УДК 519.766.4

# МЕТОДИКА ПОСТРОЕНИЯ И ПРИМЕНЕНИЯ ВЕРОЯТНОСТНЫХ СЕТЕВЫХ МОДЕЛЕЙ

**П.И. Бидюк[1], О.А. Кожуховская[2], И.А. Загорская[1]**

*[1]Национальный технический университет Украины «КПИ»*
*[2]Черкасский государственный технологический университет*

Предложена практическая методика построения графических вероятностных моделей для решения задач прогнозирования, классификации и диагностики. Методика охватывает все этапы построения модели и ее использования. Приведен пример построения вероятностной модели в форме байесовской сети для решения задачи прогнозирования финансового состояния предприятия. Выполнен сравнительный анализ результатов, полученных экспертным путем, с помощью БС, а также моделей логит и пробит.

Запропоновано практичну методику побудови графічних імовірнісних моделей для розв'язання задач прогнозування, класифікації та діагностики. Методика охоплює всі етапи побудови моделі та її застосування. Наведено приклад побудови ймовірнісної моделі у формі байєсівської мережі для розв'язання задачі прогнозування фінансового стану підприємства. Виконано порівняння результатів, отриманих експертним шляхом, за допомогою БМ, а також за моделями логіт і пробіт.

**Введение.** При решении многих практических задач моделирования процессов различной природы, прогнозирования и поддержки принятия решений возникает необходимость оценивания гипотез, в отношении которых нет полной (достаточной) информации. Часто непросто вычислить приемлемые по качеству оценки переменных и параметров, но, несмотря на наличие неопределенностей, удается найти приемлемые рациональные решения. Компьютерные системы поддержки принятия решений, которые широко используются в различных отраслях деятельности, также должны содержать средства учета неопределенностей при построении моделей исследуемых процессов и формировании альтернатив [1]. Классическими примерами задач с неопределенностями являются техническая и медицинская диагностика, анализ и менеджмент рисков в технике, экономике, финансах, прогнозировании и управлении в условиях неполной информации и т.п. [1, 2, 3]. Например, довольно часто сомнения относительно правильности диагноза возникают даже в тех случаях, когда явно проявляются соответствующие симптомы заболевания. Неопределенность может проявляться различными способами; она может существовать даже в истинности формулирования самой решаемой задачи. Например, если степень уверенности в событии $A$ составляет 90 %, то насколько можно быть уверенным в событии $B$, которое связано с $A$? Неопределенность может существовать в самом правиле формирования решения; так, в большинстве ситуаций, но не всегда, если появляется событие $A$, то появляется также и $B$.

Более сложная ситуация возникает в случае, когда правило имеет вид: если ($A$ и $B$), то $C$. Здесь необходимо учитывать истинность каждого события, которое принимается во внимание, а также истинность совместного появления $A$ и $B$. В общем случае можно выделить четыре задачи, которые возникают при использовании информации с неопределенностями [2–4]: 1. Каким образом

можно количественно определить степень определенности при установлении истинности (или неправдивости) некоторой части данных и знаний? 2. Как правильно выразить степень поддержки конечного вывода (результата) конкретным формулированием? 3. Как правильно использовать вместе две (или более) формулировки, которые независимо влияют на вывод? 4. Как корректно сформулировать вывод для того, чтобы подтвердить начальную формулировку задачи при наличии неопределенностей?

Как показано во многих исследованиях, такие задачи можно корректно формулировать и решать с помощью теории вероятностей [3–5]. В частности, путем применения к решению задач с неопределенностями байесовских методов анализа данных и конкретно байесовских сетей (БС). Именно БС и некоторые другие байесовские модели стали сегодня эффективным, широко используемым инструментом моделирования ситуаций с неопределенностями и решения задач моделирования, классификации, прогнозирования, диагностики и управления [3–7].

**Цель работы** состоит в анализе проблем, связанных с построением вероятностных моделей в форме байесовских сетей, и применении построенных моделей к решению задачи оценивания кредитного риска.

**Постановка задачи.** Разработать практическую методику построения БС в условиях полных наблюдений, которые используются для обучения структуры и параметров модели. Полученные структуры БС должны быть пригодными для решения задач прогнозирования, классификации и диагностики на основе соответствующих статистических данных и экспертных оценок. Привести пример построения сети для решения задачи прогнозирования финансового состояния предприятия и выполнить сравнительный анализ результатов, полученных с помощью различных методов.

**Методика построения сетевых вероятностных моделей**. Формально сеть Байеса представляет собой направленный ациклический граф (НАГ) $\mathbf{G}$ на множестве переменных $X_1, X_2, ..., X_n$, между которыми существуют некоторые причинно-следственные связи. Каждой переменной соответствует вершина графа, а направленные дуги, соединяющие вершины, указывают на существующие зависимости между переменными. Дочерние узловые переменные описывают таблицами условных распределений вероятностей состояний этих переменных при условии, что родительские узлы принимают определенные значения. Совместное распределение вероятностей состояний переменных НАГ определяется выражением [8, 9]:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid \mathbf{G}) = \\ = P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i \mid \mathbf{x}_{pa(i)}, \mathbf{G}), \quad (1)$$

где $\mathbf{x}_{pa(i)}$ — вектор непосредственных родительских переменных (узлов) для $X_i$; $P(x_1, x_2, ..., x_n \mid \mathbf{G})$ — вероятность появления конкретной комбинации

