1Т

УДК 004.89

В.Я. Терзиян
Харьковский национальный университет радиоэлектроники, Украина,
Университет города Ювяскюля, Финляндия
А.В. Витько
Харьковский национальный университет, Украина

Вероятностные метасети для решения задач интеллектуального анализа данных

В статье предлагается многоуровневая вероятностная модель – Байесовская метасеть – для решения задач интеллектуального анализа данных. Каждый следующий уровень метасети используется для выбора определенной подструктуры из предыдущего уровня сети на основании контекстных признаков. На базе общей модели метасети были разработаны несколько ее интерпретаций и приложений. В статье рассматривается использование вероятностных метасетей для таких важных задач, как фильтрация web-содержимого и предсказание предпочтений мобильных пользователей.

Введение

Одной из наиболее острых проблем, которые возникли на сегодняшний день в информационных технологиях, является проблема нарастающего с огромной скоростью объема хранимых данных. Основное решение этой проблемы заключается в проведении анализа данных для того, чтобы выявить структуру данных, внутренние зависимости и закономерности, основные характеристики. Полученная после анализа данных информация может заменить весь огромный объем соответствующих данных и существенно сократить время работы с этими данными.

В связи с появлением новых типов данных и новых задач обработки данных стандартного статистического анализа стало недостаточно. Появилась необходимость в интеллектуальном анализе данных, который включает интеллектуальные методы получения информации из данных. В частности, интеллектуальный анализ данных включает в себя разработку принципов выбора определенной модели данных, и принципов соответствия модели данным. Для современных задач эти методы должны быть построены с использованием искусственного интеллекта.

Стандартный статистический подход игнорирует неопределенность выбора самой модели данных. Это зачастую ведет к рискованным оценкам, основанным только на одной модели. А с учетом того, что неопределенность выбора модели вносит свою долю риска, решением представляется использование нескольких вероятностных моделей.

В качестве одного из вариантов интеграции нескольких вероятностных моделей нами предлагается построение многоуровневой вероятностной модели –

метасети Байеса, в которой каждый следующий уровень сети определяет вероятности предыдущего уровня.

Эффективное решение этой задачи особенно необходимо в таких задачах, как анализ данных в системах электронной коммерции, фильтрация содержимого сети Интернет, построение адаптивных интеллектуальных интерфейсов, то есть в задачах, в которых процесс принятия решений зависит от множества контекстных признаков, например личных предпочтений пользователя.

В разделе 1 анализируется использование вероятностных сетей для интеллектуального анализа данных. В разделе 2 мы описываем модель Байесовской метасети и даем несколько возможных интерпретаций этой модели. Раздел 3 содержит анализ возможных приложений для современных задач прогнозирования. Выводы приведены в последнем разделе.

1. Использование вероятностных (Байесовских) сетей для интеллектуального анализа данных

Вероятностная (Байесовская) сеть – это графическая модель для описания вероятностных отношений среди набора переменных [1]. В настоящее время существует множество способов представления данных для их последующего интеллектуального анализа (базы правил, деревья решений, искусственные нейронные сети, др.), а также множество методов такого анализа (классификация, регрессия, кластеризация, др.). Байесовские сети и байесовские методы интеллектуального анализа имеют ряд преимуществ:

- могут с легкостью обрабатывать неполные наборы данных;
- позволяют обучаться причинно-следственным отношениям. Изучение причинно-следственных отношений полезно, по крайней мере, по двум причинам. Это полезно для изучения проблемной области, кроме того, знание причинно-следственных отношений позволяет нам делать предсказания в присутствии помех;
- Байесовские сети совместно с байесовскими статистическими методами облегчают объединение априорных знаний о проблемной области и данных;
- байесовские методы вместе с Байесовскими сетями и другими типами моделей предлагают эффективный и принципиальный уход от излишней подгонки моделей под данные обучающей выборки.

