Классификационная характеристика для задач обработки разнородных данных

Багутдинов Р.А.

***Аннотация*—В работе рассмотрены некоторые аспекты решения задачи быстрого, верного и эффективного выбора методов обработки данных, на базе классификационной характеристики разнородных данных и соответствующих определенных критериев. Основываясь на теоретических исследованиях, в том числе в области системного анализа проведен классификационный анализ разнородных и разномасштабных данных и соответствующих методов их обработки, в том числе с использованием методов математической статистики.**

Автором предпринята попытка классификации основных, наиболее чаще встречающихся методов обработки данных для мультисенсорных систем с целью выявления рекомендаций поиска более эффективного и быстрого варианта решения необходимой для исследователя задачи. Актуальность такого подхода подкрепляется слабо формулированными задачами и универсальными рекомендациями в зависимости от степени значимости типа данных для решения конкретной практической задачи.

***Ключевые слова*—Мультисенсорные системы, обработка данных, классификация, методы обработки данных, разнородные данные.**

I. Введение

Наиболее активно развивающимися областями науки являются направления, связанные с обработкой больших массивов данных: компьютерное зрение, робототехника, физика элементарных частиц и другие.

Проблема обработки, анализа и хранения больших объемов данных, получаемых от множества различных сенсоров, является актуальной задачей. Сенсор в контексте задач получения и обработки данных может рассматриваться достаточно широко, включая в себя любой источник цифровой информации [1].

Следовательно, совокупность таких источников представляет собой мультисенсорную систему (МС). С развитием информационных технологий, наблюдается бурный интерес к решению задач обработки больших объемов данных (Big Data). Однако на сегодняшний день все еще не существует эффективного решения проблемы создания универсальных моделей, способов, алгоритмов и методов для разнородных, неформализованных и неструктурированных данных, имеющих различные типы и источники происхождения. Успешное решение этой проблемы приведет к существенному прогрессу в широком спектре прикладных задач за счет повышения эффективности, скорости работы таких систем и скорости принятия решений на основе обработки большого объема разнородных данных.

Статья получена 07 мая 2018.

Багутдинов Равиль Анатольевич, аспирант, ассистент, программист Отделения автоматизации и робототехники Инженерной школы информационных технологий и робототехники (бывший Институт кибернетики) Национального исследовательского Томского политехнического университета, г. Томск, Россия (e-mail: ravil_bagutdinov@yahoo.com).

II. Описание работы

В данной работе представлена классификационная характеристика данных и подбор соответствующих методов, рекомендуемых для обработки данных. Быстрый выбор наиболее оптимального метода для обработки больших данных имеет большое значение для многих сфер науки и техники, и являются основной частью многих интеллектуальных систем, основанных на идее работы с большими объемами гетерогенных данных. Сложность задач, решаемых такими системами, постоянно увеличивается, а требования к их техническим характеристикам повышаются. При этом все острее встают вопросы надежности и точности подобных систем, особенно для таких критичных направлений как автономные транспортные средства, системы безопасности, медицинские технологии, моделирование и прогнозирование природных и социальных событий [2].

Основу принципов обработки данных составляют различные методы классификации, моделирования и прогнозирования. К ним зачастую относят математические и статистические методы (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов) [13]. Такие методы, однако, оперируют лишь некоторыми заранее известными представлениями об анализируемых данных, что препятствует обнаружению ранее неизвестных нетривиальных и практически полезных знаний. Однако из основных задач применения и назначений методов обработки данных следует наглядное представление результатов вычислений, что позволяет использовать те или иные адаптивные инструментарии пользователями, которые могут не иметь специальной математической подготовки. В то же время, применение статистических методов анализа данных требует хорошего владения теорией вероятностей и математической статистикой.

Различные подходы и принципы обработки данных характеризуются определенными свойствами, которые

могут быть определяющими при выборе метода анализа данных. Методы можно сравнивать между собой, оценивая характеристики их свойств [14].

Анализируя существующие подходы и принципы, можно сделать выводы о необходимости более детальной проработки, как предварительного анализа данных, так и самих используемых алгоритмов [15].

