Экспертные системы, методы индуктивного вывода

УДК 519.68

А.А. ВАГИС

ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ НА БАЙЕСОВСКИХ СЕТЯХ

Приведены алгоритмы определения структуры байесовской сети в виде дерева или леса. Рассмотрены байесовские процедуры распознавания, построенные на обучающих выборках с использованием оценок переходных вероятностей заданных порядков.

Введение. В работах специалистов Института кибернетики имени В.М. Глушкова НАН Украины построена теория статистического оценивания дискретных процедур распознавания [1]. При решении задач распознавания определяющим моментом является структура описания объекта. Если она известна, то не представляет труда по обучающей выборке построить байесовские процедуры, как это проделано для цепей Маркова, и независимых признаков. Задача распознавания рассматривается с точки зрения минимизации среднего риска. Такая постановка является обобщением классических задач, решаемых на основе метода наименьших квадратов, в том смысле, что наблюдению объекта может соответствовать не одно, а несколько состояний объектов.

Каждый наблюдаемый объект описывается конечным набором признаков, принимающих конечное число значений. Описанию объекта может соответствовать конечное число состояний (исходов экспериментов), которые наблюдаются с неизвестными вероятностями. В результате проведенных опытов над объектами формируется конечная обучающая выборка, состоящая из множества описаний объектов и их состояний.

Требуется оценить эффективность процедуры распознавания, которая по измерениям набора признаков любого следующего объекта и обучающей выборке определяет состояние наблюдаемого объекта. Теория сложности дискретных задач распознавания развивается на основе байесовского подхода.

В работе [1] выбран прямой подход к построению методов распознавания: исследуются такие структуры описания объектов, для которых можно построить эффективные процедуры распознавания. Вместо выполнения достаточно сложных процедур минимизации эмпирического риска или других функционалов качества применяется простое байесовское правило классификации, которое используется для решения задач большой размерности. Показано, что байесовская процедура распознавания является оптимальной для объектов с независимыми признаками.

Байесовская процедура распознавания обладает многими свойствами, которые присущи иммунной системе человека: самообучаемая, обладает памятью, ошибкоустойчивая и самотестирующая, обладает параллельным характером вычислений и относительно проста. По нашему мнению, байесовские процедуры могут быть эффективно использованы в процессе разработки искусственных иммунных систем для решения широкого спектра прикладных задач.

Наиболее общим аппаратом для извлечения структур описания объектов из данных является аппарат байесовских сетей. За последние 10 лет выпущено более двух десятков монографий по байесовским сетям. Были проведены исследования вероятностных моделей зависимостей, структурированные ациклическими орграфами (АОГ-модели, байесовские сети). АОГ-модели являются наглядным и компактным языком представления систем зависимостей. Это эффективный аппарат отображения связей (вплоть до причинно-следственных отношений) и инструмент вероятностных рассуждений от свидетельств (в экспертных системах). Эти свойства делают АОГ-модели средством решения разнообразных задач, среди которых: медицинская и техническая диагностика, распознавание речи, прогнозирование последствий решений и действий, анализ данных в эконометрике, биоинформатике и т. д. Для решения исследовательских и прогнозно-аналитических задач необходимо вывести структуру модели на основе известной выборки данных наблюдений. Известно, что задача вывода АОГ-модели из данных в общем случае – NP-сложная. Вычислительные сложности становятся уже серьезной практической проблемой, когда число переменных модели достигает нескольких десятков.

При выводе структуры АОГ-модели из данных возникает проблема поиска и подбора сепараторов (фактов условной независимости между переменными в модели), которая решается на основе, так называемого, “constraint-based” (или сепарационного) подхода. Поиск сепараторов в сложных структурах выполняется комбинаторным путем с помощью статистических критериев. В работах [2, 3] были разработаны подходы к построению минимальных сепараторов модели. Полученные результаты открывают новые возможности по пути построения эффективных методов вывода структуры байесовской сети из имеющихся статистических данных и полученного набора минимальных сепараторов.