значений $x_1, x_2, ..., x_n$ для множества переменных $X_1, X_2, ..., X_n$. БС структурируются локально таким образом, что каждый узел взаимодействует только со своими родительскими узлами. Условные распределения вероятностей для дискретных переменных представляются множеством соответствующих (многомерных) таблиц с параметрами

$$\Theta = \{\theta_1, \theta_2, ..., \theta_n\} = \{\theta_{ik}(j)|_{k=1}^{r_i}\}_{j=1}^{q_i},$$

где $i = 1, ..., n$ — номера переменных $X_i \in \mathbf{X}$; $k = 1, ..., r_i$ — индекс, который указывает на значения переменной $X_i$; $j = 1, ..., q_i$ — индекс, указывающий на множество допустимых комбинаций значений родительских переменных для $X_i$ $(x_{pa(i)})$. Теперь уравнение (1) можно представить в виде

$$P(\mathbf{x} \mid \Theta, \mathbf{G}) = \prod_{i=1}^{n} P(x_i \mid x_{pa(i)}, \Theta_i, \mathbf{G}) = \prod_{i=1}^{n} \theta_{ik}(j). \quad (2)$$

Рассмотрим задачу построения байесовской сетевой модели на основе выборки данных мощностью $N$ значений. Обозначим через $\mathbf{D} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}\}$ множество векторов данных, сформированных из значений состояний $x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}$ переменных $X_1, X_2, ..., X_n$. При этом возможны такие случаи: 1) структура БС известна, а необходимо оценить только ее параметры (значения таблиц условных вероятностей); 2) необходимо оценить структуру и параметры сети. Возможны особые случаи построения модели, когда некоторые вершины скрыты или когда данные неполные. Поэтому, как правило, рассматривают четыре случая обучения сети, которые приведены в табл. 1.

**Таблица 1**

*Четыре случая обучения байесовских сетей*

| Структура | Наблюдения | Метод |
|---|---|---|
| Известна | Полные | Метод максимального правдоподобия |
| Известна | Неполные | Градиентные методы, EM-алгоритм (максимизация математического ожидания), применение выборки Гиббса |
| Неизвестна | Полные | Поиск в пространстве моделей |
| Неизвестна | Неполные | Структурный EM-алгоритм, алгоритм сжатия границ |

Когда есть полные наблюдения для оценивания параметров сети, можно воспользоваться методом максимального апостериорного оценивания (МАО). Для дискретных переменных такие оценки представляют собой относительные частоты появления каждого значения для каждой переменной при известной конфигурации родительских узлов. Для реализации процедуры оценивания по методу МАО необходимо знать (выбрать) априорное

распределение для параметров модели. Для этой цели часто используют распределение Дирихле.

В случае, если структура сети неизвестна, необходимо оценить сначала структуру графа **G**. Это включает в себя создание спецификаций относительно условной независимости между переменными модели и параметрами Θ. Оценивание структуры БС на основе статистических данных можно выполнить двумя путями: с помощью оптимизационных методов с учетом ограничений и с помощью поисковых процедур на основе скоринговых функций [5, 6, 9]. Методы, которые основываются на ограничениях, достаточно эффективны, но они не имеют практически необходимой робастности. Т.е. полученные в результате их применения структуры очень чувствительны к ошибкам статистического тестирования на условную независимость. Таким образом, в большинстве случаев решения практических задач используют поисковые алгоритмы на основе скоринговых функций. Среди скоринговых функций популярна байесовская, которая основывается на апостериорной вероятности графа **G**; функция на основе аппроксимации апостериорных распределений вероятностей с использованием информационного критерия Байеса; скоринговые функции на основе описания минимальной длины (ОМД) и информационно-геометрического критерия (info-geo) [8, 10–12].

Предлагаемая методика построения сети на основе статистических данных состоит из таких шагов: 1) постановка задачи исследования; 2) анализ исследуемого процесса (объекта) и выбор множества переменных для его описания, анализ глубины взаимосвязей между переменными; 3) редукция размерности задачи моделирования; 4) масштабирование и дискретизация значений переменных; 5) определение семантических ограничений; 6) оценивание сетевых моделей-кандидатов; 7) анализ качества моделей и выбор лучшей из них для решения поставленной задачи исследования. Рассмотрим эту методику подробнее.

*Шаг* 1. Постановка задачи исследования должна быть максимально конкретизированной и подробной. Она может касаться следующего: построение вероятностной модели для углубленного исследования объекта; вероятностное прогнозирование его будущего состояния; классификация некоторого множества элементов и т.п. Очевидно, что от постановки задачи зависят требования к адекватности модели и качество *окончательного результата. С другой стороны, от корректности постановки задачи* зависят временные, вычислительные и материальные затраты на получение ее решения.

*Шаг* 2. Выявление глубины взаимосвязей между переменными и редукция (сокращение) размерности задачи с помощью известных статистических методов. Для оценивания глубины взаимосвязей между переменными используют известные статистические параметры (статистики), например, коэффициенты ассоциации, контингенции Бравайса или взаимной сопряженности. В табл. 2 приведены примеры вычисления частот

взаимосвязей между двумя переменными.

**Таблица 2**

*Таблица значений частот взаимосвязей переменных* $X^{(1)}$ *и* $X^{(2)}$

| **Переменные** | $X^{(1)} = x_1^{(1)}$ | $X^{(1)} = x_2^{(1)}$ | **Сумма** |
|---|---|---|---|
| $X^{(2)} = x_1^{(2)}$ | $a$ | $b$ | $a+b$ |
| $X^{(2)} = x_2^{(2)}$ | $c$ | $d$ | $c+d$ |
| **Сумма** | $a+c$ | $b+d$ | $n$ |

В табл. 2 $a, b, c, d$ — частоты взаимосвязей двух переменных; $n$ — сумма частот.