Учитывая текущий прогресс в области вычислительной техники и сбора данных, а также связанный с этим экспоненциальный рост объемов информации, поступающей из различных источников, актуальными становятся такие проблемы, как:

– структуризация и классификация типов данных и существующего разнообразия методов обработки, определение их зависимости для выбора оптимального решения;

– разработка новых подходов к созданию универсальных технологии и систем обработки разнородных больших данных;

Цель данной работы - разработка классификационной характеристики данных и соответствующей классификации методов обработки данных в зависимости от требуемого приоритета решения задачи (определение качественных или количественных показателей). В работе рассматриваются методы структурирования, классификации, обработки гетерогенных данных мультисенсорных систем.

Значимость исследования заключается как в систематизации подходов к проблеме разработки высокопроизводительных систем обработки многомерных разнородных данных, так и в демонстрации эффективности применения этих подходов для оптимизации всех компонент таких программных систем под конкретную прикладную задачу и заданные требования. Планируется исследование методов поиска и оценки достоверности источников информации, оптимальных путей и скорости распределения информации. В дальнейшем разработанная система может легко масштабироваться различными наборами данных для применения в конкретных областях технических задач.

Существует множество исследований и практических реализаций систем обработки разнородных данных для различных специализированных задач: мониторинг технического состояния различных объектов, поиск аномальных и ложных данных среди источников информации, прогнозирование погодных явлений [3]. Несмотря на это, недостаточно внимания уделяется разработке универсального, комплексного и системного подхода к процессу разработки таких систем. Зачастую данные синхронизированы по времени, но также встречаются и другие данные, которые не могут быть синхронизированы по времени, отчетными точками для таких данных могут быть схожие параметры.

При создании такой модели классификационной характеристики можно учитывать выбор данных по определенным критериям, относительно времени, часть данных могут переходить из одного типа данных в другой или быть одновременно частью этих типов. Предлагаемая классификация данных представлена на (Рис. 1).

Данные также могут быть как регулярные, так и не регулярные, т.е. не только разнородные, но и разномасштабные по времени. Для обработки нерегулярных данных, как можно заключить, требуется наиболее сложные методы обработки или комплекс методов.

Классификация методов согласно различным практическим задачам представлена на (Рис 2).

Для того чтобы лучше понять способ применения предлагаемых классификаций на (Рис. 3) приведен пример реализации на примере обработки разнородных данных, полученных с газоанализатора и тепловизора [4].

