Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 519.2

Ярошевич
Юрий Александрович

Обучение параметров и вывод статистических суждений в вероятностных сетях

АВТОРЕФЕРАТ

на соискание степени магистра информатики и вычислительной техники по специальности 1-40-80-04 «Математическое моделирование, численные методы и комплексы программ»

Научный руководитель
Ганжа Виктор Александрович
кандидат физико-математических наук, доцент

Минск 2015

ВВЕДЕНИЕ

В век информационных технологий появляются огромные массивы данных, которые можно и нужно уметь обрабатывать с помощью вычислительной техники с целью извлечения знаний. Статистическое моделирование и интеллектуальный анализ данных представляют необходимые инструменты и способы обработки и анализа больших объемов данных. В данной работе рассматривается один из способов информационно-статистического моделирования — вероятностные сети, в частности, байесовы сети доверия.

Байесовы сети применяются для решения различных практических задач. Вероятностная природа сетей способствует их успешному применению для создания различных экспертных систем. Важным этапом в применении вероятностных сетей для решения какой-либо задачи является её построение: задание отношений независимости между парами вершин и параметров условных распределений. Применение же байесовых сетей обычно заключается в статистическом выводе различных суждений.

Одним из первых способов построения байесовых сетей было привлечение экспертов в предметной области и разработка архитектуры сети в соответствии с представлением экспертов о решаемой задаче. Появляются новые предметные области и классы задач, в которых довольно сложно найти признанного эксперта. В такой привлечение эксперта для разработки байесовой сети не всегда возможно и расточительно по времени. С другой стороны, сбор экспериментальных данных для решения какой-либо задачи обычно легко доступен. В связи с этим возникает задача обработки этих данных для построения структуры сети и нахождения параметров распределения для каждой вершины сети.

В данной работе исследуются и реализуются некоторые алгоритмы автоматического нахождения параметров распределения для вероятностной сети для случая полных и неполных данных. Также были рассмотрены вопросы статистического вывода в обученной байесовой сети, в разработанном программном обеспечении реализованы алгоритмы для статистического вывода.

В результате была разработана библиотека классов для платформы Microsoft .NET, написанная на языках программирования C# и F#, пригодная для решения практических задач в коммерческих проектах. На данный момент в библиотеке реализованы алгоритмы построения структуры сети по данным, нахождения параметров распределения по полному и неполному набору данных и известной топологии сети, а также реализованы функции для статистического вывода суждений для наиболее частых запросов к вероятностной сети.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Основной целью работы является исследование способов автоматического вывода параметров распределения для байесовых сетей по полному и неполному набору данных, проведения автоматического статистического вывода, в обученной сети, и, как итог работы, разработка инструмента для платформы Microsoft .NET для проведения вероятностного моделирования при помощи байесовых сетей и пригодного для использования в реальных проектах. Существуют различные бесплатные графические редакторы байесовых сетей, с помощью которых можно проводить моделирование в ручном режиме, но эти редакторы сложно встраивать и использовать в других проектах. Также есть библиотеки для работы с байесовыми сетями, которые можно бы было использовать в проектах, но для платформы Microsoft .NET такой доступной библиотеки нет, поэтому разработка такой библиотеки может быть востребована.

Разработанное программное обеспечение представляет из себя библиотеку для платформы Microsoft .NET. Библиотека поддерживает следующие возможности по работе с байеосвыми сетями: представление в памяти, сохранение и загрузка сетей, вывод структуры сети по данным (функциональность разработана вне рамок данной работы), нахождение параметров условных распределений для каждой из вершин сети по полному и не полному набору данных, а также функциональность проведения статистического вывода.

Для обучения параметров распределения по полным данным был исследован и реализован алгоритм использующий в качестве оценки функцию правдоподобия. При обучении по не полным данным использовался более сложных итеративный алгоритм expectation maximization, на каждой итерации которого сначала производиться оценка пропущенных данных, а затем пересчитываются текущие значения параметров, с учетом предугаданных пропусков. Функциональность статистического вывода реализована с помощью алгоритма последовательного устранения переменных. Данный алгоритм является точным методом вычисления вероятности в байесовых сетях.