*Коэффициент ассоциации Юла* для переменных $X^{(1)}$ и $X^{(2)}$ (табл. 2) вычисляется по формуле

$$K_{ac} = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}.$$

*Коэффициент контингенции (сопряженности) Бравайса* между переменными $X^{(1)}$ и $X^{(2)}$ в табл. 2 вычисляется по формуле

$$K_{kon} = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (b+d) \cdot (a+c) \cdot (c+d)}}.$$

Коэффициенты ассоциации и контингенции принимают значения в интервале $\pm 1$. При этом следует отметить, что для одних и тех же данных значение коэффициента контингенции всегда меньше коэффициента ассоциации. Значение $+1$ означает наличие полной положительной связи, а $-1$ означает полную отрицательную связь; нулевое значение соответствует случаю независимости переменных. В случае когда переменные модели имеют больше двух состояний, как например $X^{(1)}$ и $X^{(2)}$ в табл. 3, для анализа взаимосвязи применяются коэффициенты взаимной сопряженности.

Коэффициент взаимной сопряженности Пирсона вычисляется по формуле $K = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$, где $\varphi^2$ — показатель среднеквадратической сопряженности, вычисляемый по выражению:

$$\varphi^2 = \left( \sum_{j=1}^{n} \left( \sum_{i=1}^{m} \left( \frac{k_{ji}}{\left( \sum_{j=1}^{n} k_{ji} \right) \cdot \left( \sum_{i=1}^{m} k_{ji} \right)} \right) \right) \right).$$

Этот коэффициент принимает значения от нуля до единицы. При

нулевом значении переменные независимы, а единица означает, что значения одной переменной можно точно спрогнозировать по значениям другой.

**Таблица 3**

*Таблица значений частот взаимосвязи переменных* $X^{(1)}$ *и* $X^{(2)}$

| **Переменные** | $X^{(1)} = x_1^{(1)}$ | … | $X^{(1)} = x_n^{(1)}$ | **Сумма** |
|---|---|---|---|---|
| $X^{(2)} = x_1^{(2)}$ | $k_{11}$ | … | $k_{n1}$ | $\sum_{j=1}^{n} k_{j1}$ |
| … | … | … | … | … |
| $X^{(2)} = x_m^{(2)}$ | $k_{1m}$ | … | $k_{nm}$ | $\sum_{j=1}^{n} k_{jm}$ |
| **Сумма** | $\sum_{i=1}^{m} k_{1i}$ | … | $\sum_{i=1}^{m} k_{ni}$ | $\sum_{i=1}^{m}\sum_{j=1}^{n} k_{ji}$ |

В работе [13] для оценивания степени зависимости двух произвольных переменных $x^i$ и $x^j$ впервые предложено использовать значение взаимной информации (ЗВИ) $\mathrm{MI}(x^i, x^j)$:

$$MI(x^i, x^j) = \sum_{x^i, x^j} P(x^i, x^j) \cdot \log\left( \frac{P(x^i, x^j)}{P(x^i) \cdot P(x^j)} \right).$$

По своей сути ЗВИ — это аналог корреляции, но по своему содержанию — это оценка количества информации, которая содержится в переменной $x^i$ о переменной $x^j$. ЗВИ принимает только неотрицательные значения, т.е. $MI(x^i, x^j) \geq 0$, а в случае полной независимости переменной она равняется нулю, поскольку $P(x^i, x^j) = P(x^i) \cdot P(x^j)$, а отсюда имеем

$$\log\left( \frac{P(x^i, x^j)}{P(x^i) \cdot P(x^j)} \right) = \log\left( \frac{P(x^i) \cdot P(x^j)}{P(x^i) \cdot P(x^j)} \right) = \log(1) = 0.$$

В случае когда БС имеет $N$ вершин, для вычисления $MI(x^i, x^j)$ для всех пар $x^i$ и $x^j$ необходимо выполнить $\frac{N \cdot (N-1)}{2}$ вычислений; при этом необходимо учитывать, что $MI(x^i, x^j) = MI(x^j, x^i)$.

*Шаг* 3. Редукция размерности модели. Для решения задачи редукции размерности модели используют такие методы: линейный и нелинейный методы главных компонент (МГК); факторный анализ (ФА); многомерное шкалирование (ММШ); методы обучения на нелинейных структурах (например, локальное линейное погружение) и другие методы. Факторный анализ, МГК и ММШ основываются на вычислении и использовании собственных векторов. Так, МГК предполагает вычисление линейных

проекций максимальной дисперсии, которые определяются с помощью собственных векторов ковариационной матрицы измерений. Факторный анализ используется для выявления и моделирования корреляционной структуры данных, выключая при этом их случайные вариации. МГК чаще применяют для редукции размерности модели, а ФА — для выявления структурных взаимосвязей между переменными. Метод МШ обеспечивает вычисление проекций малой размерности, которые сохраняют попарные расстояния между значениями измерений.

*Шаг* 4. Большинство известных алгоритмов оценивания структуры и параметров вероятностных моделей, а также формирования вывода на их основе используют дискретные данные. Поэтому на этом шаге может выполняться масштабирование распределений с целью их приведения к удобной для дальнейших вычислений форме и дискретизация непрерывных переменных. Поскольку на качество функционирования модели влияет качество соответствующих статистических данных и экспертных оценок, то необходимо надлежащим образом подготовить данные в соответствии с требованиями теории оценивания.