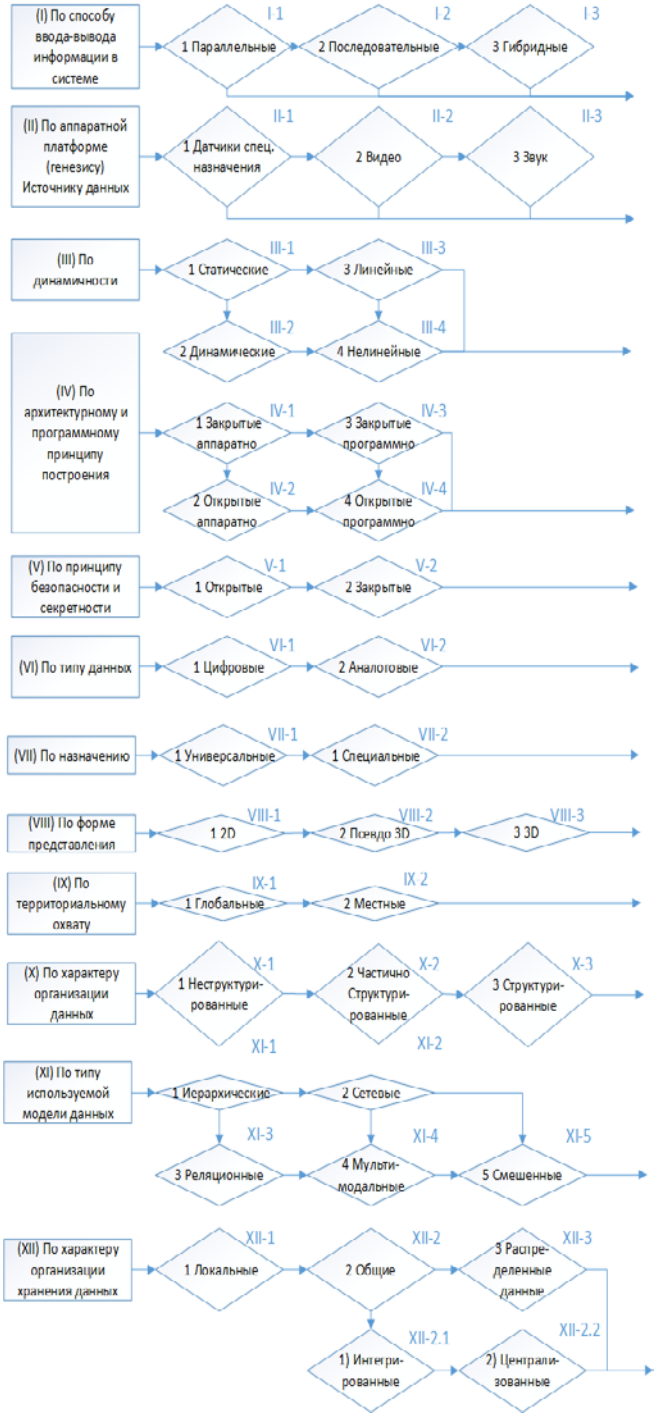

Рисунок 1. Классификация типов данных

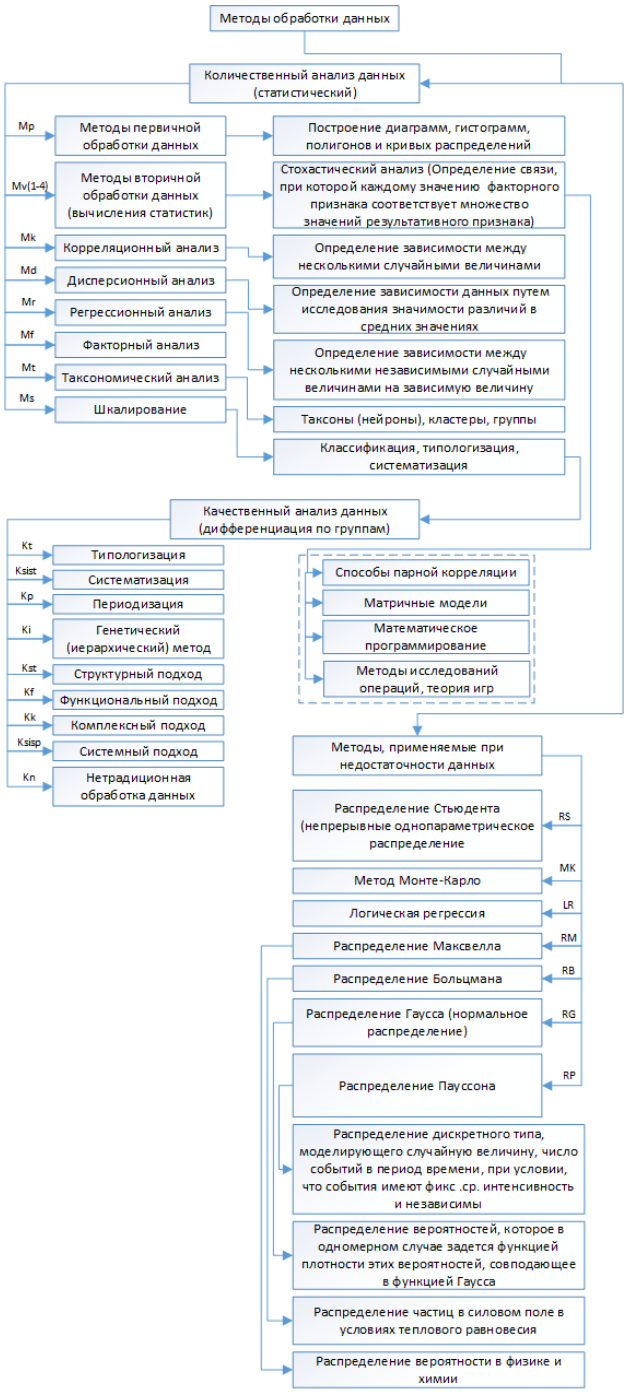

Рисунок 2. Классификация методов обработки данных в зависимости от задач.