Байесовская сеть для набора переменных $\mathbf{X} = \{X_1, \ldots, Xn\}$ – это ациклический направленный граф с сетевой структурой S, которая кодирует набор условных утверждений независимости переменных в **X** и набор локальных вероятностных распределений P, связанных с каждой переменной [1]. Оба этих компонента определяют совместное распределение вероятностей для **X**, которое определяется через условные вероятности переменных X_i при заданных значениях родительских вершин и задается формулой (1)

$$P(X) = \prod_{i=1}^{n} P(X_i \mid родители_i) \qquad (1)$$

Поскольку Байесовская сеть для **X** определяет совместное вероятностное распределение для **X**, мы можем, в принципе, вычислить любую представляющую интерес вероятность. Базовый алгоритм вероятностного вывода для ограниченного числа дискретных переменных был разработан, проанализирован и экспериментально протестирован в [2]. Он изменяет направление зависимостей (обращает дуги) в сетевой структуре с применением теоремы Байеса до тех пор, пока необходимая вероятность не может быть вычислена непосредственно из сети по формуле (1). В общем случае вероятностный вывод в произвольной Байесовской сети является NP-сложным [3].

Одной из наиболее интересных задач в интеллектуальном анализе данных является задача обучения вероятностной сети с неизвестной структурой. Если знаний о предметной области недостаточно, а также в случае, когда недоступно смысловое значение одной или нескольких переменных, эксперт не может точно оценить структуру вероятностной сети. Алгоритмическое нахождение структуры сводится к поиску в пространстве возможных структур [4]. Для решения проблемы используются два метода: выбор одной наилучшей модели и выборочное усреднение по нескольким адекватным моделям [5].

Рассматриваемые модели Байесовской сети и методы обучения направлены на случай, когда переменные в проблемной области так или иначе являются «однородными», часть из них влияет на значения других переменных, но сами условные зависимости остаются неизменными. В то же время существует ряд задач, в которых контекст ситуации влияет на условные зависимости между переменными в проблемной области. Примеры таких задач рассмотрены в разделе 3. Для решения таких задач нами предлагается модель многоуровневой вероятностной метасети.

Динамические вероятностные сети рассматриваются в качестве моделей в областях, где переменные принимают различные значения со временем [6]. Динамические вероятностные сети моделируются как скрытые марковские процессы, они кодируют стохастическое изменение состояния сети во времени. Таким образом, они существенно отличаются от предлагаемой модели Байесовских метасетей.

2. Определение и возможные интерпретации многоуровневой Байесовской метасети

Определим Байесовскую метасеть аналогично определению семантической метасети [7], [8].

Определение. Байесовская вероятностная метасеть – это набор Байесовских сетей, которые накладываются друг на друга таким способом, при котором элементы (вершины или дуги) каждой предыдущей вероятностной сети зависят от локальных распределений вероятностей, связанных с узлами следующего уровня сети.

Введение такого определения дает возможность использования нескольких интерпретаций вероятностной метасети.

Первая интерпретация Байесовской метасети – моделирование условных зависимостей. Рассмотрим Байесовскую метасеть с 2 уровнями (рис. 1)

для задачи прогнозирования. Контекстные переменные могут рассматриваться как верхний уровень управления по отношению к уровню сети с прогнозирующими переменными.

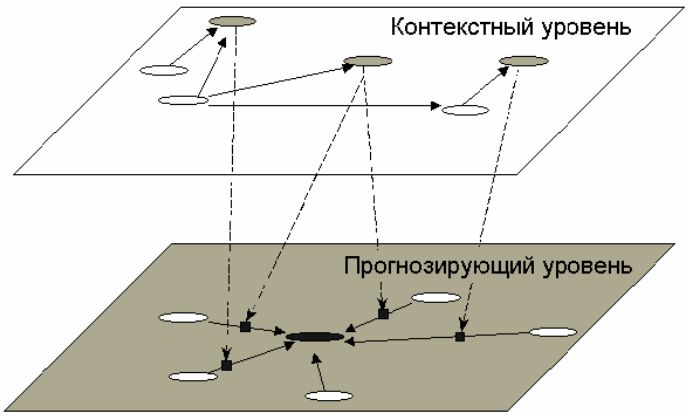

Рис. 1. Двухуровневая Байесовская метасеть для моделирования условных зависимостей

Каждый уровень метасети – это вероятностная сеть, содержащая вершины и дуги. Каждая дуга в Байесовской сети одного уровня соответствует условной зависимости между двумя переменными (вершинами), как показано на рис. 2.

