Редуктивно-аккомодативный принцип обработки гетерогенных данных мультисенсорной системы

Р.А. Багутдинов[1], С.Г. Небаба[1], А.А. Захарова[1]
ravil_bagutdinov@yahoo.com|stepanlfx@tpu.ru|zaa@tpu.ru
[1]Томский Политехнический Университет, Томск, Россия

В данной работе представлена классификационная характеристика гетерогенных данных и подбор соответствующих методов, рекомендуемых для их обработки. В работе предложен редуктивно-аккомодативный принцип обработки гетерогенных данных мультисенсорной системы. Изучена зависимость одной выборки данных от других. 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.

Ключевые слова: *разнородные данные, разномасштабные данные, датчики, робототехнические комплексы, гетерогенные данные, обработка данных, мультисенсорные системы.*

Reductive-accommodative principle of processing heterogeneous data of a multi-sensor system

R.A. Bagutdinov[1], S.G. Nebaba[1], A.A. Zakharova[1]
ravil_bagutdinov@yahoo.com|stepanlfx@tpu.ru|zaa@tpu.ru
[1]Tomsk Polytechnic University, Tomsk, Russia

In this paper, a classification characteristic of heterogeneous data is presented and the selection of appropriate methods recommended for their processing. The reductive-accommodative principle of processing heterogeneous data of a multi-sensor system is proposed in the work. The dependence of one sample of data on others was studied. Based on theoretical studies, including in the field of system analysis, a classification analysis of heterogeneous and different-scale data and corresponding methods of their processing, including using mathematical statistics methods, was carried out.

Keywords: *heterogeneous data, multiscale data, sensors, robotic complexes, heterogeneous data, data processing, multisensor systems.*

1. 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 различных сенсоров, является актуальной задачей [13]. Сенсор в контексте задач получения и обработки данных может рассматриваться достаточно широко, включая в себя любой источник цифровой информации. Следовательно, совокупность таких источников представляет собой мультисенсорную систему (МС). При этом характерной особенностью становится не только рост объема данных, но и увеличение их неоднородности. В связи с этим происходит отход от традиционных подходов обработки данных [9, 12].

С развитием информационных технологий, наблюдается повышенный интерес к решению задач обработки больших объемов данных (Big Data). Однако 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 решения прикладных задач за счет повышения эффективности, скорости работы таких систем и принятия решений на основе обработки большого объема разнородных данных.

Под термином «редукция данных» понимается совокупность аналитических методов для уменьшения размерности больших данных.

2. Редуктивно-аккомодативный принцип обработки гетерогенных данных МС.

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 [1].

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

1. структуризация и классификация типов данных и существующего разнообразия методов обработки, определение их зависимости для выбора оптимального решения;
2. разработка новых подходов к созданию универсальных технологий и систем обработки разнородных больших данных.

Предлагается редуктивно-аккомодативный принцип обработки гетерогенных данных МС, а также классификационная характеристика данных (КХД) и соответствующая классификация методов обработки

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

данных в зависимости от требуемого приоритета решения задачи (определение качественных или количественных показателей). Под предлагаемым редуктивно-аккомодативным принципом (от лат. reduction – сокращение, accommodatio – приспособление) понимается сокращение размерности больших гетерогенных данных за счет выделения ключевых параметров, объясняющих в той или иной степени изменение исходных данных. Для решения поставленной задачи используются методы структурирования, классификации, обработки гетерогенных данных МС.

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 [6, 10, 2]. 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, отсчетными точками для таких данных могут быть схожие параметры [11].

При создании такой модели КХД 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. КХД вводится для упрощения обработки данных и последующего решения конкретной практической задачи. В процессе классификации выбираются схожие параметры для того или иного типа данных. Т.е. предлагается универсальный способ описания данных и представление характеристики данных в качестве некоего комплекса знаний, в соответствии с которым можно будет предложить именно те методы обработки данных, которые наиболее подходят для решения конкретной задачи.

