

УДК 004.272.43+004.272.32

М.М. Татур

ОСОБЕННОСТИ ПОСТРОЕНИЯ ВЫЧИСЛИТЕЛЕЙ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ

Излагается подход к построению параллельных машин, проблемно-ориентированных на широкий круг задач интеллектуальной обработки данных. Представляется архитектура верхнего уровня, где в качестве функциональных единиц выступают алгоритмы Data Mining, а в качестве ускорителя – универсальный либо специализированный сопроцессор. Делается акцент на необходимость обеспечения совместимости библиотечных алгоритмов и архитектуры параллельного сопроцессора. Приводится методика корректной сравнительной оценки альтернативных аппаратных платформ в рамках предложенного подхода.

Введение

Современная теория интеллектуального анализа данных представлена широким перечнем математических методов и алгоритмов, позволяющих решать задачи кластеризации, классификации, ассоциативного поиска и др., получать количественные оценки эффективности решения таких задач. В ряде работ вскрыты принципиальные ограничения, присущие известным методам, и предложены возможные пути их преодоления [1, 2]. Однако практические достижения в области построения интеллектуальных систем остаются достаточно скромными: созданные системы назвать интеллектуальными можно лишь с большой натяжкой, а время, трудоемкость и стоимость разработки интеллектуальных систем продолжают оставаться неприемлемо большими. В настоящей работе предлагается следующая версия объяснения такого отставания в практике построения интеллектуальных систем.

Большинство актуальных прикладных задач выходят за рамки отдельно взятого формального алгоритма классификации, регрессии и т. п. Существует определенная проблема представления решаемой прикладной задачи как композиции формальных алгоритмов интеллектуального (и не только) анализа данных. Поэтому решения прикладных задач зачастую выглядят либо попыткой «лобового» применения одного из методов, известных инженеру-разработчику, либо «гремучей смеси» различных алгоритмов и эвристик. Крайне редко встречаются действительно научно и технически обоснованные композиции методов и алгоритмов, позволяющих получить гармоничное, эффективное решение прикладной задачи.

Еще с большей очевидностью дисбаланс в достижениях теории и практики проявляется в разработке высокопроизводительных аппаратных средств поддержки интеллектуальных вычислений. Это связано с тем, что создание специализированных аппаратных средств идет вслед за алгоритмическим проектированием и поэтому на порядок сложнее и затратнее во всех отношениях. Ярким примером этому является судьба нейрокомпьютеров и нейрочипов, которым пророчили большое будущее. Между тем оказалось, что помимо чисто технических ограничений (число связей, число одновременно моделируемых нейронов) модель нейронных сетей далеко не всегда обеспечивает эффективное решение реальных интеллектуальных задач. В итоге нейрокомпьютеры вроде бы есть, а покупателей нейрокомпьютеров, готовых применить их в создаваемых системах, – нет [3, 4]. Под данное утверждение не подпадают те случаи, когда отдельные производители электронных компонентов свои, по сути, сигнальные процессоры (DSP) называли нейрочипами, поскольку и в первом и во втором случаях в качестве массовой операции выступает взвешенное суммирование [5].

Между тем новый технологический виток, связанный с развитием многоядерных CPU, GPU, суперкомпьютеров, кластеров, распределенных и облачных вычислений, по сути унифицированных параллельных вычислительных архитектур, составил дополнительную конкуренцию специализированным аппаратным средствам поддержки интеллектуальных вычислений. Возникает резонный вопрос, что же может явиться альтернативой «жестким» нейронным, ассоциативным, семантическим и тому подобным процессорам, с одной стороны, и универсальным

параллельным компьютерам – с другой. Может быть, это будут интеллектуальные, или когнитивные, компьютеры, проблемно-ориентированные на широкий круг задач интеллектуального анализа данных?

1. Спецификации функций и архитектура верхнего уровня

Ограничим исследование вопросом уровнем электронных и компьютерных технологий, доступных в настоящее время в Беларуси (а также в России), в разумной ценовой категории. Это значит, что не будут рассматриваться области «живой» элементной базы, самоорганизующихся синергетических архитектур, трехмерных чипов, нано-, квантовых, фотонных и тому подобных компьютеров. Целью настоящей работы является определение прагматичного подхода, который откроет путь к созданию эффективных программно-аппаратных комплексов как основы для построения интеллектуальных систем. Предлагаемые идеи и пути решения являются дальнейшим развитием авторских результатов (математических моделей нечеткой классификации, семантической обработки знаний и их программно-аппаратных реализаций) [6–8] в направлении унификации.