Стандартный байесовский вывод применяется в Байесовской сети каждого уровня. Например, P(Y) вычисляется следующим образом:

$$P(Y) = P(X) \times P(Y \mid X). \qquad (2)$$

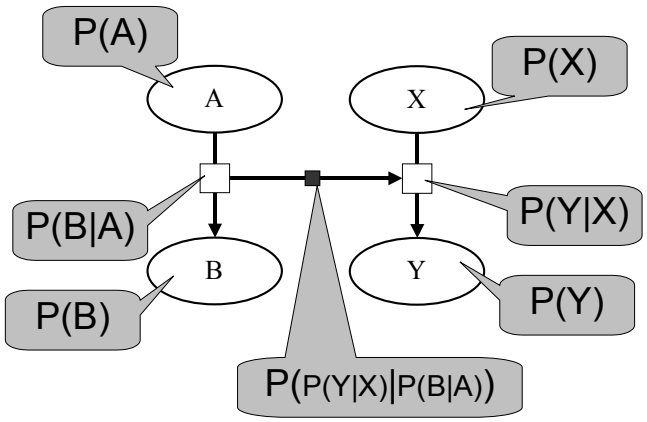

Рис. 2. Условные и безусловные вероятности в Байесовской метасети

Узлы 2-го уровня сети соответствуют условным вероятностям 1-го уровня сети P(B|A) и P(Y|X). Дуга на 2-ом уровне сети соответствует условной вероятности P(P(Y|X)|P(B|A)). Логический вывод показан соотношением (3):

$$P(Y) = P(X) \times P(P(Y \mid X) \mid P(B \mid A)) \times P(B \mid A). \qquad (3)$$

Двухуровневая метасеть может быть легко расширена до многоуровневой (многоконтекстной) метасети [8]. Многоуровневое представление контекста позволяет решать следующие проблемы:

- получать интерпретируемое знание, используя все известные уровни его контекста;
- получать неизвестное знание, когда интерпретация его в некотором контексте и сам контекст известны;
- получать неизвестное знание относительно контекста, когда известно, как знание интерпретируется в этом контексте;
- преобразовать знание из одного контекста в другой;
- в пределах любой проблемы получать зависимости при рассмотрении нескольких контекстов и использовать такие зависимости, чтобы вычислить более точные решения проблемы.

В принципе, мы можем допустить, что Байесовская метасеть может иметь столько уровней, сколько необходимо. На рис. 3 приведен пример трехуровневой Байесовской метасети.

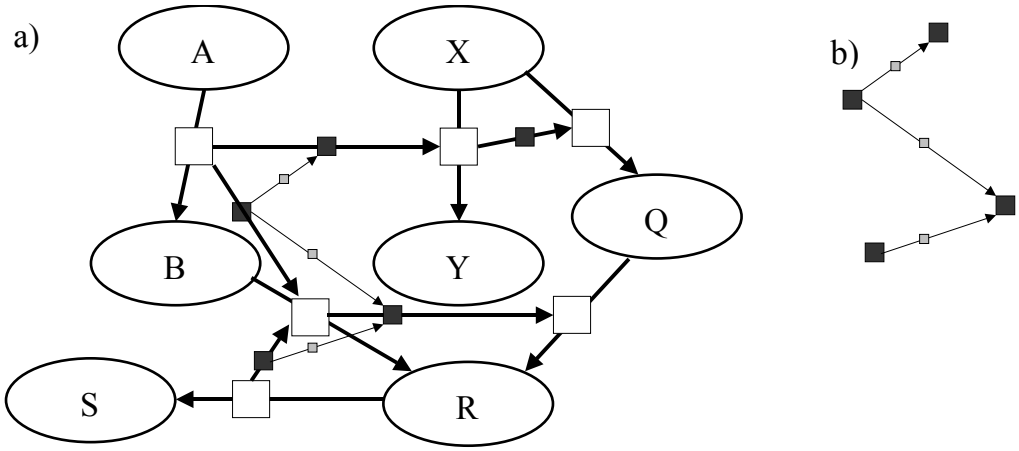

Рис. 3. Пример трехуровневой Байесовской метасети. Третий уровень (b) управляет условными зависимостями второго уровня сети, которая управляет условными зависимости первого уровня (a).