В пионерской работе [4] рассматривались вопросы извлечения структуры описания из статистических данных. Используя методологический принцип экономии в научном объяснении, известный как «бритва Оккама», вначале

предпочтение было отдано простейшей структуре, которая описывается лишь небольшим числом зависимостей. Такой разреженной структурой является сеть – дерево, она имеет минимальное число связей между переменными. Дерево – единственная структура графа, в которой связи могут быть ориентированы таким образом, что каждая переменная зависит не более чем от одной родительской переменной. Заметим, что аналогичная ситуация присуща однородным цепям Маркова.

Если каждая переменная X_i принимает r значений, то случайный вектор $\mathbf{x} = (x_1,...,x_n)$ может принимать r^n значений. Если вероятностное распределение P неизвестно и имеется независимая выборка $x^1,\ldots,\ x^s$, то для оценивания $P(x)$ требуется вычислить и сохранить r^n величин относительных частот векторов $\mathbf{x}$ среди наблюдений $x^1,\ldots,\ x^s$. Однако, если компоненты $X_1,\ \ldots,\ X_n$ независимы, т. е.

$$P(x_1,...,x_n) = \prod_{i=1}^{n} P(x_i),$$

то каждая вероятность $P(x_i)$ вычисляется с помощью относительной частоты $X_i = x_i$, поэтому для определения $P(x)$ требуется провести только $n \cdot r$ вычислений. Поскольку предположение о независимости часто не выполняется, то целесообразно исследовать модели, для которых требуется умеренное число вычислений и хранений данных. Вероятностное распределение для зависимых переменных в виде дерева является одной из таких моделей.

Вероятностное распределение дерева $P^t(x)$ записывается в виде произведения $n-1$ условных вероятностей

$$P^t(x) = \prod_{i=1}^{n} P\left(x_i \middle| x_{j(i)}\right), \qquad (1)$$

где $X_{j(i)}$ – родительская переменная для X_i при некоторой ориентации дерева. Корень X_1 не имеет родителей, $P\left(x_1 \middle| x_0\right) = P\left(x_1\right)$. Поэтому число параметров, необходимых для определения вероятностного распределения дерева, равно $(r-1)r(n-1)+(r-1)$, а именно $r(r-1)$ параметров для каждой из $(n-1)$ матриц условных (переходных) вероятностей и $(r-1)$ параметров для корня дерева.

Таким образом, байесовская процедура для сети – дерево строится аналогично процедуре распознавания для цепи Маркова [1, 5]. Для ее построения используются оценки переходных вероятностей

$$\hat{P}\left(x_i \middle| x_j,\ f = l\right) = \frac{k\left(x_j, x_i, l\right)}{k\left(x_j, l\right)}, \qquad (2)$$

где $f = l$ – состояние объекта, $k\left(x_j, l\right)$, $k\left(x_j, x_i, l\right)$ – соответственно число наблюдений значений x_j переменной X_j (пар $\left(x_j, x_i\right)$) в обучающей выборке V_l, $l \in \{1,...,h\}$.

Исследование эффективности байесовской процедуры распознавания на цепях Маркова основано на исследовании свойств оценок переходных вероятностей. В отличие от независимых бернуллиевых величин математическое ожидание оценок переходных вероятностей, построенных в виде частот, смещено и не совпадает с точными значениями вероятностей. Т. Андерсон и Л. Гудмен показали, что оценки переходных вероятностей асимптотически нормальны и вывели формулы дисперсии и ковариации оценок для этого предельного распределения [1, 6]. На основе этих результатов для цепей Маркова получены асимптотические оценки погрешности байесовской процедуры распознавания, которые аналогичны оценкам для независимых признаков [1, 5].

Байесовская процедура распознавания состояния l объекта $\mathbf{x} = (x_1,..., x_n)$, $l \in \{1,..., h\}$ строится на основе формулы Байеса

$$P(f = l | x_1,..., x_n) = \frac{P(x_1,..., x_n | f = l) P(f = l)}{P(x_1,..., x_n)}. \qquad (3)$$

Оценки вероятностей в числителе (3) строятся с помощью соотношений (1), (2) и обучающих выборок описаний объектов $\mathbf{x} = (x_1,..., x_n)$, имеющих заданные состояния. Знаменатель в (3) не используется, поскольку байесовская процедура определяет такое состояние наблюдаемого объекта $\mathbf{d} = (x_1 = d_1,..., x_n = d_n)$, которое имеет максимальную оценку в (3).