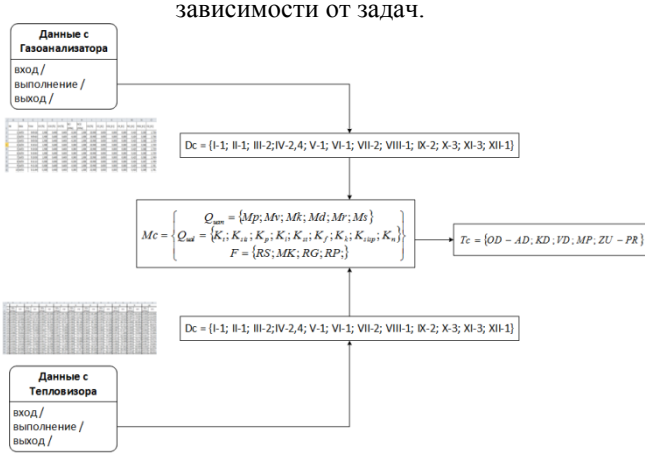

Рисунок 3. Модель применения классификационной характеристики данных.

Следуя логики вещей, задачи обработки данных имеют соответствующие методы их решения, а методы в свою очередь характеризуются подмножеством разнородных данных. Соответственно, классификацию данных можно представить в виде:

$$D_1 = \begin{Bmatrix} I1; II1; III1; IV1; V1; VI1; \\ VII1; VIII1; IX1; XI1; XII1 \end{Bmatrix} \quad (1)$$

Классификацию методов обработки данных в зависимости от задачи можно представить в виде:

$$Mc = \begin{Bmatrix} Q_{uan} = \{M_p; M_v; M_k; M_{d;} M_r; M_s\} \\ Q_{ual} = \begin{Bmatrix} K_t; K_{sis}; K_p; K_{t;} K_{st}; \\ K_f; K_{k;} K_{sisp}; K_{nod} \end{Bmatrix} \\ Q_f = \{R_s; R_{mk}; R_g; R_p\} \end{Bmatrix} \quad (2)$$

Классификацию задач обработки данных можно представить в виде:

$$Tc = \{T_{od-ad(1,2)}; T_{kd}; T_{vd}; T_{mp}; T_{zd}; T_{hd}; T_{pd}; T_{zu-pr}\} \quad (3)$$

В результате получаем следующую расшифровку классификационной характеристики данных: По способу ввода-вывода информации в системе данные поступают параллельно, источник данных – датчики (конкретизация типа датчиков описывается дополнительно), данные имеют статический вид, данные открыты аппаратно, имеют открытый доступ, данные цифровые, универсальные, по форме представления двумерные, глобальные, неструктурированные, модель данных иерархическая, хранятся локально.

III. Заключение

Из-за постоянной компликации задач обработки, поиска, сбора и распределения большого объема данных возникает необходимость универсальной системы классификации и их взаимосвязи, а также хорошо проработанных сценариев их обработки. Предлагаемая классификационная характеристика способствует быстрому поиску решения задач, т.к. позволяет единовременно увидеть всю картину существующих связей.

Значимость проведенного исследования отражена как в систематизации подходов к проблеме мультисенсорной обработки многомерных разнородных данных, так и в применении этих подходов для оптимизации всех компонент таких систем под конкретную прикладную задачу и заданные требования.

Библиография

[1] *Ахо Альфред, Хопкрофт Д., Ульман Д.* Структуры данных и алгоритмы / Пер. с англ.: М: Издательский дом «Вильямс», 2003. 384 с.

[2] *Bagutdinov R.A., Zaharova A.A.* The task adaptation method for determining the optical flow problem of interactive objects recognition in real time. Journal of Physics: Conference Series. 2017;803(1):012014.https://doi.org/10.1088/17426596/803/1/012014

[3] *Багутдинов Р.А.* Гносеологические аспекты к определению назначения и состава СТЗ в задачах проектирования и разработки робототехнических комплексов. Программные

системы и вычислительные методы. 2017;1:39-45. https://doi.org/10.7256/2454-0714.2017.1.20372

[4] *Багутдинов Р.А., Небаба С.Г., Захарова А.А.* Алгоритм обработки разнородных данных для мультисенсорной СТЗ на примере анализа температуры и концентрации газа / В сборнике: ГРАФИКОН-2017 Труды 27-й Международной научной конференции. 2017.. С. 97-100.