«Нестандартные» распределения, например распределения с явно выраженной асимметрией, логарифмируют или преобразуют по методу квадратного корня с целью их приближения к распределениям известных форм. Очевидно, что при этом теряется исходный масштаб данных; что необходимо учитывать при интерпретации результатов оценивания структур и параметров моделей. Для дискретизации данных разработано несколько эффективных схем, которые обеспечивают получение рациональных интервалов в процессе дискретизации (например, интервалов одинаковой ширины или интервалов одинаковых частот попадания значений).

Очевидно, что значения выборки должны быть представлены в каждом интервале. Количество интервалов необходимо сокращать, поскольку это дает возможность уменьшить количество оцениваемых параметров. Так, если сеть состоит из десяти бинарных переменных и каждая переменная имеет в среднем по три родительских узла, то для такой сети необходимо оценить около $10 \cdot 2^3 = 80$ параметров. Если же сеть состоит из тернарных переменных, то для нее необходимо оценить $10 \cdot 3^3 = 270$ параметров.

С другой стороны, сокращение размерности данных приводит к уменьшению их разрешающей возможности (точности представления измерений и экспертных оценок), т.е. уменьшается точность представления входных и выходных данных, а также связанных с ними оценок вероятностей. Гранулярность (глубина) любого анализа определяется количеством имеющихся вариантов событий и подробностью данных. Отсюда следует, что для углубленного ситуационного анализа процессов и объектов произвольной природы необходимо иметь большие массивы данных, обеспечивающих высокую точность измерений. При наличии пропусков данных необходимо использовать более сложный в реализации метод максимизации математического ожидания (EM-алгоритм) [14, 15].

*Шаг* 5. Формулирование семантических ограничений. В процедуре поиска лучшей структуры сети необходимо задавать контекстно-направленные семантические ограничения, касающиеся ограничения области поиска возможных структур. Эта рекомендация касается процедур поиска любого типа (т.е. полного и неполного перебора). Поскольку размерность пространства поиска растет экспоненциально с увеличением количества переменных, то полный перебор практически невозможен. Так, для трех переменных пространство поиска сетевых структур ограничено 25 направленными ациклическими графами (11 эквивалентных марковских классов), а для 10 переменных это число возрастает до $3 \cdot 10^{17}$ ($1 \cdot 10^{17}$ эквивалентных марковских классов) [16]. Семантические ограничения дают возможность сократить пространство поиска только теми структурами моделей, которые согласуются с выбранными временными прецедентами или иными требованиями к зависимостям между переменными. Это автоматически сокращает время, необходимое для обработки данных. Вполне приемлемые основания для формулирования семантических ограничений дают базовая теория каузальных структур и оценки экспертов. Семантические ограничения дают возможность сократить количество возможных реализуемых структур и повысить вероятность построения рациональной структуры благодаря использованию знаний о предметной области [14, 15].

*Шаг* 6. Поиск возможных структур моделей-кандидатов. На этом шаге из множества возможных структур моделей необходимо выбрать несколько лучших моделей-кандидатов с помощью соответствующих критериев качества. Поиск основывается, как правило, на эвристических алгоритмах с использованием скоринговых функций (СФ), которые позволяют сравнить построенные модели. Результатом использования каждой комбинации скоринговой функции, алгоритма поиска структуры и соответствующей выборки данных является модель-кандидат, т.е. БС определенной структуры. Таким образом, задача оценивания структуры имеет оптимизационный характер благодаря комбинации скоринговой функции и алгоритма поиска. Целью решения этой оптимизационной задачи является оценивание структуры направленного ациклического графа $\mathbf{G}$ в пространстве допустимых структур $\Omega^G$, который минимизирует значение скоринговой функции и соответствует обучающим данным $\mathbf{D}$.

В качестве скоринговой функции естественно использовать апостериорное распределение вероятностей

$$P(\mathbf{G} \mid \mathbf{D}, \Theta) \propto P(\mathbf{D} \mid \mathbf{G}) \cdot P(\mathbf{G}).$$

Однако вычисление точных значений этой функции даже для сетей небольшой размерности требует значительных вычислительных затрат. Поэтому при оценивании распределения $P(\mathbf{G} \mid \mathbf{D})$ делают упрощения, например, относительно типа распределения. Так, в работе [4] предложен

алгоритм К2, для реализации которого принято равномерное априорное распределение для $P(\mathbf{G})$, а маргинальное правдоподобие $P(\mathbf{D}|\mathbf{G})$ рассчитывается с использованием сопряженного распределения Дирихле для параметров сети. К2 основывается на алгоритме «жадного» поиска локального экстремума и таком упорядочивании структуры сети, что для каждой переменной $X_i$ добавляется такой родительский узел, который способствует максимизации значения скоринговой функции. Эта процедура повторяется для каждой переменной $X_i$, до тех пор пока не прекратится увеличение значения скоринговой функции или количество переменных $X_i$ не превысит заданный порог.

Простую аппроксимацию апостериорного распределения вероятностей БС можно обеспечить с помощью других скоринговых функций. Так, байесовский информационный критерий (БИК) представляет собой оценку маргинального правдоподобия модели на больших выборках. Отметим, что для достижения аппроксимации приемлемого качества не нужны большие выборки данных. Кроме того, в этом случае не нужно задавать априорное распределение для параметров. Скоринговая функция info-geo, представляющая собой модификацию БИК, имеет вид [17]

$$Ig = -\log P(\mathbf{D}|\hat{\Theta}) + \frac{|\Theta|}{2}\log\frac{N}{2p} + \log\int(\det\mathbf{I}(\Theta))^{1/2}\,d\Theta,$$

где первый член — это логарифм правдоподобия с использованием оценок $\hat{\Theta}$, вычисленных по методу максимального правдоподобия; другой член характеризует меру сложности модели, которая определяется количеством ее параметров; последняя составляющая критерия включает детерминант информационной матрицы Фишера $\mathbf{I}(\Theta)$ и интерпретируется как «геометрическая» мера сложности. Первые два члена функции info-geo соответствуют БИК с обратным знаком.

