УДК 004.9

Т.В. Юр

Запорожский национальный технический университет, Запорожье

ОБЗОР ПРИМЕНЕНИЙ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ В ЗАДАЧАХ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

В последнее время все более широкое применение находит математический аппарат вейвлет-анализа в различных сферах науки и при решении практических задач. Данная работа посвящена обзору применения вейвлетов в методах интеллектуального анализа данных, т.е. методах обнаружениях в данных новых знаний. Для структурирования обзора весь процесс извлечения новых знаний разбит на последовательность этапов, для каждого из которых показаны возможности вейвлетов в данной области. Показаны потенциальные направления дальнейшего исследования и применения вейвлетов.

***Ключевые слова**: вейвлет, анализ данных, извлечение знаний, временной ряд.*

Введение

Вейвлет-преобразование является синтезом идей, возникших на протяжении многих лет в различных областях науки. Вейвлет-преобразование является средством многоуровневого разбиения данных на различные частотные компоненты и дальнейшее изучение этих компонент с разрешением, соответствующим масштабу [1, 2]. Вейвлет-преобразование позволяет получить компактное и информативное математическое представление многих исследуемых объектов. Сегодня многие пакеты математического программного обеспечения предоставляют быстрые и эффективные библиотеки, которые выполняют вейвлет-преобразование. Благодаря такому легкому доступу вейвлеты быстро получили популярность среди исследователей и инженеров как теоретических, так и практических задач.

Интеллектуальный анализ данных (ИАД) – это собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [3, 4]. Он с успехом находит свое применение в различных областях: в биоинформатике, фармацевтике, банковском деле, торговле, сферах развлечения и т.д. Множество важных задач науки и техники решаются при помощи таких методов ИАД, как нейронные сети, нечеткая логика, деревья решений, генетические алгоритмы, статистические методы [5].

Основное применение вейвлетов – это анализ данных с временными или пространственными особенностями (например, временные ряды, потоки данных, изображения). Однако в последнее время сферы применения теории вейвлет-анализа все больше расширяются.

Вейвлеты имеют множество благоприятных свойств, таких как: нулевые моменты, компактный носитель, иерархическая и многомасшабная структура декомпозиции, линейная временная и пространственная сложность преобразования, декорре-

ляция коэффициентов и большое многообразие базисных функций [1, 2].

Как правило, вейвлет-преобразования подразделяют на дискретное (ДВП) и непрерывное (НВП). Путем разложения данных на компоненты в базисе анализирующего вейвлетаВейвлет-преобразование позволяет представлять исходные данные с различной степенью детализации.

При этом дискретное вейвлет-преобразование позволяет получить сжатое представление данных и используется для преобразований и кодирования сигналов, а непрерывное вейвлет-преобразование обладает некоторой избыточностью представления исходных данных и поэтому применяется для более глубокого анализа особенностей сигналов и их обработки [1, 2].

Приведенные свойства вейвлет-преобразования позволяют получить значительно более эффективные решения многих проблем, решаемых в ИАД. Теория вейвлетов может сыграть достаточно важную роль в ИАД и может стать ядром многих методов.

Постановка задачи. Целью работы является обзор применения вейвлетов в методах, решающих задачи ИАД. Для структурирования обзора весь процесс извлечения новых знаний разбит на последовательность шагов, для каждого из которых показаны возможности вейвлетов в данной области. Показаны потенциальные направления дальнейшего исследования и применения вейвлетов.

Процесс интеллектуального анализа данных

Интеллектуальный анализ данных связан с отысканием шаблонов, трендов, ассоциаций, аномалий, важных атрибутов, структур и зависимостей в данных. Он является мультидисциплинарной областью, которая использует и усовершенствует идеи из различных областей знаний, таких как обработка сигналов и изображений, машинное обучение, распознавание образов, экспертные системы, оптимизация, статистика и другие. ИАД анализирует большие объемы сложных необработанных данных, помогая принимать на их основе решения [3 – 5].