[5] *Островский О.А.* Алгоритмы проведения осмотров цифровых носителей информации для предотвращения компьютерных преступлений. Военно-юридический журнал. 2017;11:3-6.

[6] *Островский О.А.* Принцип объектной декомпозиции в систематизации идентификационных кодов, характеризующих преступления в сфере компьютерной информации. Полицейская деятельность, 2017;3(3):10–18. https://doi.org/10.7256/2454-0692.2017.3.21869

[7] *Галямов А.Ф., Ризванов Д.А., Сметанина О.Н., Юсупова Н.И.* Модели и алгоритмы глобально распределённой обработки слабоструктурированных данных на основе микроразметки для поддержки принятия решений // Фундаментальные исследования. – 2017. – № 1. – С. 27-35

[8] *Мухитова А.А., Жижимов О.Л.* Адаптивные технологии при построении административных графических интерфейсов для гетерогенных информационных систем для ввода и редактирования данных / В сборнике: Распределенные информационные и вычислительные ресурсы. Наука – цифровой экономике (DICR-2017) Труды XVI всероссийской конференции. Институт вычислительных технологий СО РАН. 2017. С. 142-149.

[9] *Сибиряков М.А., Васяева Е.С.* Модификация и моделирование алгоритмов обработки данных в кэш-памяти систем хранения данных / Кибернетика и программирование. — 2016. - № 4. - С.44-57. doi: 10.7256/2306-4196.2016.4.18058.

[10] *Юревич Е. И.* Сенсорные системы в робототехнике: учеб. пособие / СПб.: Изд-во Политехн. ун-та, 2013. -100 с

[11] *Hastie, T., Tibshirani, R., & Friedman, J. H. (2001).* The elements of statistical learning : Data mining, inference, and prediction. New York: Springer.

[12] Witten, I. H., & Frank, E. (2000). Data mining. New York: Morgan-Kaufmann.

[13] Y. Maheo, F. Massi, N. Bouscharain, S. Milana, G. Le Jeune and Y. Berthier Degradation of high loaded oscillating bearings: Numerical analysis and comparison with experimental observations, Wear, vol. 317, pp. 141-152, Sep 2014.

[14] *Z. Pan, J. Polden, N. Larkin, S. Van Duin, and J. Norrish* Recent progress on programming methods for industrial robots // Robotics and Computer-Integrated Manufacturing, vol. 28, pp. 87-94, 2012.

[15] *Zaytsev A.* Variable fidelity regression using low fidelity function blackbox and sparsification // Lecture Notes in Computer Science,2016. V. 9653, P. 147–164.

Багутдинов Равиль Анатольевич родился в г. Караганде 26.04.1985, закончил с отличием Карагандинский высший политехнический колледж в 2006 году по специальности «Сети связи и системы коммутаций», получил степень бакалавра по специальности «Радиотехника, электроника и телекоммуникации» в Карагандинском государственном техническом университете в 2009 году, получил степень магистра по специальности «Техническая физика» в Национальном исследовательском Томском государственном университете в 2013 году, обучается в аспирантуре Национального исследовательского Томского политехнического университета.

Область научных интересов: теоретические основы и методы системного анализа, оптимизации, управления, принятия решений и обработки информации; методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений в технических системах; визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации; вопросы технического зрения и телекоммуникации в робототехнике для космической и военной отрасли; вопросы модернизации и оптимизации системы образования.

Classification characteristic for heterogeneous data processing tasks

Bagutdinov R.A.

***Abstract*—The paper considers some aspects of solving the problem of fast, correct and efficient choice of data processing methods based on the classification characteristics of heterogeneous data and the corresponding specific criteria. Based on theoretical studies, including in the field of system analysis, a classification analysis of heterogeneous and different-scale data and related methods of their processing, including using mathematical statistics methods, was carried out.**

The author made an attempt to classify the main, most frequently encountered data processing methods for multisensory systems in order to identify recommendations for finding a more efficient and quicker solution to the problem that is necessary for the researcher. The relevance of this approach is supported by poorly formulated tasks and universal recommendations, depending on the degree of significance of the type of data for solving a particular practical problem.

***Keywords*—Multisensory systems, data processing, classification, data processing methods, heterogeneous data.**