В настоящей работе предлагается оригинальная архитектура проблемно-ориентированной машины, которая представляет собой надстройку над архитектурой обычного параллельного либо специализированного компьютера. Архитектура включает следующие составные части: центральный процессор или супервизор, параллельный сопроцессор в качестве ускорителя, а также библиотеку программ интеллектуальной обработки данных (рис. 1). Связка центральный процессор – сопроцессор является общеизвестной и широко применяемой, а отличительные особенности заключаются в появлении библиотеки программ алгоритмов. (Конструктивные варианты реализации вычислительной системы могут быть различными и в настоящей работе не рассматриваются.)

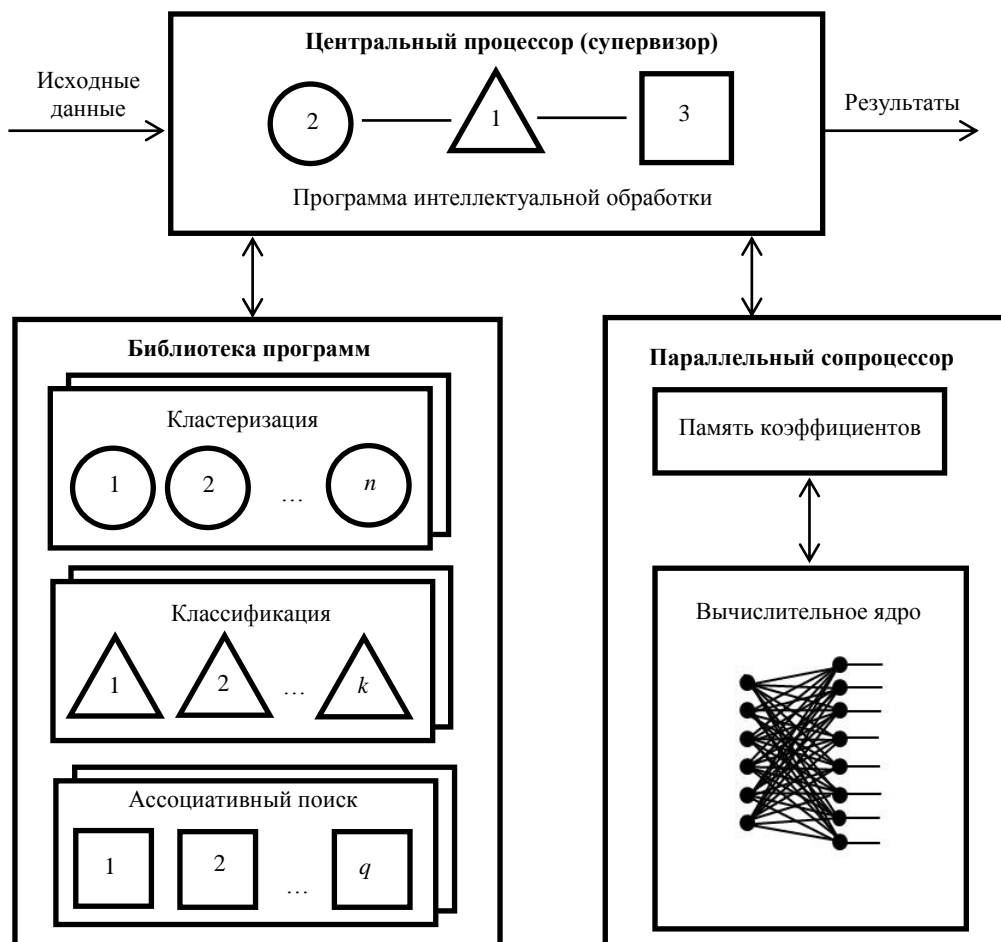


Рис. 1. Обобщенная схема интеллектуального, или когнитивного, суперкомпьютера

Концептуально предлагаемый подход реализует принцип программного управления высокого уровня. Полагается, что прикладная задача интеллектуальной обработки данных может быть представлена как композиция формальных задач кластеризации, классификации, регрессии, ассоциативного поиска, семантической обработки и некоторых других, включая задачи математической статистики. Для каждой из задач известен перечень алгоритмов реализации, причем большинство из них являются переборными, а значит, вычислительная сложность возрастает нелинейно в зависимости от объема входных данных. Каждый из алгоритмов реализован в виде программы, оптимизированной под конкретную аппаратную реализацию параллельного сопроцессора. Также в библиотеке содержатся программы алгоритмов оценки эффективности решения задач кластеризации, классификации и др.