Известной мерой качества обучения является так называемое описание минимальной длины (ОМД). В соответствии с теорией кодирования Шеннона, если известно распределение $P(X)$ случайной переменной $X$, то длина оптимального кода для передачи значения $X = x$ через канал связи определяется выражением $L(x) = -\log P(x)$. Энтропия источника $S(P) = -\sum_x P(x)\cdot\log P(x)$ представляет собой минимальную ожидаемую длину закодированного сообщения. Любой иной код, который основывается на некорректном представлении источника сообщений, приведет к появлению сообщения большей длины. Т.е. чем лучше модель источника, тем компактнее можно закодировать данные.

В задаче обучения БС источник информации — некоторая неизвестная функция распределения $P(D|h_0)$, где $D = \{d_1, ..., d_N\}$ — данные; $h_0$ — гипотеза относительно вероятностной природы данных. Если ввести

функцию эмпирического риска $L(D|h) = -\log P(D|h)$, пропорциональную эмпирической погрешности оценивания распределения, то разница между $P(D|h_0)$ и модельным распределением $P(D|h)$ в соответствии с мерой Кульбака-Лейблера определяется так:

$$\left|P(D|h) - P(D|h_0)\right| = \sum_D P(D|h_0) \cdot \log \frac{P(D|h_0)}{P(D|h)} =$$

$$= \sum_D P(D|h_0) \cdot \left|L(D|h) - L(D|h_0)\right| \geq 0 .$$

Таким образом, эта мера представляет собой разницу между ожидаемой длиной кода (в соответствии с выдвинутой гипотезой) и минимально возможной. Эта разница всегда неотрицательна и равняется нулю только при полном равенстве двух распределений. В обобщенной формулировке принцип ОМД означает, что из множества моделей необходимо выбрать ту, которая дает возможность максимально компактно описать данные без потери информации.

Для поиска глобального оптимума можно использовать генетический алгоритм или поисковые методы на основе процедуры Монте-Карло для марковских цепей [8, 9, 17]. При использовании алгоритма моделирования отжига каждая сетевая структура интерпретируется как состояние марковской цепи. На каждом шаге поисковой процедуры сетевая модель вынужденно переводится из одного состояния в другое. При этом возмущения для БС реализуются с помощью трех операций: добавление дуги, исключение дуги или изменение ее направления. Эти операции дают возможность создавать множество потенциальных структур, из которых случайным образом выбирается одна для исследования с помощью скоринговой функции. Таким образом, алгоритм поиска дает возможность выбирать сети с улучшенными значениями скоринговых функций для дальнейшей обработки. При этом пространство поиска сокращается благодаря исключению ациклических структур и применению семантических ограничений. Этот алгоритм требует больших вычислительных затрат, чем алгоритм «жадного» поиска, но он имеет более высокую вероятность сходимости к глобальному максимуму.

*Шаг* 7. На этом этапе сравниваются характеристики моделей-кандидатов с целью выбора лучшей для описания выбранных характеристик исследуемого процесса. Для оценивания качества моделей такого типа применяют критерии точности прогнозирования с использованием соответствующих тестовых данных. Если модель строится для решения задачи классификации, то для оценивания ее качества используют усредненную полезность (или стоимость), полученную с помощью вероятностных прогнозов. Такой подход может быть применен в тех случаях, когда есть возможность получить информацию о стоимости возможных потерь от некорректной классификации или о полезности, достигнутой благодаря правильной обработке данных. Приемлемой метрикой для

сравнения истинного совместного распределения вероятностей (а оно всегда неизвестно) с его оценкой является расстояние Кульбака-Лейблера, которое часто рассматривают как некую стандартизованную оценку качества построенной модели.

Очевидно, что основным критерием качества модели является ее использование для решения поставленной задачи. БС дают возможность формировать вероятностные выводы различными методами и получать, таким образом, альтернативные результаты, из которых можно выбрать лучший.

**Пример построения байесовской сети для анализа банкротства предприятия**

Для построения модели использованы 14 финансовых показателей функционирования предприятия. Это фактическая информация, полученная от банковского учреждения, выдававшего кредиты юридическим лицам в 2009 г. Выборка содержит данные о 395 фирмах, в основном представителей крупного и среднего бизнеса. Рассмотрим кратко использованные показатели: ANS (Annual sales) — объем продаж товаров и услуг за год; INDEF (In default) – индикатор, указывающий на пребывание (или нет) предприятия в состоянии дефолта в предыдущем году; LGD (Loss Given Default) — индикатор, указывающий на часть капитала, вложенного в предприятие, которую потеряет инвестор в случае дефолта; BAL (Balance) — баланс — сумма, снятая с линии кредитования; LBD (Leverage Buyout Deal) — индикатор, указывающий на то, что договор является левериджевым, т.е. кредит берется с целью покупки контрольного пакета акций или производственных мощностей другого предприятия; ROA (Return on Assets) — индикатор прибыльности акций предприятия в процентах; DSCR (Debt Service Ratio) — коэффициент обслуживания задолженности; это показатель объема денежных ресурсов, доступных для обслуживания долга; чем выше значение этого показателя, тем проще получить кредит; CIC (Cash Interest Cover Ratio) — коэффициент наличного покрытия процентов; это мера возможности выплаты компанией процентов по задолженности до выплаты налогов и процентов; чем ниже этот показатель, тем выше вероятность банкротства; DDER (Debt to Debt + Equity Ratio) — долги/(долги + собственный капитал); CFOL (Cash Flow from Operations to Liabilities) — отношение среднего взвешенного операционного денежного потока к общим обязательствам; EBITDA (Earnings before Interest, Taxes, Depreciation and Amortization Volatility) – волатильность прибыли до выплаты процентов, налогов, снижения стоимости и амортизации; NWC (Net worth CPI) — собственный капитал относительно индекса рыночных цен, это показатель чистой стоимости компании за последний год с учетом инфляции; NPAT (Negative Net Profit After Tax Flag) — индикатор отрицательной чистой прибыли после выплаты налогов за последний год; MP (Market Position) — положение на рынке.