Вторая интерпретация Байесовской метасети – моделирование отбора соответствующих атрибутов для прикладных целей. Методы отбора атрибутов извлекают подмножество атрибутов, которые влияют на целевой концепт. Сила и слабость каждого из этих методов базируется на характеристиках домена и типах данных. Как известно, нет единого метода отбора атрибутов, который может применяться во всех прикладных областях. Выбор метода отбора атрибутов зависит от различных характеристик набора данных.

Байесовская метасеть может быть инструментом для моделирования отбора релевантных атрибутов. Переменные контекста рассматриваются тоже как верхний уровень управления по отношению к уровню сети с прогнозирующими переменными. Значения переменных контекста влияют на релевантность переменных в прогнозирующей модели, как показано на рис. 4.

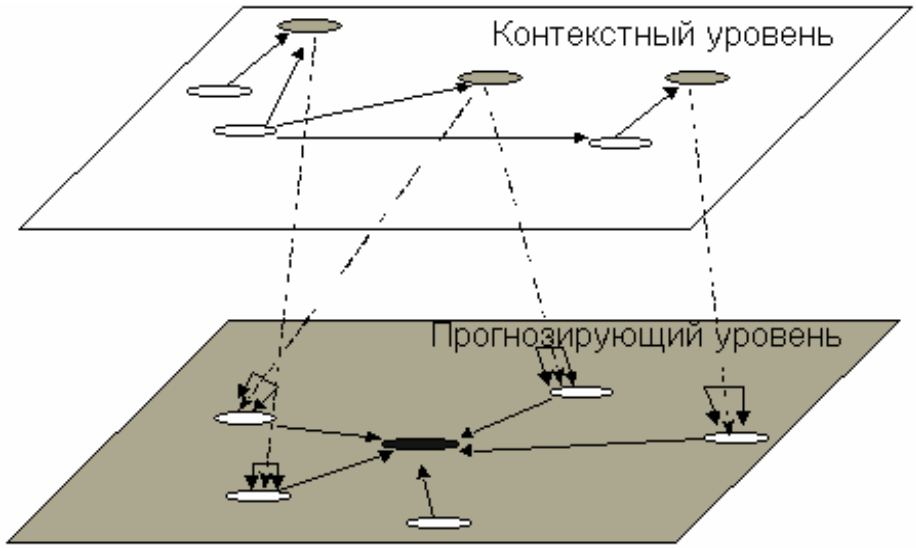

Рис. 4. Двухуровневая Байесовская метасеть для моделирования выбора релевантных атрибутов

Мы рассматриваем релевантность как вероятность важности переменной для вывода целевого атрибута в данном контексте. При таком определении релевантность наследует все свойства вероятности, как показано на рис. 5. Условная релевантность Ψ(Y|X) – это есть релевантность Y при заданной релевантности X. Стандартный байесовский вывод применяется к релевантности, как показано в соотношении (4).

$$P(Y) = \psi(Y) \times P(X) \times P(Y \mid X) = \psi(X) \times \psi(Y \mid X) \times P(X) \times P(Y \mid X) \qquad (4)$$

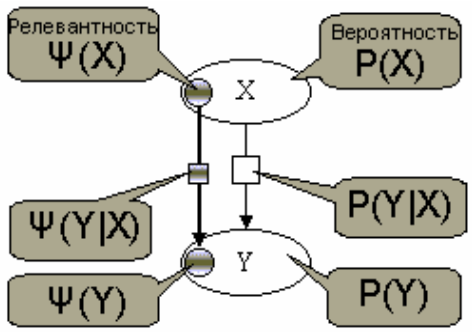

Рис. 5. Определение релевантности подобно определению вероятности

Сеть релевантности определена над заданной прогнозирующей вероятностной сетью (рис. 6). Она кодирует условные зависимости между релевантностями. Сеть релевантности содержит первичные релевантности и условные релевантности. При рассмотрении такого определения сети релевантности над прогнозирующей сетью очевидно, что между узлами обеих сетей существует строгое соответствие, но дуги не обязательно должны соответствовать друг другу (как показано на рис. 6).