Данные также могут быть как регулярные, так и не регулярные, т.е. не только разнородные, но и разномасштабные по времени [14]. Для обработки нерегулярных данных, как можно заключить, требуется наиболее сложные методы обработки или комплекс методов [3]. Для того чтобы лучше понять способ применения предлагаемых классификаций на рисунке 1 приведен пример реализации обработки разнородных данных, полученных с газоанализатора и тепловизора [8]. Данные поступают с датчиков, проводится их первичная обработка, в соответствии с каждым набором данных выводится их КХД, в соответствии с которой впоследствии предлагается наиболее подходящий метод или методы обработки данных, которые приведут к получению наиболее полной ценной информации обо всей совокупности разнородных данных.

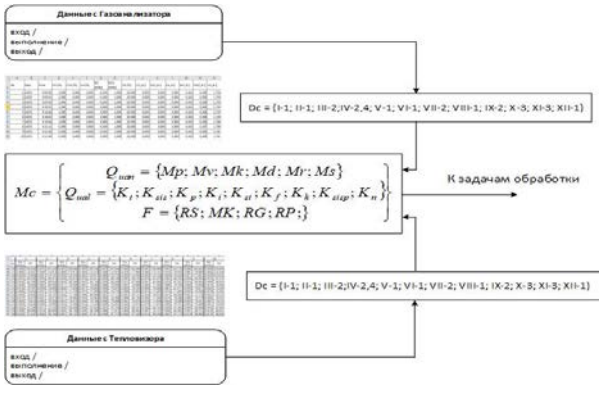

Рис. 1. Модель применения КХД.

Пример расшифровки описания КХД:

$D_1 = \{I1; II1; III1; IV1; V1; VI1; VII1; VIII1; IX1; X1; XI1; XII1\}$

Расшифровка: 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.

Проблема обработки больших разнородных данных зависит от постановки задач, методов обработки, и подходов к решению. Требуется объединение данных, полученных с различных источников, приведение их к общей понятной форме, для проведения последующих операций обработки. Решение такой проблемы обработки данных позволяет уйти от необходимости самостоятельно отбирать источники с наиболее важной информацией.

В научных исследованиях сформулировано множество условий, определений и методов, обработки больших данных. Но даже сейчас нет четкого соглашения о том, как ускорить этот процесс, упростить работу пользователя с данными, выявить наиболее ценную информацию из огромного потока данных, не потеряв при этом часть информации, которая может быть важна.

3. Математический аппарат обработки гетерогенных данных мультисенсорной системы, проведение эксперимента и анализ результатов.

Для проведения эксперимента используется МС на базе чипа Atmega 2560, на которой установлены различные датчики, данные с которых считываются через авторский скетч и поступают либо в таблицу Excel с помощью соответствующего скрипта, либо в БД для последующей обработки.

Упрощенно весь процесс обработки данных можно разделить на несколько этапов, приведенных ниже.

1. Сырые данные приводим к единой форме.
2. Проводим синхронизацию по времени.
3. Проводим сортировку данных, чтобы поток данных делился на интервалы по 100 замеров каждый для удобства расчетов.
4. Проводим комплексный корреляционный анализ.
5. Проводим комплексный регрессионный анализ.

Т.к. регрессионный анализ данных проводят по двум выборкам, то можем сравнить последовательно каждую выборку данных.

Для решения задачи регрессионного анализа отбросим те данные, которые не влияют на конечный результат анализа (всего 12 столбцов): последний столбец (т.к. там данные порядкового характера), первый столбец (т.к. там данные времени), второй столбец, т.к. там данные температуры, которые в нашем случае не изменялись). Итого 9 столбцов.

Сравниваем последовательно каждую выборку:

Для удобства обозначим каждый столбец буквой латинского алфавита (таблица 1).