Разработанная функциональность была протестирована на примере одной хорошо описанной в литературе байесовой сети. С помощью стороннего ПО (данная функциональность не реализована в разработанной библиотеке) были сгенерированы наборы данных для обучения. По этим наборам были обучены параметры распределения для сети, результаты обучения были сверены с параметрами исходной сети.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе произведён обзор предметной области задач, решаемых в рамках данной работы; рассмотрены вопросы о сущности байесовых сетей и принципе их работы; приведена оценка сложности различных проблем, возникающих при применении вероятностных сетей для решения прикладных задач. Рассмотрены принципы работы алгоритмов обучения параметров вероятностной сети. В рамках рассмотрения данного вопроса были показаны принципы обучения по полным и неполным данным. Также рассмотрен алгоритм точного статистического вывода в обученной сети. В главе также произведен обзор существующих общедоступных программ, решающих схожие задачи.

Во второй главе произведено обоснование выбора использованных для реализации технологий. Произведен краткий обзор программной платформы Microsoft .NET и ее возможностей. Кратко рассмотрен, использовавшийся при разработке, язык программирования C#. Более детально рассмотрен язык F# и его функциональные возможности, т. к., например, неизменяемые типы использовались при реализации библиотеке очень интенсивно.

Третья глава посвящена описанию деталей реализации разработанной библиотеки. В ней обосновывается выбор в пользу собственной реализации графов, вместо использования готовых решений, приводится описание разработанного представления для графов. В главе сформулированы требования к реализации представления байесовых сетей, такие как возможность версионирования структуры сети, возможность легкой отмены и повторения внесенных изменений в сеть, возможность расширения структуры сети дополнительными атрибутами, пригодными для увеличения количества сценариев применения, от «голой» структуры, до атрибутов графического представления. С учетом сформулированных требований был произведен выбор способа реализации. Также в главе описаны функции для импорта и экспорта байесовых сетей в существующие форматы представления, которые поддерживаются многими другими программами. Немаловажный вопрос описанный в данной главе — вопрос эффективного представления эксперементальных данных в оперативной памяти компьютера. Данный аспект влияет на эффективность реализации алгоритмов вывода структуры сети, а также алгоритмов обучения параметров сети по данным. Завершается глава разделами про тестирование реализации алгоритмов вывода параметров распределения по полным и неполным данным, а также тестированием алгоритма статистического вывода суждений. Тестирование показало приемлимую точность вычислений, как для обучения параметров, так и для статистического вывода.

ЗАКЛЮЧЕНИЕ

В работе были рассмотрены вопросы автоматического нахождения параметров распределения для вероятностной сети с известной структурой на основе экспериментальных данных. Также был исследован вопрос точного статистического вывода суждений на основе обученной байесовой сети. Была разработана библиотека типов и функций для представления, автоматического построения структуры сети, нахождения параметров распределения и статистического вывода суждений. Разработанная библиотека функций может быть использована при разработке коммерческих продуктов на платформе Microsoft .NET для реализации вероятностных моделей, созданных в более удобных графических редакторах вроде SAMIAM, Netica или GeNIe. Единственным и наиболее близким по функциональности продуктом со схожей областью применения для платформы Microsoft .NET является библиотека Infer.NET, разработанная в Microsoft Research Cambridge, но она, к сожалению, имеет крайне ограниченную лицензию, запрещающую коммерческое использование вне Microsoft.

В целом разработанная библиотека включает требуемую функциональность, необходимую для практического использования. В библиотеке присутствуют функции для решения всего цикла задач связанных с применением байесовых сетей: построение структуры сети по эксперементальным данных, обучение параметров распределение по полным и неполным данны, а также функции для статистического вывода в готовой сети. Помимо этого реализована необходимая функциональность по сохранению, загрузке и модификации сетей.

В результате цель работы была достигнута. Было создано программное обеспечение решающие, хотя и не всеми возможными способами, все типы задач, связанных с применением баейсовых сетей на практике. За рамками проделанной работы остались некоторые специфические вопросы, например, неточный вывод суждений, использование метода градиентного спуска на этапе оптимизации функции правдоподобия, для решения обучения параметров и другие менее распространённые задачи. Эти вопросы возникают не во всех практических задача, но при необходимости разработанная библиотека может быть доработана. Разработка дополнительных методов решения задачи статистического вывода и обучения параметров является возможным путем развития уже созданного ПО.