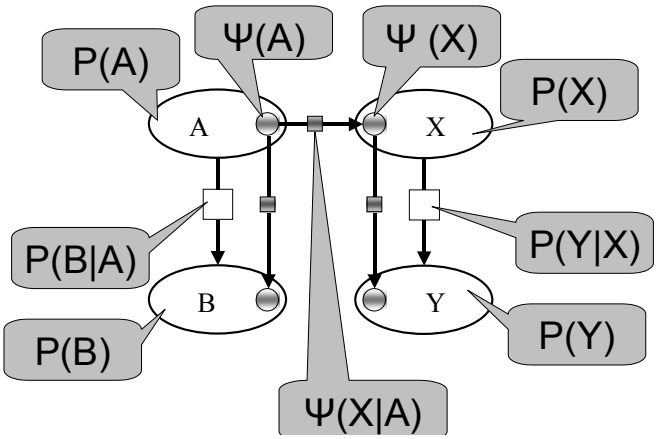

Рис. 6. Сеть релевантности определена над прогнозирующей сетью. Выход целевого атрибута показан соотношением (5)

Это означает, что релевантности двух переменных могут быть зависимы, хотя их значения условно независимы и наоборот (рис. 7). Топология этих сетей в общем случае отлична.

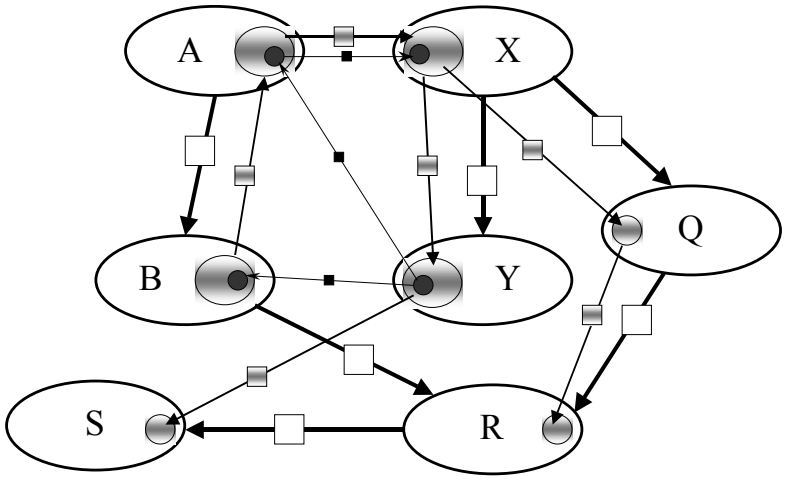

Рис. 7. Пример трехуровневой Байесовской метасети. Третий уровень управляет релевантными атрибутами на втором уровне, который управляет релевантными атрибутами на первом уровне

$$P(Y) = \psi(A) \times \psi(X \mid A) \times \psi(Y \mid X) \times P(X) \times P(Y \mid X). \quad (5)$$

Связь между вероятностью и релевантностью может быть определена в соответствии с правилом:

$$P(X) = \begin{cases} 0, if \psi(X) < \theta; \\ P(X) \times \psi(X), if \psi(X) \geq \theta, \end{cases} \quad (6)$$

где θ – порог релевантности. Порог релевантности выбирается согласно формулировке задачи и существующему контексту.

Мы можем использовать более простое правило типа (7):

$$P(X) = \begin{cases} 0, if \psi(X) < \theta; \\ P(X), if \psi(X) \geq \theta. \end{cases} \quad (7)$$

Сеть релевантности и прогнозирующая сеть – два уровня структуры метасети. В принципе, мы можем предполагать, что Байесовская метасеть, моделирующая релевантность атрибутов, может иметь столько уровней, сколько необходимо. На рис. 7 Байесовская метасеть состоит из 3 уровней.

Третья интерпретация Байесовской метасети – комбинированная метасеть первого и второго видов. В общем случае оба из уровней управления, введенные выше, могут существовать в проблемной области.

Все свойства входящих сетей те же самые, как мы их описали выше. Фактически в комбинированной метасети два уровня управления воздействуют на основной уровень (рис. 8).