Структура сети – дерево выводится на основе формулы взаимной информации между парами переменных

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \ln \frac{P(x_i, x_j)}{P((x_i) P(x_j)} \geq 0. \qquad (4)$$

С помощью (4) вычисляют веса всех $n(n-1)/2$ ребер и упорядочивают их по величине. Алгоритм построения структуры – дерево выполняется на основе следующих шагов.

1. Выбирают два ребра, имеющих наибольшие веса.

2. Проверяют следующее по порядку ребро и добавляют его в дерево, если при этом не образуется петля. В противном случае ребро исключается и выбирается следующее по порядку величины веса ребро.

3. Повторяют шаг 1 до тех пор, пока не будут выбраны $n-1$ ребер.

Алгоритм вывода структуры сети – дерево в [4] был обобщен на схему байесовской сети, в которой вершина может иметь нескольких родителей и между двумя вершинами существует единственный путь. Такую сеть называют лесом (polytree), так как ее можно рассматривать как объединение нескольких деревьев, сливающихся между собой через вершины, имеющие связи вида $X \to Y \leftarrow Z$.

Вероятностное распределение сети – лес определяется соотношениями

$$P(x_1,...,x_n)=\prod_{i=1}^{n}P\left(x_i\left|x_{j_1(i)},x_{j_2(i)},...,x_{j_m(i)}\right.\right), \tag{5}$$

где $\left\{x_{j_1(i)},x_{j_2(i)},...,x_{j_m(i)}\right\}$ – множество (возможно пустое) прямых родителей переменной x_i, причем родители каждой переменной взаимно независимы, т. е.

$$P\left(x_{j_1(i)},x_{j_2(i)},...,x_{j_m(i)}\right)=\prod_{k=1}^{m}P\left(x_{j_k(i)}\right). \tag{6}$$

Структура сети – лес в отличие от структуры – дерево не всегда восстанавливается однозначно из эмпирических данных, поскольку в ней могут присутствовать три вида триплетов соседей:

1. $X\to Y\to Z$
2. $X\leftarrow Y\to Z$ (7)
3. $X\to Y\leftarrow Z$.

Триплеты 1 и 2 представляют собой один вид зависимости между переменными модели, поэтому они не различимы в сети. Наоборот, триплет 3 определяется однозначно, так как переменные X и Z независимы, пары $\{X,Y\}$, $\{Y,Z\}$ являются зависимыми. Поэтому скелет этих трех триплетов одинаковый, а направления дуг можно определить лишь частично.

Структура зависимостей всех потомков заданной вершины X в сети – лес такая же, как и в случае простой сети – дерево, и вершины, содержащие родителей X, независимы друг от друга. Поэтому алгоритм аналогичный сети – дерево способен определить скелет структуры сети – лес. Если скелет структуры задан, то зависимость вида 3 в (7) дает возможность для переменной, которая имеет нескольких родителей (коль скоро определены первые два родителя), определить направление всех ребер родитель – потомок. В частности, заметим, что частично ориентированный триплет $X\to Y-Z$ может быть завершен с помощью тестирования взаимной независимости переменных X и Z. Если X и Z независимы, то Z – родитель переменной Y, в противном случае Y – родитель переменной Z.

В терминах количества информации для некоторой пары переменных $\left(X_i,X_j\right)$, которые не имеют общих потомков, выполняются соотношения

$$I(X_i,X_j)>0, \tag{8}$$

и если X_k не блокирует путь между переменными X_i и X_j, то

$$I(X_i,X_j\left|X_k\right.)>0, \tag{9}$$

где

$$I(X_i,X_j\left|X_k\right.)=\sum_{x_i,x_j,x_k}P(x_i,x_j,x_k)\ln\frac{P(x_i,x_j\left|x_k\right.)}{P(x_i\left|x_k\right.)P(x_j\left|x_k\right.)}. \tag{10}$$

Дополнительно к этому родители любой переменной взаимно независимы, поэтому

$$I(X_{j_1(i)}, X_{j_2(i)}) = 0, \tag{11}$$

и если X_k блокирует путь между переменными X_i и X_j, то из (5)

$$I(X_i, X_j | X_k) = 0. \tag{12}$$

С помощью неравенств (8), (9) и соотношений (11), (12) можно провести точное восстановление скелета структуры и частично направления дуг. Алгоритм построения структуры – лес состоит из следующих шагов.