Humidity (%)	PhotoPin (lux)	MQ2 (ppm)	Ratio (Om)	LPG (ppm)	Methane (ppm)	Smoke (ppm)	Hydrogen (ppm)	PIR [0;1]
A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**

Таблица 1. Условное обозначение каждой выборки данных МС.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

Для того чтобы провести регрессионный анализ со всеми данными, необходимо сравнить значение каждого столбца последовательно со значениями каждого последующего столбца (таблица 2):

AB	-	-	-	-	-	-	-
AC	**BC**	-	-	-	-	-	-
AD	**BD**	**CD**	-	-	-	-	-
AE	**BE**	**CE**	**DE**	-	-	-	-
AF	**BF**	**CF**	**DF**	**EF**	-	-	-
AG	**BG**	**CG**	**DG**	**EG**	**FG**	-	-
AH	**BH**	**CH**	**DH**	**EH**	**FH**	**GH**	-
AI	**BI**	**CI**	**DI**	**EI**	**FI**	**GI**	**HI**

Таблица 2. Матрица условных обозначений каждой выборки данных МС для последующей обработки.

Уравнение регрессии имеет вид $y = bx + a$. Оценочное уравнение регрессии по выборочным данным будет иметь вид $y = bx + a + \varepsilon$,
где ε_i – случайная ошибка (отклонение), a и b соответственно оценки параметров α и β регрессионной модели, которые следует найти.

Причины существования случайной ошибки: отброс значимых данных, аппроксимация данных с различными параметрами, неверное описание модели, ошибки измерения [7].

Так как ε_i для наблюдения i – случайны и их значения в выборке неизвестны, то по наблюдениям x_i и y_i можно получить оценки параметров α и β регрессионной модели, которыми являются случайные величины a и b, т. к. соответствуют случайной выборке.

Для оценки параметров α и β - используют МНК (метод наименьших квадратов). МНК дает наилучшие результаты, но только в том случае, если выполняются определенные предпосылки относительно случайного члена (ε) и независимой переменной (x).

Формально критерий МНК можно записать так:

$$S = \sum(y_i - yi)2 \to min.$$

Система нормальных уравнений:

$$\begin{cases} an + b\sum x = \sum y \\ a\sum x + b\sum x2 = \sum xy \end{cases}.$$

Следующим этапом строится расчетная таблица для расчета параметров регрессии, далее получаем эмпирическое уравнение регрессии для каждой из выборок:

1) y=0.3026x+781.64 (PhotoPin);
2) y=0.408x+153.99 (MQ2 (вся газовая смесь));
3) y=0.0008x+0.4086 (Ratio);
4) y=13.376x+3645.2 (LPG);
5) y=140.87x+33264 (Methane);
6) y=104.03x+26952 (Smoke);
7) y=16.108x+4747.3 (Hydrogen);
8) y=0.014x-0.3294 (PIR).

Прогнозируемый уровень показывает коэффициент a, но только в тех случаях, когда x находится близко с выборочными значениями, если нет, то буквальная интерпретация данных может привести к неверным результатам, и даже если линия регрессии точно описывает значения наблюдаемой выборки, то это не значит, что аналогично будет при экстраполяции влево или вправо. Подставив в уравнение регрессии соответствующие значения x, можно определить выровненные (предсказанные) значения результативного показателя $y(x)$ для каждого наблюдения [4, 5].

Связь между y и x определяет знак коэффициента регрессии b (если > 0 – прямая связь, иначе - обратная). В нашем примере связь прямая. Далее находим числовое значение регрессии для каждой из пар выборки, а также корреляцию (соответствующие матрицы корреляционных и регрессионных данных представлены в таблице 3, а карты данных корреляции и регрессии представлены на рисунке 2). Карты корреляции и регрессии в данном случае строятся для быстрого визуального анализа применения предлагаемого принципа. На карте корреляции видно, что наибольшее сгущение данных происходит в области водорода, дыма, метана и LPG, карта регрессии подтверждает, что наибольшее значение регрессии наблюдается именно в области углеводородов (пик значений прослеживается в области дыма и метана).