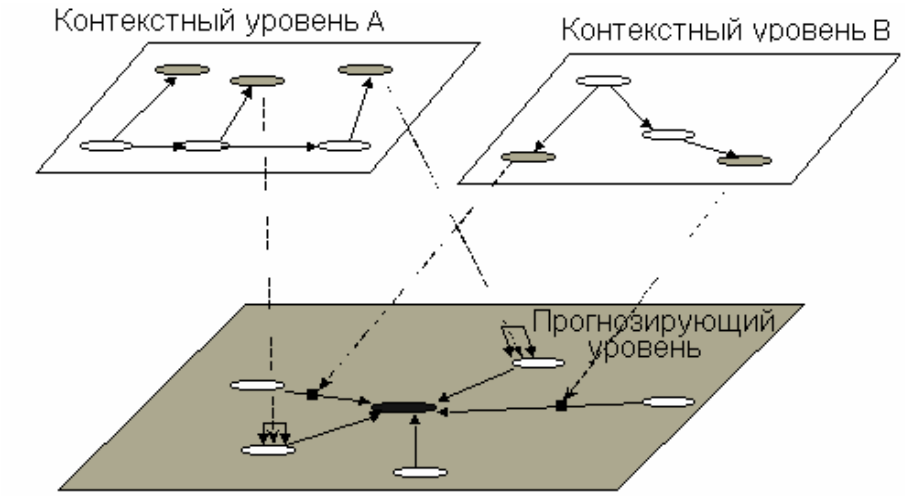

Рис. 8. Комбинированная Байесовская метасеть с двумя уровнями управления

3. Применение вероятностных метасетей

Как уже было замечено, применение предлагаемой модели (в различных ее интерпретациях) эффективно для задач анализа данных в проблемных областях, где контекст ситуации влияет на условные зависимости между переменными в проблемной области. Авторами был проведен анализ таких задач и выделены следующие актуальные задачи в области построения интеллектуальных информационных систем, направленных на пользователя:

– фильтрация содержимого сети Интернет для конкретного пользователя;
– моделирование пользовательских предпочтений;
– предоставление пользователю информационных и др. видов услуг.

Наибольший экономический эффект решение таких задач принесет в мобильной информационной среде (мобильном Интернет) для пользователей

мобильных телефонов, имеющих выход в глобальную сеть. Достижения в беспроводных сетевых технологиях и непрерывно увеличивающееся число пользователей мобильных телефонов делает мобильный терминал возможным каналом для предложения индивидуализированных услуг мобильным пользователям и создает возможность быстрой разработки мобильной электронной коммерции [9]. Скорость и точность предоставления вышеуказанных услуг еще более актуальны из-за высокой стоимости беспроводного сетевого трафика, да и размеры терминала ограничены по сравнению с доступом в Интернет через персональные компьютеры, что требует тщательной фильтрации информации.

Одна из наиболее важных особенностей мобильной среды – мобильность. Мобильные сетевые серверы и даже мобильные терминалы в состоянии теперь определить свои координаты с высокой точностью. Это дает возможность предоставления пользователям оперативных и локализованных услуг [10].

Цель предоставления локализованных услуг – снабдить пользователя информацией о необходимых ему объектах, принимая во внимание пространственное расстояние между ним и объектами. Расположение пользователя в данном случае является существенным контекстным атрибутом, отличающимся от атрибутов потребности и желания, и такой атрибут желательно вынести в отдельный уровень вероятностной метасети – управляющий. Схема многоуровневой структуры профайлов пользователя была предложена в [11].

В последние годы было разработано множество методов фильтрации [12]. Фильтрация содержимого сети и адаптация под пользователя достаточно актуальны и в привычной Интернет-среде в настоящую эпоху переполнения ее информацией. Информационные системы требуют такого пользовательского интерфейса, который может «разумно» определять предпочтения пользователя и использовать их для предоставления необходимой информации.

В общем случае задача моделирования предпочтений пользователей и фильтрации информации имеет набор переменных (атрибутов), которые влияют на выбор или предпочтение (целевой атрибут) пользователя определенным образом. Вероятностные связи между этими атрибутами кодируются в первом, прогнозирующем уровне Байесовской метасети. Кроме того, на механизм принятия решения, то есть зависимости между атрибутами первого уровня, оказывает влияние контекст ситуации (настроение, самочувствие, присутствие других людей рядом, пространственное расположение, время дня и т.д.). В свою очередь связи между контекстными атрибутами также могут быть вероятностными, и их предлагается кодировать вторым, управляющим уровнем Байесовской метасети.