1. Выполняются шаги 1 – 3 описанной выше процедуры для структуры сети – дерево.

2. Находят внутреннюю вершину скелета, начиная поиск с наиболее удаленных слоев, до тех пор, пока не будет найдена вершина со многими родителями, используя правило тестирования триплета $X \to Y \leftarrow Z$.

3. Для вершины со многими родителями определяют направления всех ее ребер, используя тест триплета $X \to Y \leftarrow Z$.

4. Для каждой вершины, которая имеет, по крайней мере, одну входящую дугу, определяют направления всех ее соседних ребер, используя $X \to Y \leftarrow Z$ тест.

5. Повторяют шаги 2 – 4, пока не будут проставлены, по возможности, все направления ребер.

6. Ребра, для которых не удалось определить направления, помечаются как «неопределенные». Для завершения процесса определения направлений оставшихся ребер необходимо привлекать ряд дополнительных соображений.

Выбор направлений ребер проводится путем решения серии задач распознавания на тестовых примерах. Останавливаются на такой структуре, которая обладает наилучшими результатами тестирования.

В байесовской процедуре распознавания на структуре сети – лес используются оценки переходных вероятностей, построенные в виде частот, для цепей Маркова определенных порядков. Если переменная имеет нескольких родителей, то при вычислении вероятности (5) применяются оценки переходных вероятностей заданных порядков, построеннные аналогично (2).

Заключение. В статье приведены полиномиальные алгоритмы определения структуры байесовской сети, имеющей вид дерева или леса (polytree). Для заданной структуры байесовской сети рассмотрены байесовские процедуры распознавания, построенные на обучающих выборках и использующие оценки переходных вероятностей заданных порядков.

О.А. Вагіс

ПРОЦЕДУРИ РОЗПІЗНАВАННЯ НА БАЙЄСІВСЬКИХ МЕРЕЖАХ

Описано поліноміальні алгоритми визначення структури байєсівської мережі у вигляді дерева або лісу. Для відомої структури байєсівської мережі розглянуто байєсівські процедури розпізнавання, які побудовано на навчаючих вибірках за допомогою оцінок перехідних ймовірностей.

A.A. Vagis

RECOGNITION PROCEDURES ON BAYESIAN NETWORKS

The polynomial algorithms of determination of Bayesian network structure are described as a tree or polytree. For the known Bayesian networks structure, Bayesian recognition procedures built on teaching samples with the use of estimations of transitional probabilities are considered.

1. *Гупал А.М., Сергиенко И.В*. Оптимальные процедуры распознавания. – Киев: Наук. думка, 2008. – 232 с.
2. *Балабанов А.С.* Минимальные сепараторы в структурах зависимостей. Свойства и идентификация // Кибернетика и системный анализ. – 2008. – № 6. – С. 17 – 32.
3. *Балабанов А.С.* Формирование минимальных *d*-сепараторов в системе зависимостей // Кибернетика и системный анализ. – 2009. – № 5. – С. 38 – 50.
4. *Pearl J*. Probabilistic reasoning in intelligent systems: networks of plausible inference. – San Mateo: Morgan Kaufmann, 1988. –552 p.
5. *Гупал А.М., Вагис А.А.* Статистическое оценивание марковской процедуры распознавания // Проблемы управления и информатики. – 2001. – № 2 – С. 62–71.
6. *Anderson T.W., Goodman L.A.* Statistical inference about Markov Chains // The Annals of Mathematical Statistics. – 1957. – 28. – P. 89–110.

Получено 22.12.2009

Об авторе:

Вагис Александра Анатольевна,
кандидат физико-математических наук, старший научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины.