

Анализ многоагентных систем для анализа данных

Юлейси Гонсалес Перес¹, И. И. Холод²

Санкт-Петербургский государственный электротехнический университет

«ЛЭТИ» им. В.И. Ульянова (Ленина)

Санкт-Петербург, Россия

¹yuleisy2688@gmail.com, ²iiholod@mail.ru

Аннотация. Агенты являются важной частью решения многих актуальных проблем в области распределенных вычислений, включая сложность и информационную перегрузку. Существует необходимость интеграции многоагентной технологии и интеллектуального анализа данных, чтобы противостоять росту числа гетерогенных и распределенных источников данных. В данной статье описываются основные понятия: агентов, интеллектуального анализа данных, анализируются возможности их интеграции. Также приведен сравнительный анализ существующих систем интеллектуального анализа данных на базе многоагентных систем.

Ключевые слова: агент; агент интеллектуального анализа данных; интеллектуальный анализ данных; многоагентная система

I. ВВЕДЕНИЕ

В настоящее время, по мере совершенствования телекоммуникационных сетей, наблюдается тенденция к развитию распределенных систем, обеспечивающих большую гибкость и потенциал решения сложных задач. В контексте искусственного интеллекта многоагентные системы (МАС) были описаны как возможное решение для решения сложных задач с распределенными функциями. Развитие МАС предполагает заметное увеличение сложности решаемых задач, а также необходимость адаптации существующих методов или разработки новых для их решения [1].

Одна из основных проблем, стоящая в настоящее время перед интеллектуальным анализом данных (ИАД), это резкий рост объемов анализируемой информации, а также числа распределенных и гетерогенных источников. Для ее решения используют распределенные и параллельные методы ИАД, которые предполагают централизованное управление, что снижает их надежность. МАС предлагает управление, которое не является централизованным и, следовательно, такие системы могут быть гораздо более надежными и универсальными [2].

Интеграция МАС и ИАД это прекрасная возможность создать инновационные инструменты и системы, которые могут дать значительные результаты в распределенных средах. В случае распределенных данных, агенты могут

размещаться на источниках, обрабатывая их, как множество взаимодействующих сущностей [3].

II. АГЕНТ

Для объединения усилий и стандартизации разработки МАС была создана международная организация Foundation for Intelligent Physical Agents (FIPA). В последние годы спецификации FIPA стали фактически стандартами, используемыми разработчиками МАС в компьютерном сообществе и за его пределами. FIPA определяет интеллектуального агента, как систему, обладающую следующими свойствами [4]:

- **Автономность:** агенты могут работать без вмешательства пользователя и иметь контроль над своими действиями и внутренним состоянием.
- **Общительность:** агенты могут общаться с другими агентами с помощью какого-то языка.
- **Реактивность:** агенты могут воспринимать свое окружение и своевременно реагировать на изменения в нем.
- **Проактивность:** агенты контролируют свои планы или цели, несмотря на изменения в окружающей среде.
- **Адаптивность:** агент может обучаться и изменять свое поведение на основе этого обучения.
- **Мобильность:** способность агента перемещаться в сети.

Для взаимодействия агентов FIPA был предложен язык общения агентов (ACL), называемый FIPA-ACL.

A. Распределенный интеллектуальный анализ данных

Григорий Пятецкий-Шапиро дал следующее определение ИАД – исследование и обнаружение «машины» в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны и доступны для интерпретации человеком [5].

К задачам ИАД относят: классификация, регрессионный анализ, кластеризация, поиск ассоциативных правил, извлечение наиболее значимых атрибутов и др.

Распределенный интеллектуальный анализ данных (РИАД) определяется как ИАД в больших объемах данных, распределенных между несколькими узлами [6]. Предполагается, что данные распределяются между двумя или более узлами, и эти узлы взаимодействуют друг с другом для получения глобальных результатов.

Выделяют несколько стратегий РИАД [7]:

- **Стратегия централизованного обучения:** предполагает сбор всех данных в центральном хранилище и построение в нем единой модели. Эта стратегия используется, когда объемы распределенных данных невелики.
- **Стратегия мета-обучения:** она предлагает извлечения локальных знаний из однородно распределенных данных, их сборе на центральном узле и последующем объединении в единый результат.
- **Гибридная стратегия обучения:** предполагает оба способа обучения: централизованный и мета-обучение.

Стратегия централизованного обучения может быть неэффективной или неосуществимой по следующим причинам [8]:

- **Высокая стоимость связи.** Передача огромных объемов данных по сети с ограниченной пропускной способностью может занять чрезвычайно много времени, а также потребовать высоких финансовых затрат.
- **Высокая стоимость вычислительных ресурсов.** Вычислительная стоимость анализа центрального хранилища данных намного больше, чем сумма затрат на анализ небольших частей данных, которые могут быть выполнены параллельно.
- **Конфиденциальные данные.** Передача конфиденциальных данных (например, медицинская и финансовая информация о людях) по сети нежелательна, поскольку ставит под угрозу их безопасность.

Эти факторы являются причиной использования альтернативных стратегий, которые реализуются за счет интеграции технологий МАС и ИАД в виде агентов ИАД.

В. Агенты интеллектуального анализа данных

По отдельности каждая из технологий ИАД и МАС имеет специфические проблемы [1]:

- МАС сталкиваются с проблемой повышения способности агентов к обучению;
- ИАД – сталкивается с проблемой распределенных источников данных, поскольку данные становятся все более разнородными и размещаются в разных местах, число которых постоянно растет.

В результате интеграция МАС и ИАД может выполняться в интересах [3]:

- агентов (data mining-driven agents), которые используют ИАД для поддержки таких своих свойств как: адаптивность, обучаемость, реактивность и проактивность.
- ИАД (agent-driven data mining), использующие такие свойства агентов как: общительность, автономность, адаптивность и мобильность, в основном для решения задач в распределенных средах.

Во втором случае создаются агенты интеллектуального анализа данных (АИАД), которые благодаря своей автономной, гибкой, мобильной, адаптируемой и рациональной природе являются отличным выбором для параллельного и распределенного анализа.

Следующие характеристики агентов могут быть использованы для решения задач ИАД [7]:

- **Автономность.** АИАД является модульным расширением системы анализа данных, позволяющим осуществлять доступ к источнику информации в соответствии с заданными ограничениями.
- **Общительность.** АИАД могут работать независимо на локальных источниках. В процессе они могут взаимодействовать, чтобы обмениваться промежуточными результатами, если это требуют реализуемые алгоритмы анализа.
- **Реактивность.** АИАД могут реагировать на изменения, происходящие с анализируемыми данными. Например, дообучить модель, в случае поступления новых данных.
- **Проактивность.** АИАД могут планировать свою работу на основе получаемой информации. Например, перемещаться к другим источникам информации, если это требуется для уточнения строящейся модели.
- **Адаптивность.** АИАД может адаптировать свое поведение в зависимости от промежуточных результатов. Например, выбирать источник в зависимости от качества данных в нем, каналов связи и т.п.
- **Мобильность.** АИАД может перемещаться к источнику данных, выполняя анализ на нем и перенося с собой только результаты.

Интеграция МАС и ИАД позволяет реализовать две стратегии РИАД: мета-обучение и гибридное обучение. По сравнению с централизованным обучением это дает следующие преимущества:

- уменьшение сетевого трафика – пропускная способность уменьшается за счет передачи между узлами только кода, а не данных;
- эффективное масштабирование без увеличения сложности системы, поскольку агент может породить другого агента;

- асинхронная работа, благодаря чему агенты отключаются от сети и автоматически подключаются после выполнения своих задач;
- надежность, поскольку агенты могут обходить источники данных с ошибками и использовать только надежные источники;
- отказоустойчивость, поскольку выход из строя одного агента не приводит к отказу всей системы, в том числе и за счет замены отказавшего агента.

Необходимо отметить, что мета-обучение при использовании MAC может быть реализовано двумя способами:

- параллельная работа нескольких агентов, когда агенты извлекают локальные модели из данных и передают их агенту-мастеру для их объединения (рис. 1.а).
- последовательная работа одного агента, когда он перемещается от источника к источнику, достраивая модель (рис. 1.б).

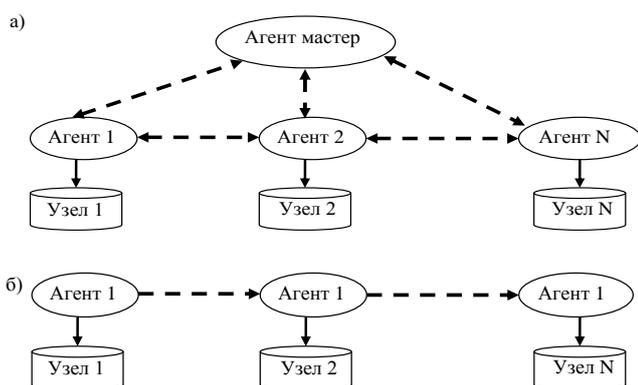


Рис. 1. Стратегии мета-обучения а) параллельный, б) последовательный

Параллельная стратегия применима, когда данные могут быть обработаны независимо друг от друга. Последовательная стратегия применима к данным, обработка которых зависит от результатов, полученных на предыдущих источниках.

III. МНОГОАГЕНТНЫЕ СИСТЕМЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

В настоящее время существует большое число MAC ИАД. Они реализуют разные стратегии и решают разные задачи анализа.

Parallel DM Agents (PADMA): основана на стратегии централизованного обучения. Агенты PADMA используются для доступа к локальным данным и решения задачи кластеризации и используют параллельную стратегию работы [2].

PAPYRUS: основана на стратегии гибридного обучения. Конечным результатом являются файлы PMML

(Predictive Model Markup Language). Papyrus поддерживает различные методы объединения и обмена локально добытыми прогностическими моделями и метаданными, требующимися для их описания с помощью специального языка разметки [6], [9].

Extendible Multi-Agent Data Mining System (EMADS): это коллекция независимых постоянных, автономных (но сотрудничающих) агентов KDD, действующих через Интернет. Она основана на стратегии централизованного обучения, и агенты используют стратегию параллельной работы [10].

Multi-Agent Data mining Architecture using Meta-learning (MADAM): это архитектура РИАД, которая сочетает MAC и мета-обучения с целью повышения производительности системы РИАД [11].

Mobile Agent using Data mining-based Intrusion Detection System (MAD-IDS): использует метод обнаружения неконтролируемых аномалий, основанный на алгоритме кластеризации. Эта система основана на стратегии мета-обучения, и используют стратегию параллельной работы [12].

Pikater: позволяет пользователю получить знания, предложив наилучший возможный метод ИАД для анализа набора данных на новом источнике, полученного путем мета-обучения. Агенты используют стратегию параллельной работы и OWL-DL онтологии для прямого взаимодействия между агентами [13].

Multi Agents based Data Mining for Intelligent Decision Support Systems (DMMAS): разрабатывалась как библиотека, которая может быть встроена и использована другими приложениями. Основана на стратегии гибридного обучения и агенты используют стратегию параллельной работы [14].

Agent-based Data Mining Info-structure (ADMI): разработана для предоставления стратегических услуг в контексте здравоохранения. Все модули агентов ADMI разработаны в отдельном автономном потоке (многопоточном), и поэтому все модули выполняются одновременно и непрерывно, за исключением инициализации агента [15].

Java Agents for Meta-learning (JAM): предоставляет набор агентов для объединения нескольких моделей, изученных на разных сайтах, в мета-классификатор, который во многих случаях повышает общую точность прогнозирования [16].

WODNI: это мобильная агентная РИАД система, использующая фреймворк коллективного ИАД на гетерогенных сайтах данных. Эта система основана на стратегии мета-обучения, и используют стратегию параллельной работы [17].

В таблице показаны MAC на основе ИАД и их характеристики.

ТАБЛИЦА I. MAC ОСНОВАННЫЕ НА ИАД

Характеристики \ АИД	PADMA [2]	Papyrus [6], [9]	EMADS [10]	BODHI [17]	JAM [16]	Pikater [13]	DMMAS [14]	ADMI [15]	MADAM [11]	MAD-IDS [12]
Задачи ИАД: классификация (1), регрессионный анализ (2), кластеризация (3), поиск ассоциативных правил (4), извлечение наиболее значимых атрибутов (5)	3	3	1, 4	5	1, 4	5	1, 4	4, 5	1	3, 4
Стратегия обучения: централизованного (1), мета-обучения (2), гибридная (3)	1	3	1	2	2	2	3	2	2	2
Стратегия работы агентов: параллельная (1), последовательная (2)	1	1	1	1	1	1	1	1	1	1
Язык программирования	C++	Java	Java	Java	Java	Java	Java	Java	Java	Java
MAC система	-	IBM Aglets	JADE	IBM Aglets	-	JADE	JADE	JADE	JADE	JADE
FIPA	-	-	+	-	-	+	+	+	+	+
DM библиотека	Parallel Portable File System (PPFS)	-	-	-	-	-	WEKA	-	Learning Algorithm Repository	JPCAP
Язык, используемый в сообщениях	-	Special markup language	FIPA-ACL	KQML format	-	OWL-DL	FIPA-ACL	FIPA-ACL	FIPA-ACL	FIPA-ACL
PMML	-	+	-	-	-	-	-	-	-	-