Заключение

В работе предложен формализм Байесовской метасети как модели интеграции нескольких вероятностных сетей для решения задач интеллектуального анализа данных. В основе такой модели лежит стандартная вероятностная сеть. Каждый дополнительный уровень в метасети используется, чтобы выбрать соответствующие контексту зависимости атрибутов базовой сети или соответствующую подструктуру из базового уровня сети. Предложены три

интерпретации вероятностной метасети. Общий формализм предложенной Байесовской метасети можно использовать для решения нескольких задач моделирования: моделирование условных зависимостей в профайле пользователя, моделирование выбора релевантных атрибутов и комбинированное моделирование.

Наиболее актуальными областями применения предложенной модели вероятностной метасети являются фильтрация содержимого сети Интернет для конкретного пользователя и моделирование пользовательских предпочтений в обычных и мобильных распределенных информационных системах. Ожидается, что использование Байесовских метасетей принесет наибольший экономический эффект в мобильной электронной коммерции.

Литература

1. Heckerman D. A Tutorial on Learning with Bayesian Networks // Technical Report MSR-TR-95-06. – Microsoft Corporation, Redmont, WA. – 1995.
2. Витько А.В. Базовый алгоритм вероятностного вывода на Байесовых сетях // Вестник ХГПУ. Сер. Новые решения в современных технологиях. – 1999. – Вып. 75. – С. 23-26.
3. Cooper G. Computational complexity of probabilistic inference using Bayesian belief networks // Artificial Intelligence. – 1990. – № 42. – P. 393-405.
4. Витько А.В. Проблемы обучения сетей Байеса с неизвестной структурой // Труды 4-го Междунар. молодеж. форума «Радиоэлектроника и молодежь в XXI веке». – Харьков: ХТУРЭ. – 2000. – Ч. 2. – С. 223-224.
5. Heckerman D., Meek C. Models and Selection Criteria for Regression and Classification // Technical Report MSR-TR-97-08. – Microsoft Corporation, Redmont, WA. – 1997.
6. Kanazawa K., Koller D., Russel S.J. Stohastic simulation algorithms for dynamic probabilistic networks // Uncertainty in Artificial Intelligence: Proc. of the 11th Conference. – San Francisco: Morgan Kaufmann Publishers, 1995. – P. 346-351.
7. Терзиян В.Я. Многоуровневые модели управления базами знаний и их приложения для автоматизированных информационных систем: Дис… доктора техн. наук. – Харьков: ХТУРЭ, 1993.
8. Terziyan V., Puuronen S. Reasoning with Multilevel Contexts in Semantic Metanetworks // Formal Aspects in Context. – Kluwer Academic Publishers, 2000. – P. 107-126.
9. The MeT Initiative - Enabling Mobile E-Commerce // Met Overview White Paper. – 2000, October, 2 // http: //www.mobiletransaction.org/pdf/MeT_White_Paper.pdf.
10. Chadha K. Location-Based Services: The Next Differentiator, Mobile Internet & Inf. Services, San Diego, 2000, March, 28-30 //http://www.the-arc-roup.com/ebrief/2000/mobileinternetis/executive_summary.htm.
11. Терзиян В.Я, Витько А.В. Интеллектуальное управление информацией в мобильной электронной коммерции // Новости искусственного интеллекта. – 2001. – № 5-6.
12. Kutschinski E., Poutre H.L. Scientific techniques for interactive profiling // Technical Report (ASTA project). – Enschede: Telematica Instituut, 2001.

In this paper the multilevel probabilistic model – Bayesian metanetwork – is proposed for the tasks of intelligent data analysis. The extra level in a metanetwork is used to select appropriate substructure from the basic network level based on contextual features. Based on the metanetwork model we consider several possible interpretations and implementations. The use of probabilistic metanetworks for such important tasks as filtering of the Web content and prediction of mobile user's preferences is discussed.

Статья поступила в редакцию 01.07.02.