов сети. Для каждого коэффициента сохраняется процент ошибок при тестировании связей. Этот процент ошибки используется в алгоритме пересчета как коэффициент затухания. Природа появления этого коэффициента заключается в ограничении количества факторов, влияющих на решение, и рассмотрении взаимного влияние факторов независимо. Хотя использование коэффициента затухания устраняет зацикливание алгоритма пересчета, это не увеличивает точность пересчета.

Основным противоречием сетей Байеса является конфликт между количеством связей сети и точностью получаемого результата. Интерес заключается в том, что не только увеличение количества связей действительно приводит к увеличению точности, но и более грамотное сочетание этих связей позволяет добиться аналогичного результата. В свою очередь, увеличение количества или объединение связей (от многих характеристик к единой, содержащей все возможные комбинации условных вероятностей объединенных характеристик, и учитывающей влияние факторов на решение совместно) однозначно замедляет пересчет.

С целью повышения точности сетей Байеса и уменьшения трудоемкости в работе предлагается ввести соотношение точность/трудоемкость для целевой сети, производить объединение характеристик или связей согласно CI-тесту предложенному в [1], а также производить d-разделение характеристик.

Основа CI-теста в том, что она рассматривает отношение истинного значения условной вероятности, к вероятности, полученной при помощи перемножения вероятностей, т. е. считая факторы независимыми. Данная мера позволяет оценить, насколько независимы характеристики X_i и X_j:

$$CI(X_i, X_j \mid X_{i+1},\ldots,X_{j-1}) =$$

$$= \sum_{x_i,\ldots,x_j} P(x_i,\ldots,x_j)\log\frac{P(x_i,x_j \mid x_{i+1},\ldots,x_{j-1})}{P(x_i \mid x_{i+1},\ldots,x_{j-1})P(x_j \mid x_{i+1},\ldots,x_{j-1})}$$

Для направленного ациклического графа G = (V,E), и характеристик X, Y, и $C \in V \setminus \{X,Y\}$, говорят, что X и Y d-разделены относительно C в G тогда и только тогда, когда не существует путь смежности P между X и Y. Если X и Y

не d-разделены относительно C, говорят, что X и Y d-связаны относительно C. Определение d-разделения двух узлов может легко быть распространено на d-разделение двух непересекающихся множеств узлов составляющих путь в графе [7, 8].

То, что для разделения, объединения и добавления связей используется одна и та же мера позволяет сравнивать пользу от этих операций друг с другом и, соответственно трудоемкость при пересчете.

Достоинством описанного алгоритма тестирования работы сети является его гибкость. Алгоритм допускает исключение некоторого множества путей пересчета из списка с целью повышения точности других путей пересчета.

Литература

1. Jie Cheng. David Bell, Weiru Liu, Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory UCLA, 1998
2. Witten, I. H. (Ian H.). —Data Mining. Practical Machine Learning Tools and Techniques
3. Han J., Kamber M. Data Mining: Concepts and Techniques. — Morgan Kaufmann Publishers, 2000.
4. С. Хабаров. - Конспект лекций. 2003.
5. Hand D. J., Mannila H., Smyth P. Principles of Data Mining. — MIT Press, 2000.
6. Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // J. of the Royal Statistical Society, Series B.
7. Learning Bayesian Network Structure from Distributed Data, R. Chen K. Sivakumar, 2002
8. Pearl, J., Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann, 1988
9. Искусственный интеллект, современный подход. Стюарт Рассел, Питер Норвиг. М. 2006

РЕАЛИЗАЦИЯ РОЛЕЙ В ЯЗЫКАХ ПРОГРАММИРОВАНИЯ СО СТАТИЧЕСКОЙ ТИПИЗАЦИЕЙ

А. П. Побегайло
Беларусь, г. Минск

Введение

Прогресс в развитии отрасли программного обеспечения привел к разработке все более совершенных и технически сложных программных систем. Реализация таких систем требует глубокого понимания и точной формализации процессов, которые автоматизирует или которыми управляет программная система. Очевидно, что разработка таких систем невозможна без их моделирования. В связи с этим все более актуальным становится онтологический подход к проектированию программных систем. Как правило, при таком подходе к моделированию системы требуется определиться со способом реализацией ролей.