Головная программа интеллектуальной обработки запускается в центральном процессоре и состоит из программ отдельных алгоритмов из библиотеки, а также вспомогательных команд, не включенных в библиотеку. На рис. 1 показан упрощенный пример такой программы, состоящей из алгоритма кластеризации 2, алгоритма классификации 1 и алгоритма ассоциативного поиска 3. Эта же программа осуществляет прием исходных данных, организацию ускорения вычислений на параллельном сопроцессоре и выдачу результатов обработки.

С функциональной точки зрения анонсируемая вычислительная система напоминает программные системы Wolfram Mathematica и Weka [9, 10]. Отличие состоит в применении параллельного универсального сопроцессора, например кластера GPU, и адаптации алгоритмов под его архитектуру. Для каждого алгоритма библиотеки будет определен расчетный показатель эффективности параллельной реализации на сопроцессоре, а это значит, что в головной программе, как правило, будут применяться алгоритмы с высокими показателями эффективности.

В случае оригинальной программно-аппаратной реализации сопроцессора следует ожидать более высокую степень согласованности алгоритмов библиотеки и архитектуры параллельного спецвычислителя, а в результате и эффективность вычислительной системы в целом.

2. Сравнение эффективности аппаратных платформ при построении интеллектуальных компьютеров

Общепринято под эффективностью понимать отношение достигаемого качества решения задачи к стоимости или другим затраченным ресурсам. Применительно к компьютерам это традиционно отношение производительности к стоимости или объему оборудования. Из опыта оценки эффективности нейрокомпьютеров известно, что оценивать производительность в флоспах оказалось не совсем корректным и поэтому были предложены другие меры, например такие, как число моделируемых синапсов (умножений со сложением), нейронов (то же плюс пороговая функция) в единицу времени. В ряде работ по нейронным сетям заявляется, что они по определению предполагают эффективную параллельную реализацию. Однако при этом обычно опускается немаловажный вопрос, сколько нейронов могут моделироваться одновременно. Простейшие эксперименты по моделированию нейронных сетей на параллельных компьютерах показывают, что закономерности Амдаля – Густавсона распространяются и на параллельную реализацию нейросетей, т. е. увеличение числа процессорных элементов не приводит к линейному росту производительности вычислений в целом. Это и понятно, так как проблема распараллеливания при моделировании имеет отношение не только к параллельной компьютерной архитектуре, но и к реализуемой модели.

Какую эффективность можно ожидать от проблемно-ориентированного когнитивного компьютера? Чтобы ответить на этот вопрос, сначала укажем и прокомментируем параметры, которые в той или иной мере определяют эффективность: время вычисления тестовой задачи, объем входных данных тестовой задачи, точность вычисления тестовой задачи и число процессорных элементов. Стоимость, естественно, является интегральным параметром, характеризующим эффективность вычислительных машин. Однако в настоящей работе стоимость как параметр рассматриваться не будет, а будет полагаться, что число процессорных элементов характеризует общие затраты ресурсов на достижение определенной производительности.

Объем входных (обрабатываемых) данных – это величина, которая определяется числом образов, числом классов и (или) кластеров, числом информативных признаков (размерностью

пространства признаков). Для упрощения рассуждений абстрагируемся от влияния разрядности и форм представления данных в вычислительной машине на объем вычислений.

Точность вычислений (в нашем случае качество, достоверность принятия решений) – это величина, которая определяется математическим методом и алгоритмом вычислений, репрезентативностью обучающей выборки, критерием принятия решений и т. п. Для упрощения рассуждений абстрагируемся от влияния разрядности и форм представления данных в вычислительной машине на точность вычислений.

Число процессорных элементов характеризует напрямую потенциальную возможность распараллеливания вычислительного процесса, а косвенно, как было отмечено выше, – стоимость вычислительной системы.