В качестве прогнозируемой переменной выбрана PD (Probability of

Default) — вероятность банкротства. Для оценивания точности прогнозирования с помощью БС в дальнейшем будут использованы также экспертные оценки вероятности банкротства предприятий, входящих в обучающую выборку.

**Оценивание структуры модели**. В работе выполнено сравнение двух моделей — простой и каскадной БС. Выбранные переменные разделены на две группы в зависимости от степени корреляции с основной переменной PD. К первой группе отнесены переменные с коэффициентом корреляции большим, чем 0,1 (табл. 4). Поскольку эти переменные имеют наибольшее влияние на риск банкротства, то они использованы на первом уровне БС. Ко второй группе отнесены переменные с меньшим значением корреляции (табл. 5).

**Таблица 4**

*Корреляция между PD и предикторами первого уровня*

| | |
|---|---|
| NPAT | 0,440220551 |
| IN DEF | 0,319509857 |
| EBITDA | 0,252328868 |
| AN S | 0,147820080 |
| LBD | 0,108973829 |
| CFOL | -0,164349051 |
| NWC | -0,396602581 |
| BAL | -0,095535831 |

**Таблица 5**

*Корреляция между PD и предикторами второго уровня*

| | |
|---|---|
| DDER | 0,06566500 |
| LGD | 0,06549991 |
| MP | 0,03289091 |
| ROA | 0,00874046 |
| CIC | -0,02270559 |
| DSCR | -0,02446745 |

Узлы второго уровня должны давать недостающую информацию узлам первого уровня. Поэтому дочерними вершинами для определенных вершин первого уровня должны быть те, которые имеют максимальный коэффициент корреляции с родительскими вершинами (табл. 6). Из анализа табл. 6 следует, что для некоторых предикторов первого уровня (например, для NW) существует больше одной дочерней вершины, а для некоторых предикторов второго уровня (например, для DSCR) имеется больше одного родительского узла.

В результате проведенного анализа можно построить простую и каскадную БС для оценивания состояния предприятия.

Простая сеть включает только предикторы первого уровня, которые являются наиболее значимыми (рис. 1). Каскадная модель включает

предикторы первого и второго уровней (рис. 2).

**Таблица 6**

*Корреляция между предикторами первого и второго уровней*

| Переменная | DDER | LGD | MP | ROA | CIC | DSCR |
|---|---|---|---|---|---|---|
| **NPAT** | -0,00894 | 0,117601 | -0,11384 | -0,02001 | -0,03475 | **-0,14772** |
| **IN DEF** | -0,00897 | -0,10259 | -0,06572 | -0,00279 | -0,00485 | 0,001083 |
| **EBITDA** | 0,037038 | -0,06717 | -0,07193 | 0,071313 | **0,15956** | **-0,12545** |
| **AN S** | -0,09649 | 0,0812 | **0,190491** | -0,00978 | -0,01551 | -0,03875 |
| **LBD** | 0,127637 | -0,06611 | 0,145117 | -0,04283 | 0,116547 | -0,02602 |
| **CFOL** | -0,1302 | 0,17587 | 0,143272 | -0,02644 | **0,204045** | 0,064543 |
| **NWC** | **-0,45732** | **0,248318** | **0,293683** | **-0,10257** | 0,020369 | -0,04113 |
| **BAL** | **0,258413** | **-0,29473** | -0,0573 | -0,02587 | -0,0123 | -0,02841 |

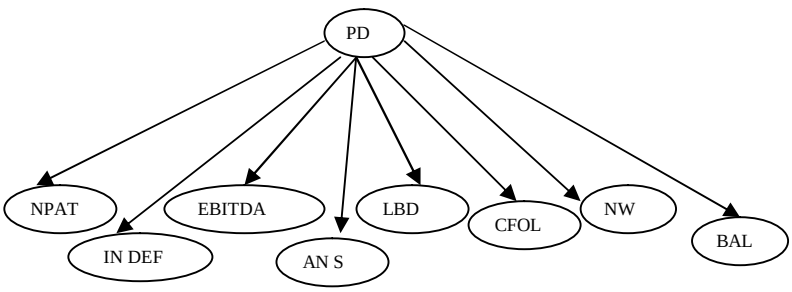


Рис.1. Схема простой БС для оценивания состояния предприятия

**Анализ результатов применения модели**. Для уточнения результатов моделирования обучающая выборка была разделена на пять групп по 65 значений. Отдельно рассмотрена выборка из 25 значений, которая включала 21 фирму, не склонную к банкротству, и четыре фирмы — банкроты. На каждом из пяти обучающих подмножеств построена модель в виде простой (наивной) БС. Эта модель использована для прогнозирования дефолта предприятий из обучающей выборки. Затем с помощью пяти моделей было вычислено среднее значение вероятности дефолта для каждой из 25 фирм из обучающей выборки. Этот результат принят как общий прогноз, полученный с помощью простой БС. Корректная идентификация состояния получена для всех предприятий (21 + 4), т.е. ошибка классификации равнялась нулю. Это свидетельствует о правильно выбранной структуре модели в форме БС.