Интеллектуальный анализ данных является итеративным процессом, который можно разбить на следующие этапы [4, 5]:

1) хранение и управление данными,

2) предварительная обработка и подготовка данных к анализу,

3) процессы непосредственного анализа данных и извлечения знаний,

4) оценка и интерпретация полученных результатов.

Последний из перечисленных этапов в основном касается нетехнической работы, такой как документация и оценка результатов аналитиком. Поэтому остановимся более подробно только на первых трех этапах.

Этап хранения и управления данными

На этапе хранения и управления данными задаются механизмы и структуры для организации доступа к данным. Данный этап является крайне важным в практических применениях, поскольку для решения задач ИАД могут использоваться огромные объемы информации, накопленной в течение длительного периода времени.

Цель этапа состоит в отыскании таких методов хранения данных, которые позволили бы обеспечить быстрый и эффективный доступ к данным больших объемов.

Благодаря тому, что вейвлет-преобразование предоставляет иерархическую многомасштабную структуру представления данных, он может быть применен для эффективного хранения сжатых данных [6, 7].

Вейвлет-преобразование широко используется для сжатия изображений, аудио и видео данных, временных сигналов. Вейвлет-преобразование позволяет получить высокое соотношение сжатия в сочетании с хорошим качеством восстановленного сигнала. Так вейвлет-преобразование было выбрано для стандартов сжатия изображений JPEG2000 и ICER.

Кроме того в качестве примера можно привести TSA-деревья и 2D TSA-деревья, предоставляющие эффективные структуры хранения временных рядов и пространственно-временных данных, которые поддерживают многоуровневые запросы (на разных уровнях абстракции) о тренде в данных и внезапных изменениях их поведения [8, 9].

Этап предварительной обработки и подготовки данных к анализу

Предварительная обработка данных является важным шагом, который обеспечивает качество данных и повышает эффективность и качество процесса анализа.

Данные из реального мира склонны быть неполными, зашумленными, противоречивыми и избыточными и поэтому непосредственно не подходят для проведения анализа.

Подготовка данных для анализа включает очистку для удаления шума и выбросов, интеграцию данных из различных информационных источников, прореживание/сжатие данных для уменьшения размерности и сложности данных, преобразование данных в подходящий для анализа формат [4,5].

Благодаря свойству нулевых моментов вейвлетов, в большинстве случаев только некоторые из

вейвлет-коэффициентов являются значимыми и несут полезную информацию, а большая часть пренебрежимо малы и могут быть отнесены к шуму. Сохраняя определенные вейвлет-коэффициенты можно применить вейвлет-преобразование для очистки данных от шума (пороговая обработка вейвлет-коэффициентов) либо уменьшения размерности данных.

К примеру, в работе [10] приведен анализ особенностей применения вейвлет-преобразования для очистки сигналов и изображений от шума, а в работе [11] предложен метод, использующий вейвлет-преобразование для выделения значимых атрибутов в данных.

Этап анализа данных и извлечения знаний

Непосредственный анализ – это основной процесс, при котором применяются различные интеллектуальные методы для извлечения полезных шаблонов из данных.

Основная идея применения вейвлет-преобразования на данном этапе состоит в следующем.

Как известно, вейвлет-преобразование имеет свойство уменьшать временную корреляцию данных, т.е. корреляция вейвлет-коэффициентов гораздо меньше, чем корреляция соответствующего исходного временного ряда [2]. Благодаря этому свойству простые традиционные модели ИАД, которые не могут быть применены на исходных сложных данных, могут быть достаточно точны в области вейвлетов.

Другими словами мы можем преобразовать исходные данные в область вейвлетов и в нем проводить последующий анализ.

Кроме того,вейвлет-преобразование предоставляет средства выделения характерных особенностей в данных, которые могут быть использованы в качестве признаков в существующих традиционных методах ИАД.