В результате получаем:

Корреляция данных

	Humidity (%)	PhotoPin	MQ-2	Ratio	LPG	Methane	Smoke	Hydrogen	PIR	
Humidity (%)	1									
PhotoPin	0,095	1								
MQ-2	0,864	-0,152	1							
Ratio	-0,099	-0,996	0,143	1						
LPG	0,155	0,984	-0,054	-0,984	1					
Methane	0,153	0,961	-0,031	-0,964	0,992	1				
Smoke	0,179	0,924	0,008	-0,932	0,967	0,989	1			
Hydrogen	0,199	0,892	0,038	-0,903	0,942	0,972	0,995	1		Среднее
PIR	0,061	-0,345	0,036	0,328	-0,373	-0,373	-0,358	-0,331	1	0,310646

Регрессия данных

	Humidity (%)	PhotoPin	MQ-2	Ratio	LPG	Methane	Smoke	Hydrogen	PIR	
Humidity (%)	1									
PhotoPin	7,081	1								
MQ-2	-9,547	-304,682	1							
Ratio	-0,018	0,599	0,101	1						
LPG	-313,008	9988,786	1663,881	-662,824	1					
Methane	-3296,218	105197,630	17522,959	-6980,321	6000671,229	1				
Smoke	-2434,462	77886,323	12940,132	-5155,107	443584,800	4187973,001	1			
Hydrogen	-376,917	12028,981	2003,684	-798,182	68984,683	648465,565	517125,765	1		Среднее
PIR	-0,327	10,455	1,742	-0,693	59,697	563,604	449,452	77,038	1	281078,6

Таблица 3. Матрица корреляционных и регрессионных данных.

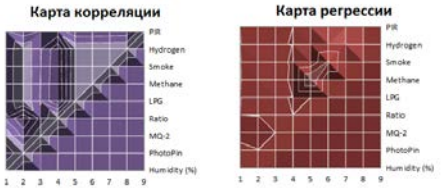

Рис. 2. Карты корреляции и регрессии.

Линейный коэффициент корреляции принимает значения от -1 до +1. Связи между признаками могут быть слабыми и сильными (тесными). Их критерии оцениваются по шкале Чеддока. В данном случае среднее значение корреляции составляет 0,311, т.е. связь между данными слабая, но прямая.

Коэффициент линейной парной корреляции может быть вычислен через коэффициент регрессии b и в среднем составит:

$$r_{x,y} = b\frac{S(x)}{S(y)} = 0.311,$$

где S – среднеквадратичное отклонение.

Общий коэффициент детерминации по всем выборкам в этом случае будет равен:

$$R^2 = 1 - \frac{\sum(y_i - y_x)^2}{\sum(y_i - \bar{y})^2} = 0.431,$$

т.е. на основании выше сказанного можно заключить, что связь между разнородными данными слабая, но прямая, в 43.1% случаев изменения одних данных приводят к изменению других данных. Точность подбора уравнения регрессии - высокая. Остальные 56.9% изменения данных объясняются факторами, не учтенными в модели (а также ошибками спецификации).

Наибольшее влияние в ходе эксперимента прослеживается между значениями углеводородной смеси, метана, дыма и водорода (в общем, порядка 60-70%). Наименьшее значение между данными влажности, сопротивлением МС и значениями инфракрасного датчика присутствия (ИКД).

Исходя из рассмотренного примера, можно выдвинуть следующую гипотезу: во многих прикладных задачах, требующих больших объемов разнородных данных для нахождения решения приемлемой точности, есть возможность получить то же решение даже по информации от существенно ограниченного набора датчиков МС при условии правильной интерпретации имеющихся данных и

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

нахождении новых, неявных зависимостей между данными. Другими словами, если с помощью имеющейся МС необходимо решить практическую задачу на основании лишь двух-трех выборок значений данных с газоанализатора, то эту задачу можно решить без установки дорогостоящего дополнительного оборудования. Решение было получено по ограниченному набору датчиков, в данном случае удалось обойтись без датчика температуры, а также доказано, что значения ИКД и влажности незначительно влияют на результат.