IV. ЗАКЛЮЧЕНИЕ

В статье рассмотрены технологии ИАД и MAC, а также преимущества от их интеграции для анализа данных хранящихся в распределенных источниках. Основными достоинствами такой интеграции являются: снижение сетевого трафика, простота масштабирование, повышение безопасности данных и др. Проведен сравнительный анализ существующих MAC, выполняющий ИАД. В результате можно сделать вывод, что они не являются достаточно универсальными и гибкими. Они не позволяют: решать всех задач анализа данных, использовать все стратегии анализа, интегрировать стандартные языки общения агентов (FIPA ACL) и описания моделей знаний (PMML).

Дальнейшие наши исследования будут направлены на устранение этих недостатков.

СПИСОК ЛИТЕРАТУРЫ

- [1] Molero C.G., Meda C.M.E., "Integración de Minería de Datos y Sistemas Multiagente: un campo de investigación y desarrollo," Ciencias de la Información, T. 41, вып 3, septiembre-diciembre 2010. P. 53-56.
- [2] Chaimontree S., Atkinson K., Coenen F., "A Multi-Agent Based Approach To Clustering: Harnessing The Power of Agents," Springer, 2012. P. 1-2.
- [3] Mesbahi N., Zoubeidi M., Merizig A., Kazar O., "An Agent-Based Approach for Extracting Business Association Rules from Centralized Databases Systems," Journal of Digital Information Management, T. 17, вып 5, October 2019. P. 2.
- [4] Agüero M. J. L., "Diseño de organizaciones virtuales ubicuas utilizando desarrollo dirigido por modelos," Valencia, 2014. P. 36-37.
- [5] Распределенные объектные технологии / Радченко Г.И., Захаров Е.А. Южно-Уральский государственный университетИздательский центр ЮУрГУ, 2013. С 21.
- [6] Urmela S., Nandhini M., "Approaches and Techniques of Distributed Data Mining: A Comprehensive Study," International Journal of Engineering and Technology (IJET), T. 9, вып 1, 2017. P. 69.
- [7] Naidu P.A., "Distributed Data Mining and Multi Agent-Based Distributed Data Clustering," International Journal of Science and Research (IJSR), T. 2, вып 1, 2013. P. 96-97.
- [8] Ganga D. S. V. S., "A Survey on Distributed Data Mining and its Trends," IMPACT: International Journal of Research in Engineering & Technology, T. 2, вып 3, 2014. P. 107-108.
- [9] Klusch M., Lodi S., Moro G., "Agent-Based Distributed Data Mining: The KDEC Scheme," Springer, March 2003. P. 104-122.
- [10] Albashiri K.A., Coenen F., Leng P., "EMADS: An Extendible Multi-Agent Data Miner," Springer, 2008. P. 2-7.
- [11] Kumar S.S., Kumar P.S., Charan O.A., Dash S., "MADAM: A Distributed Data Mining System Architecture Using Meta-Learning," The IUP Journal of Information Technology, T. X, вып 4, 2014. P. 5-11.
- [12] Brahmi I, Yahia S. B., Aouadi H., Poncelet P., "Towards a Multiagent-Based Distributed Intrusion Detection System Using Data Mining Approaches," Springer, 2012. P. 3-5.
- [13] Hegde N.P., Varija B., "Data Mining and MultiAgent Integration," International Journal of Engineering Trends and Technology (IJETT), 2017. P. 10-12.
- [14] Sharma D., Shadabi F., "Multi-agents based data mining for intelligent decision support systems," IEEE, January 2015. P. 241-243.
- [15] Hassan Z. S. Z., R. A, S. S., Manikam S., Yu-N Ch., "ADMI: A Multi-Agent Architecture to Autonomously Generate Data Mining Services," in Proceedings of the 2do International IEEE Conference on Intelligent Systems, T. 1, 2004. P. 273-279.
- [16] Stolfo S., Prodromidis A. L., Tselepis Sh., Lee W., Fan D. W., Chan P. K., "JAM: Java Agents for Meta-Learning over Distributed Databases," Proceedings of the International Conference on Knowledge Discovery and Data Mining, 1997. P. 74-81.
- [17] Kargupta H., Park B. H., Hershberger D., Johnson E., "Collective Data Mining: A New Perspective Toward Distributed Data Analysis," in Advances in Distributed and Parallel Knowledge Discovery, 2000. P. 131-178.