Вейвлет-анализ в таком виде применяется для решения задач кластеризации данных, сегментации изображений, классификации, регрессии, предсказания числовых рядов, выявления аномалий и выбросов в данных, поиска похожих объектов по образцу (аудио файлов, изображений), визуализации данных.

Среди проанализированных методов ИАД, использующих ВП, необходимо отметить наиболее интересный метод кластеризации WaveCluster [12]. Данный метод рассматривает всю совокупность данных как сигнал в N-мерном пространстве атрибутов и пытается на основании вейвлет-преобразования выделить в этом сигнале низкочастотные поддиапазоны частот, в которых связанные компоненты и будут кластерами. Эксперименты показали, что метод WaveCluster по своей эффективности превосходит такие методы Birch и CLARANS.

Интеллектуальный анализ временных рядов

Одним из направлений ИАД, перспективных с точки зрения исследований, является анализ временных рядов. Временной ряд – это последовательность собранных в разные моменты времени значений каких-либо параметров исследуемого процесса или объекта [13].

Методы интеллектуального анализа призваны исследовать временные ряды с целью нахождения в них скрытых шаблонов. Не смотря на то, что все эти методы используются различные математические подходы, все они имеют одну общую особенность: для их применения скорее необходимо некоторое высокоуровневое представление данных, чем исходные сырые данные.

Такое высокоуровневое представление необходимо как для выделения характеристик временного ряда, так и для его эффективного хранения, передачи и обработки.

Вейвлет-преобразование по своей природе призвано предоставить такое высокоуровневое представление временных рядов и поэтому может быть эффективно применено в следующих направлениях.

1. Поиск подобия во временных рядах предполагает отыскание полного или частичного совпадения с заданным рядом (шаблоном). При этом вводится параметр допустимого расстояния между рядами.

Решение данной задачи состоит из двух этапов: индексирования рядов и выполнения запросов. Индексирование – это процесс создания указателей для ускорения доступа к данным, предполагающий выделение признаков временного ряда и его сжатие. ДВП в данной задаче может быть применено для выполнения индексирования данных и создания метрик расстояния между рядами.

2. Классификация временных рядов состоит в назначении ряду одной из заранее известных меток класса. ДВП может быть интегрировано в классификацию временных рядов двумя путями: применение методов классификации к результатам вейвлет-преобразования и применение многомасштабного представления данных.

3. Кластеризация временных рядов состоит в их разбиении на группы по подобию. Кластеризация позволяет идентифицировать шаблоны и тренды для каждой из групп.

4. Выявление аномалий в поведении временных рядов включает: определение внезапных изменений в рядах, выявление аномалий путем сравне-

ния расстояний между несколькими рядами и выявление нерегулярных шаблонов.

5. Прогнозирование значений временного ряда в будущем на основании исторических данных.

Выводы

Если охарактеризовать данные, как зафиксированные факты, тогда знания – это множество шаблонов или зависимостей, которые лежат в основе этихданных. В базах данных, которые накапливаются по всему миру, спрятано огромное количество потенциально важных, но еще не выявленных знаний.

Для каждого этапа интеллектуального анализа данных в роботе проведен обзор возможностей применения вейвлет-преобразования. Показаны характеристики вейвлетов, которые могут помочь привнести улучшения в существующие методы данной сферы науки. Проведенный анализ позволил выделить наиболее перспективное направление дальнейших исследований, связанное с анализом числовых рядов.

Список литературы

1. Малла, С. Вэйвлеты в обработке сигналов [Текст] / С. Малла. – М.: Мир, 2005. – 671 с.

2. Addison, P.S. Illustrated wavelet transform handbook. Introductory Theory and Applications in Science, Engineering, Medicine and Finance / Paul S. Addison. – Bristol: Institute of Physics Publishing, 2002. – 400 p.

3. Барсегян, А.А. Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: БХВ-Петербург, 2007. – 384 с.