Время вычисления тестовой задачи (имеется в виду время кластеризации, классификации, ассоциации и т. п. или их композиции) определяется с момента подачи входных данных до получения результата. Этот параметр может измеряться как в абсолютных величинах, так и в условных – модельных тактах. В первом случае на оценку будут влиять технические характеристики аппаратной платформы (тактовая частота процессора, время доступа к данным в оперативной памяти и др.), во втором случае – способ определения модельного времени. Заметим, что производительность будет представлена не в гига-, терафлопсах и не в кило-, мега-нейронах, а в задачах кластеризации и (или) классификации и т. д. в единицу времени. Время будет складываться как из времени непосредственных вычислений на параллельном сопроцессоре, так и времени «накладных расходов», связанных с загрузкой либо выгрузкой данных сопроцессора, а также реализации участков программ, не поддающихся распараллеливанию, подготовительных операций, сервисных функций и т. п.

Влияние указанных параметров друг на друга будет характеризовать эффективность интеллектуальных вычислений. Графики качественных зависимостей такого влияния для различных аппаратных платформ показаны на рис. 2.

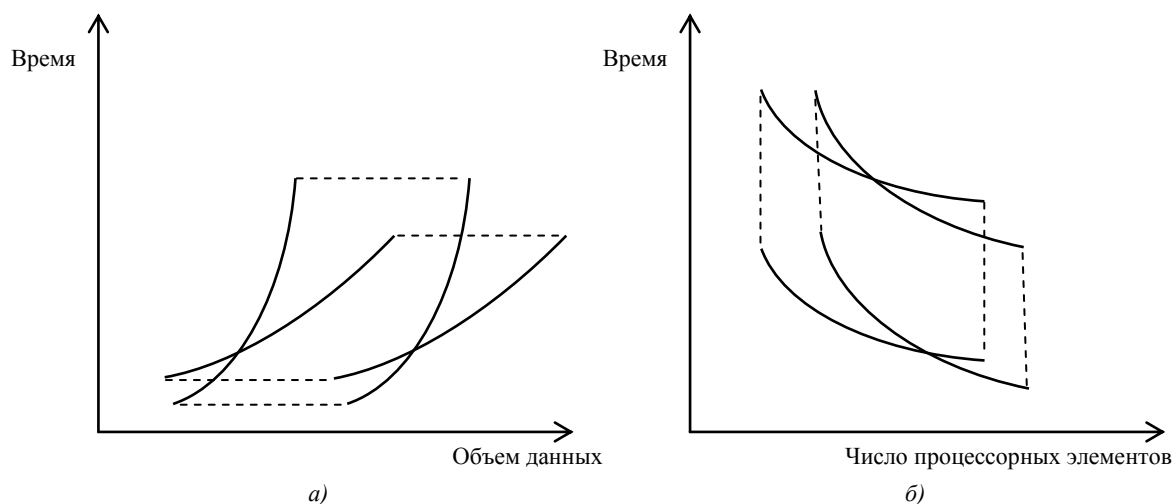


Рис. 2. Качественные зависимости времени вычислений тестовой задачи: а) от объема обрабатываемых данных; б) от числа процессорных элементов

На рис. 2, а изображен вполне ожидаемый вид экспоненциальной зависимости, поскольку задачи интеллектуального анализа данных носят переборный характер. Однако характер наклона может существенно изменяться в зависимости от того, насколько качественно выполнено распараллеливание алгоритмов применительно к конкретной аппаратной платформе или насколько удачно спроектирована архитектура проблемно-ориентированной машины. На рис. 2, б показано проявление закона Амдаля – Густафсона, согласно которому производительность параллельной машины не может возрастать линейно с увеличением числа процессорных элементов. Между тем за счет предлагаемого подхода к построению когнитивных компьютеров наклон этой кривой можно значительно изменить (снизить в первом случае и повысить

во втором). Для обоих графиков полагается, что другие существенные параметры (точность, объем данных или число процессорных элементов) принимаются постоянными.

На рис. 2 одноименные кривые образуют некоторый диапазон пространства, отмеченный пунктиром, в котором может проходить конкретная кривая, а ее реальное положение будет определено техническими характеристиками аппаратной платформы. Важно отметить, что графики, характеризующие эффективность различных альтернативных платформ, могут пересекаться и это будет крайне полезной информацией при системном проектировании вычислительной машины в целом.