Полученные результаты были сравнены с экспертными оценками вероятности дефолта для каждого предприятия из обучающей выборки. При этом наблюдалось полное совпадение результатов оценивания состояния предприятий.

При использовании каскадной модели были корректно идентифицированы успешные предприятия, а для одного из предприятий-

банкротов спрогнозирована вероятность на уровне 0,475. Таким образом, точность определения дефолтов на выбранных данных составила 75 %. Соответствующий график сравнения результатов с экспертными оценками приведен на рис. 3.

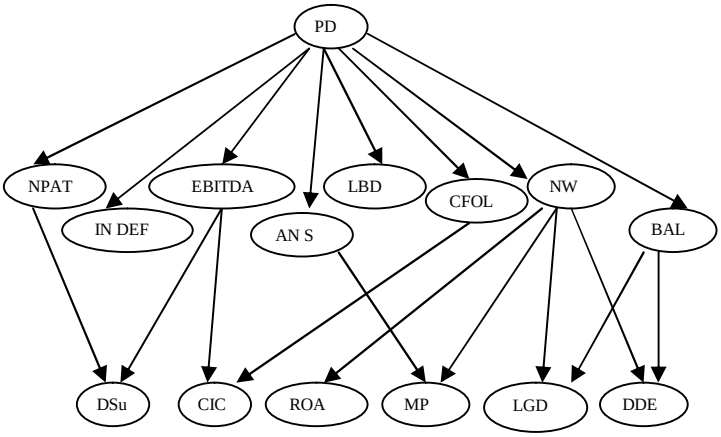


Рис. 2. Схема каскадной БС для оценивания состояния предприятия

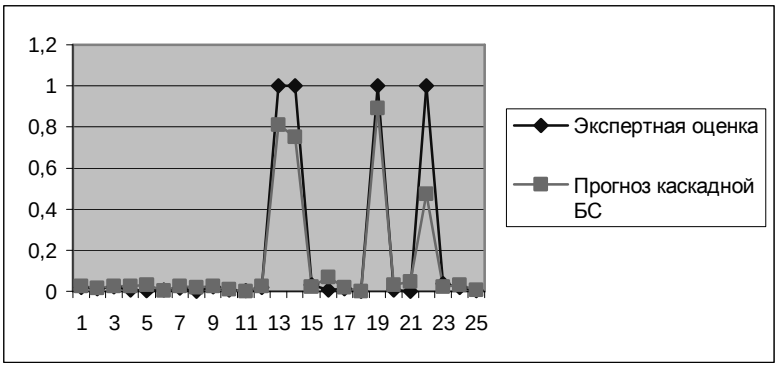


Рис. 3. Сравнение прогнозов (каскадная БС) с экспертными оценками вероятности дефолта

Результаты, полученные с помощью БС, сравнивались с результатами применения моделей двоичного выбора логит и пробит (табл. 7). Как видно из табл. 7, использование БС в данном случае дает возможность существенно улучшить результаты прогнозирования состояния предприятия.

**Таблица 7**

*Сравнение точности прогнозов по БС, логиту и пробиту*

| Модель | Корректно идентифицированные успешные предприятия, % | Корректно идентифицированные предприятия-банкроты, % | Общая точность, % |
|---|---|---|---|
| **Простая БС** | 100 | 100 | 100 |
| **Каскадная БС** | 100 | 75 | 96 |
| **Логит** | 95,24 | 50 | 88 |
| **Пробит** | 95,24 | 50 | 88 |

**Выводы**

Выполнен анализ проблем, связанных с построением вероятностных моделей в форме байесовских сетей. Предложена практическая методика построения БС, пригодных для решения задач прогнозирования, классификации, диагностики и т.п. Методика охватывает все этапы построения модели и ее дальнейшее использование. Значительное внимание уделено обучению структуры модели на основе статистических данных и экспертных оценок. Показано, что существует множество критериев оптимизации структуры (скоринговые функции) и алгоритмов поиска экстремума, которые обеспечивают возможность получения альтернативных структур-кандидатов и выбрать лучшую из них для практического использования.

Следует отметить, что в целом процесс построения вероятностных моделей в форме БС основывается на применении нетрадиционных методов и подходов к анализу данных на всех этапах моделирования. Это обусловлено сложностью процесса выявления возможных причинно-следственных связей между переменными, высокой размерностью моделей, объединением в одной структуре дискретных и непрерывных переменных, необходимостью использования альтернативных оптимизационных методов и функционалов качества, а также возникающей иногда сложностью интерпретации полученных результатов.

Рассмотрен пример применения методики построения БС к решению практической задачи прогнозирования финансового состояния предприятия на основе фактических данных. Показано, что качество результатов прогнозирования, полученных с помощью БС, может превосходить качество результатов, полученных с помощью нелинейных регрессионных моделей логит и пробит.

В дальнейших исследованиях целесообразно усовершенствовать предложенную методику путем введения автоматизированных процедур поиска лучших структур и формирования окончательного результата – вероятностного вывода. Это даст возможность существенно ускорить процесс моделирования и практического применения построенной модели. Также необходимо модифицировать методику построения динамической БС на прогнозирование временных рядов.