4. Han,J. Data Mining: Concepts and Techniques / J. Han, M. Kamber. – MorganKaufmann Publishers, 2000. – 550 p.

5. Witten,I.H. Data Mining: Practical Machine Learning Tools and Techniques / I.H. Witten, E.Frank, M.A. Hall.– MorganKaufmann, 2011. – 664.

6. Arunalatha, G. A Survey of Wavelet Compression Methods / G. Arunalatha, S. Sarathraj, K.Manodurga // International Journal of Scientific & Engineering Research. – – 2013. – Vol. 4, Is. 3. – P. 1-4.

7. Remya, S. Wavelet Based Compression Techniques: A Survey / S. Remya, V.A. Dilshad Rasheed // Advances in Communication, Network, and Computing.– 2012. – Vol.108. – P. 394-397.

8. Shahabi,C. TSA-tree: A Wavelet-Based Approach to Improve the Efficieny of Multi-Level Surprise and Trend Queries / C. Shahabi, X. Tian, W. Zhao // Proceedings of 12th Int. Conf. ofScientificandStatisticalDatabaseManagement, 2000. – P. 55-68.

9. Shahabi, C. 2D TSA-tree: a wavelet-based approach to improve the efficiency of multi-level spatial data mining / C. Shahabi, S. Chung; M.Safar, G. Hajj // Proceedings of 13th Int. Conf. ofScientificandStatisticalDatabaseManagement 2001. – P. 59-68.

10. Ergen, B.Signal and Image Denoising Using Wavelet Transform / Burhan Ergen // Signal and Image Denoising Using Wavelet Transform, Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology. – InTech, 2012. – P.495-514.

11. Tripathy, A. Dimensionality Reduction of Data Warehouse Using Wavelet Transformation: An Enhanced Approach for Business Process / A.Tripathy, D.Kaberi, S. Tripti// Computer Networks and Information Technologies. – 2011. – Vol. 142. P. 523-525.

12. Sheikholeslami G.WaveCluster: A multi-resolution clustering approach for very large spatial databases / G. Sheikholeslami, S. Chatterjee, A. Zhang // Proc. 24th Int. Conf. VeryLargeDataBases. – 1998. – P. 428–439.

13. Витязев, В.В. Вейвлет-анализ временных рядов / В.В.Витязев. – Спб.: Изд-во С.-Петерб. унт-та, 2001. – 58 с.

Поступила в редколлегию 17.09.2015

Рецензент: д-р техн. наук проф. А.В. Переверзев, Запорожский национальный технический университет, Запорожье.

ОГЛЯД ЗАСТОСУВАНЬ ВЕЙВЛЕТ-ПЕРЕТВОРЕННЯ В ЗАДАЧАХ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Т.В. Юр

Останнім часом все більш широке застосування знаходить математичний апарат вейвлет-аналізу в різних сферах науки і при вирішенні практичних завдань. Дана робота присвячена огляду застосувань вейвлетів в методах інтелектуального аналізу даних, тобто в методах виявлення в даних нових знань. Для структурування огляду весь процес вилучення нових знань розбитий на послідовність етапів, для кожного з яких показані можливості вейвлетів в даній області. Показані потенційні напрямки подальшого дослідження і застосування вейвлетів.

Ключові слова: *вейвлет, аналіз даних, вилучення знань, часовий ряд.*

REVIEW OF WAVELET TRANSFORM APPLICATIONS IN DATA MINING TASKS

T.V. Yur

In recent years, the increasing use of mathematical apparatus is wavelet analysis in various fields of science and in solving practical problems. This paper provides an overview of the use of wavelets in the methods of data mining, ie, detection methods in these new knowledge. For structuring review all new knowledge extraction process is divided into a sequence of stages, each of which shows the possibility of wavelets in the art. Showing potential areas for further research and application of wavelets.

Keywords: *wavelet, data analysis, knowledge discovery, time series.*