Заключение

В работе в качестве формальных задач Data Mining рассматриваются кластеризация, классификация, регрессия и ассоциативный поиск. В качестве исходных данных выступают образы, представленные в виде многомерных численных векторов информативных признаков. Для каждой из задач известен широкий перечень алгоритмов решения, которые имеют принципиальные отличия. Так, например, одни связаны с вычислением расстояний между образами в определенной метрике, другие – с аппроксимацией дискриминантной или регрессионной функций, третьи – с анализом и обработкой графов. Эти различия обуславливают очень узкую специфику применения известных нейрокомпьютеров и ассоциативных процессоров.

Предложен подход, который позволяет унифицировать программно-аппаратную реализацию задач регрессии, классификации, кластеризации и ассоциативного поиска посредством приведения различных алгоритмов к единой аппаратной платформе. В качестве аппаратной платформы может выступать как универсальная параллельная машина (например, многоядерная CPU, GPU, суперЭВМ), так и специализированная. В любом случае необходимо обеспечить эффективную реализацию библиотечных алгоритмов интеллектуальной обработки на параллельном сопроцессоре. Это открывает путь к созданию аппаратного вычислительного ядра, общего для широкого круга задач Data Mining и Knowledge Discovery, а следовательно, к созданию нового типа компьютеров. Предлагаемая модель вычислительной машины является более универсальной по сравнению с нейронными, ассоциативными, графовыми, семантическими и некоторыми другими, так как ориентирована на более высокий уровень представления формальных математических задач. Поэтому для ее определения использован термин «когнитивная», или «интеллектуальная», ЭВМ в связи с тем, что она охватывает широкий круг задач из области интеллектуального анализа данных.

Предложенный подход позволяет осуществить корректную сравнительную оценку альтернативных аппаратных средств, применяемых для построения интеллектуальных систем. Для этого в качестве функциональной единицы необходимо использовать алгоритм реализации одной из задач – кластеризации, классификации, ассоциативного поиска или их комбинации.

Список литературы

1. Data Mining: A Knowledge Discovery Approach / K.J. Cios [et al.]. – Springer, 2007.
2. Том, И.Э. Методы интеллектуального анализа многомерных данных для решения задач классификации / И.Э. Том, Н.А. Новоселова, О.В. Красько. – Минск : ОИПИ НАН Беларуси, 2011. – 233 с.
3. Кирсанов, Э.Ю. Нейрокомпьютеры с параллельной архитектурой / Э.Ю. Кирсанов. – М. : Радиотехника, 2004. – 221 с.
4. Аляутдинов, М.А. Нейрокомпьютеры: от программной к аппаратной реализации / М.А. Аляутдинов, А.И. Галушкин, П.А. Казанцев. – М. : Горячая линия – Телеком, 2008. – 152 с.
5. Реализация искусственных нейронных сетей в НТЦ «Модуль» // Компоненты и технологии. – 2005. – № 4. – С. 98–102.
6. Synthesis and Analysis of Classifiers Based on Generalized Model of Identification / D. Adzinets [et al.] // Advances in intelligent and soft computing. – Springer, 2010. – Vol. 71. – P. 529–536.
7. Tatur, M. Problem-Oriented Processors for the Solving of Classification Tasks / M. Tatur // J. of Information, Control and Management Systems. – 2013. – Vol. 11, no. 2. – P. 155–164.

8. Cognitive information processing based on a parallel processor / N. Verenik [et al.] // Proc. of 10th Intern. Conf. on Digital Technologies. – Žilina, 2014. – P. 367–371.

9. Wolfram Mathematica [Electronic resource]. – Mode of access : <http://www.wolfram.com/mathematica>. – Date of access : 12.01.2015.

10. Weka 3: Data Mining Software in Java [Electronic resource]. – Mode of access : <http://www.cs.waikato.ac.nz/ml/weka>. – Date of access : 12.01.2015.

Поступила 29.01.2015

*Белорусский государственный университет
информатики и радиоэлектроники,
Минск, ул. П. Бровки, 6
e-mail: tatur@bsuir.by*

M.M. Tatur

CONSTRUCTION PRINCIPLES OF COMPUTING UNITS FOR INTELLECTUAL DATA PROCESSING

An approach to the development of problem-oriented parallel computers for a wide range of tasks of intelligent data processing is described. A high level architecture is shown, where Data Mining algorithms are considered as functional elements and a general or specialized co-processor as an accelerator for computer performance. An accent has been made on the necessity to provide compatibility of the librarian algorithms and the architecture of the parallel co-processor. The technique for correct comparative evaluation of the alternative hardware platforms in the framework of the proposed approach is presented.