Полученные в ходе эксперимента теоретические результаты и формулируемые на их основе выводы подтверждаются строгостью математических выкладок, базирующихся на аппарате интегрального и дифференциального исчисления, теории вероятностей и математической статистики. Справедливость выводов относительно эффективности предложенной системы подтверждена статистическим моделированием и опытно-методической обработкой реальных результатов.

4. Заключение

В работе предложен редуктивно-аккомодативный принцип обработки гетерогенных данных мультисенсорной системы. Другими словами предлагается некий унифицированный подход к проблеме разработки высокопроизводительных систем обработки многомерных разнородных данных под конкретную прикладную задачу и заданные требования. В дальнейшем разработанная система может легко масштабироваться различными наборами данных для применения в конкретных областях технических задач.

Изучена зависимость одной выборки данных от других. Для расчетов была выбрана парная линейная регрессия. Оценены её параметры методом наименьших квадратов. Статистическая значимость уравнения проверена с помощью коэффициента детерминации. Установлено, что в исследуемой ситуации 0.9% общей вариабельности одних данных объясняется изменением других данных. Установлено также, что параметры модели статистически не значимы.

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 классификация способствует быстрому оптимальному поиску решения задач, т.к. позволяет единовременно увидеть всю картину существующих связей.

Результаты исследования применимы в сфере мониторинга и обработки разнородных данных для получения быстрой информации о согласованности данных с целью получения значимой информации, на основании которой можно быстро принять правильное решение. Данные исследования могут быть использованы при моделировании ситуаций, требующих быстрого реагирования, таких как: осуществление ликвидации аварий, моделирование эвакуации людей из зданий в чрезвычайных ситуациях, моделирование ситуаций при террористических атаках.

5. Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-41-700001 р_а.

1. Литература

[1] 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/10.1088/1742-6596/803/1/012014

[2] Kashnikov A., Lyadova L. Integration of heterogeneous sources of data based on recurrent decomposition / International Journal "Information Technologies & Knowledge" Vol.5, Number 3, 2011, P.274-284.

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

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

[5] Witten, I. H., & Frank, E. (2011). Data Mining: Practical Machine Learning Tools and Techniques.

[6] 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

[7] Буре В.М. Евсеев Е.А. Основы эконометрики: Учеб. Пособие. - СПб.: Изд-во С.-Петерб. ун-та, 2004.- 72 с.

[8] 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

[9] Клеменков П.А., Кузнецов С.Д. Большие данные: современные подходы к хранению и обработке. //Труды ИСП РАН. Том: 23. 2012. - С. 143-158

[10] Лысенко Н.В., Мончак А.М. Анализ эффективности гетерогенных видеоинформационных систем на основе критерия доминирования. / Радиоэлектроника. 2018. С. 57-62.

[11] 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.

[12] Скобло Т.С., Клочко О.Ю., Белкин Е.Л., Сидашенко А.И. Новые подходы в исследовании неоднородности гетерогенных структур / Металлофизика и новейшие технологии, 2018.Т. 40, № 2. — С. 255-280.

[13] Финогеев А.А., Финогеев А.Г., Нефедова И.С. Технология конвергентной обработки данных в защищенной сети системы мониторинга/ Фундаментальные исследования. – 2015. – № 11-5. – С. 923-927

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

2. Об авторах

Багутдинов Равиль Анатольевич – аспирант, программист ОАР ИШИТР НИ ТПУ. E-mail: ravil_bagutdinov@yahoo.com.

Небаба Степан Геннадьевич – к.т.н., инженер Научно-образовательной лаборатории 3DM и ПД ОАР ИШИТР НИ ТПУ. E-mail: stepanlfx@tpu.ru.

Захарова Алена Александровна – профессор, д.т.н., зав. Научно-образовательной лаборатории 3D-моделирования и промышленного дизайна ОАР ИШИТР НИ ТПУ. E-mail: zaa@tpu.ru.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18