1. Pearl J. *Causality*: *models, reasoning and inference*. Cambridge, Cambridge University Press Publ., 2000. 400 p.
2. Zgurovskyj M.Z., Bidyuk P.I., Terentyev A.N. Method of constructing Bayesian networks based on scoring functions. *Cybernetics and System Analysis*, 2008, vol. 44, no. 2, pp. 219–224.
3. Bidyuk P.I., Davydenko V.I., Trofimenko D.V., Terentyev A.N. Comparative analysis of estimation methods of vertices correlation while Bayesian networks construction. *Journal of Automation and Information Sciences*, 2009, vol. 42, no. 11, pp. 36–45.
4. Dagum P., Galper A., Horvitz E., Seiver A. Uncertain reasoning and forecasting *Int. J. of Forecasting*. 1995, vol. 11, pp. 73–87.

5. Cooper G.A., Cooper G., Herskovits E. Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 1992, vol. 9, no. 4, pp. 309–347.
6. Cowell R.G., Dawid A.Ph., Lauritzen S.L., Spiegelhalter D.J. *Probabilistic networks and expert systems*. New York, Springer-Verlag Publ., 1999. 323 p.
7. Fan Chin Chang, Yu. Yuan-Chang BBN-based software project risk management. *Journal of Systems Software*, 2004, vol. 73, pp. 193–203.
8. Guidici P., Castelo R. Improving MCMC model search for data mining. *Machine Learning*, 2003, vol. 50, pp. 127–158.
9. Hastings W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 1970, vol. 57, no. 1, pp. 97–109.
10. Heckerman D., Geiger D., Chikering D. *Learning Bayesian networks: the combination of knowledge and statistical data*. Technical report MSR-TR-94-09, 1994. 54 p.
11. Jordan M. *Learning in Graphical Models.* MIT Press Publ., 1998. 644 p.
12. Korb A., Nicholson A. *Bayesian Artificial Intelligence*. London, Chapman & Hall Publ., 2004. 458 p.
13. Chow C.K., Liu C.N. Approximating discrete probability distributions with dependence trees. *IEE Transactions on Information Theory*. May 1968, vol. IT–14, no. 3, pp. 462–467.
14. Neapolitan R.E. *Learning Bayesian Networks*. Chicago (Illinois), Northeastern University Publ., 2004. 703 p.
15. Lee Sik-Yum. *Structural equation modeling*. New York, John Wiley & Sons, Ltd. Publ., 2008. 432 p.
16. Lam W., Bachus F. Learning Bayesian networks. An approach based on the MDL principle. *Computational Intelligence*, 1995, vol. 10, no. 3, pp. 269–293.
17. Lauritzen S., Spiegelhalter D. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B*, 1988, vol. 50, no. 2, pp. 157–224.

**P.I. Bidyuk, O.A. Kozhukhivska, I.O. Zahirska**

**A METHODOLOGY FOR CONSTRUCTION AND APPLICATION OF PROBABILISTIC NETWORK MODELS**

**Introduction**: While solving many practical problems of forecasting, classification and diagnostics researchers encounter with uncertainties of various nature and type. To achieve high quality of a final result the uncertainties should be taken into account with appropriate models. The following approaches are appropriate in such cases: fuzzy sets, neural networks, neural-fuzzy combinations, Bayesian models in the form of Bayesian networks (BN) and other types.

**Purpose**: The purpose of the work is to analyze the problems combined with probabilistic networks construction in the form of Bayesian networks, and to propose a methodology for the nets constructing, and to construct a specific model for estimation and prediction of an enterprise state using relevant statistical data.

**Methods of a research**: To fulfill the task outlined above a practically oriented method is proposed for Bayesian belief networks construction. The method includes the following steps: (1) – a detailed analysis of specific features of a process (object) under study and selection of a set of variables for its description; estimation of a degree of interrelations between the variables using a set of appropriate statistics; (2) – detailed problem statement on the basis of preliminary study of the process under consideration; (3) – if necessary and possible, reduction of a model dimensionality using a set of statistical procedures for multivariate data analysis; (4) – preliminary processing of selected variables: scaling and transforming into discrete form of the selected variables values; (5) – finding and determining constraints that can be formulated analytically or by experts; (6) estimation of a probabilistic model structure and its parameters, i.e. possible candidate models using appropriate optimization procedures; (7) – analysis of the candidate models quality (adequacy) and selection of the best one for a specific application defined by the problem statement.

**Results**: Using statistical data relevant to functioning of selected number of enterprises a simplified and more complicated Bayesian networks were constructed for forecasting of financial state of an enterprise. A comparative analysis of the results received from three sources is given: from experts, with Bayesian networks and with nonlinear logit and probit models. Comparison of the results showed that Bayesian networks models generally provide quite acceptable quality of prediction results that stay in line with other modern techniques or generate higher quality estimates for enterprise states.

**Conclusions**: Thus, we proposed a practically oriented multi-step methodology for Bayesian belief networks construction that was successfully applied for solving practical problem of forecasting an enterprise state. The results achieved with the belief network have been compared to expert estimates and logit/probit models. The BN models showed quite acceptable estimation results that stay in line with other modern techniques. It should also be noted that the model construction methodology presented in this paper requires further refinement directed towards improvement and automation of the entire model construction process.

***Keywords:*** probabilistic modeling, Bayesian networks, modeling methodology, structure and parameter estimation, practical application, comparative analysis.

Получено 11.02.2013