

УДК 004

РАЗРАБОТКА И ИССЛЕДОВАНИЕ ИТЕРАТИВНОГО ПАРАЛЛЕЛЬНОГО АЛГОРИТМА ФОРМИРОВАНИЯ СКРЫТЫХ ФУНКЦИОНАЛЬНО-ДЕТЕРМИНИРОВАННЫХ СТРУКТУР ДЛЯ КЛАССИФИКАЦИИ И АНАЛИЗА МЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ

Сорокин В.А., Рохмистров Д.С., Ботыгин И.А.

Научный руководитель: Ботыгин И.А.

*Национальный Исследовательский Томский политехнический университет,**634050, Россия, г. Томск, пр. Ленина, 30**E-mail: qwertvasilii@gmail.com*

An algorithm of allocation structures in the initial geophysical field, allowing for the spatial characteristics of the test decompose space into regions with a similar behavior of the studied parameters. The algorithm is based on the structuring of the various expansions (season, anomaly, and so on. D.) Geophysical fields, and provides a variety of information about the object as a parameter set selected structures.

Key words: *cluster analysis, parallel computing, meteorological data, big data.*

Ключевые слова: *кластерный анализ, параллельные вычисления, метеорологические данные, больше данные.*

Кластерный анализ – задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались [1]. Применение кластерного анализа в общем виде сводится к следующим этапам:

- отбор выборки объектов для кластеризации;
- определение множества переменных, по которым будут оцениваться объекты в выборке;
- вычисление значений меры сходства между объектами;
- применение метода кластерного анализа для создания групп сходных объектов;
- представление результатов анализа [2].

Предметом исследования в данной статье является набор данных среднемесячной температуры, полученных за 55 лет на 333 метеостанциях. Среднемесячные изменения таких данных за исследуемый период формируют колебательный процесс с квазипериодом в один год. Для описания характера погоды представляют интерес температурные изменения, отклоняющиеся от годового цикла. Однако их непротиворечивое выделение затруднено [3].

Для расчета фаз каждого температурного ряда был реализован алгоритм: входные данные в виде температурного ряда интерполируются сплайном до размера массива равному степени двойки (функция Spline), далее применяется алгоритм быстрого преобразования Фурье (FFT), далее используется фильтрация окном для отсеивания помех (Filtration) и применяется обратное преобразование Фурье (IFFT), далее с помощью преобразования, основанного на арктангенсе отношения действительной части к мнимой части полученного аналитического сигнала получается фаза ряда (Tangent), полученная фаза на данном этапе представляет собой главные значения функции \arctg в пределах годового периода, следующий этап алгоритма – вычисление непрерывной функции фазы на всем интервале наблюдения (Linking), затем из полученной функции вычитается линейная составляющая (LineDiff), после чего на выходе получают фазовые флуктуации данного температурного ряда.

На рис. 1 приведена общая блок-схема расчета типовых фаз. Вычисленные фазы для каждой из групп подаются на вход узла CorrTable, где высчитывается попарный коэффициент корреляции. На следующем этапе полученная таблица корреляции фаз передается в узел CorrComparison. Здесь формируются группы из тех температурных рядов, которые имеют с формирующим группу рядом коэффициенты корреляции выше заданного уровня g . На основе сформированных групп в узле AvPhasesCalc рассчитываются типовые фазы текущей итерации как среднеарифметическая оценка первого уровня.

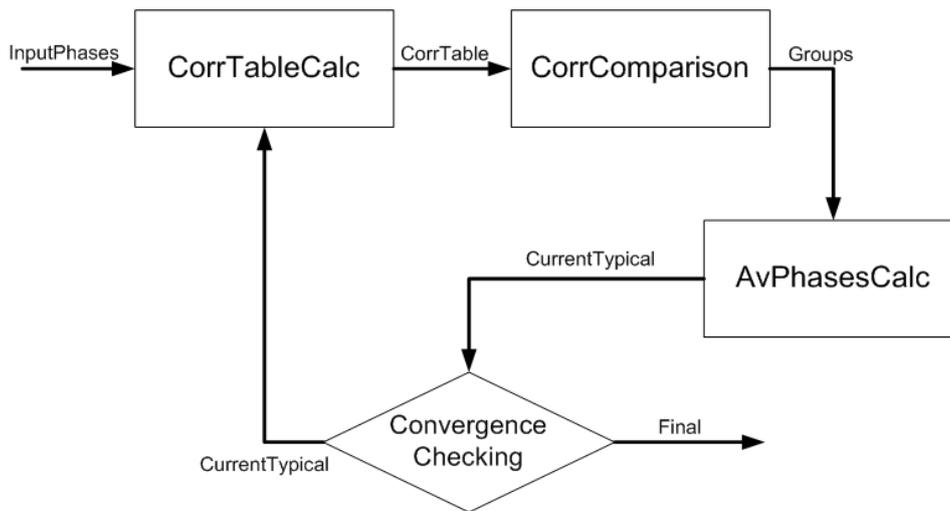


Рис. 1. Общий алгоритм расчета типовых фаз

Все вычисления алгоритма делятся на два этапа: подготовительная часть – расчет фазы температурных рядов, основная часть – фазовая группировка.

В общем случае параллельное вычисление, основанное на модели MapReduce, представлено на рис. 2.

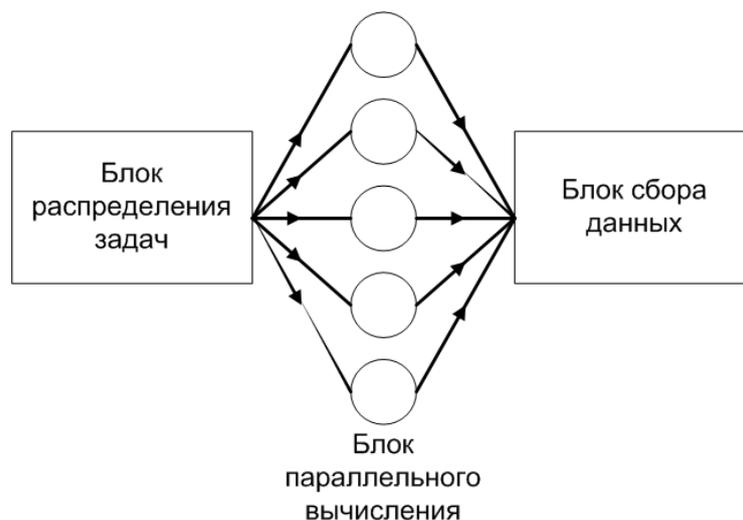


Рис. 2. Общая блок-схема параллельного расчета

Таким образом блок распределения задач распределяет входные данные по доступным мощностям для параллельного расчета фазы температурных рядов, после чего полученные результаты на каждом узле собираются в блоке сбора данных и подготавливаются для дальнейшего анализа.

Для параллельного расчета фазовой группировки, так как этот процесс (рис. 1) итеративен, на каждой итерации используются данные, полученные на предыдущем этапе, что значит, что параллельное выполнение всего блока невозможно.

На данном этапе представляется возможным параллельное вычисление внутри алгоритма. Выделим на блок-схеме фазовой группировки этапы, подходящие для распределения цветом (рис. 3), в каждом таком блоке подразумевается использование общей схемы параллельного расчета (рис. 2).

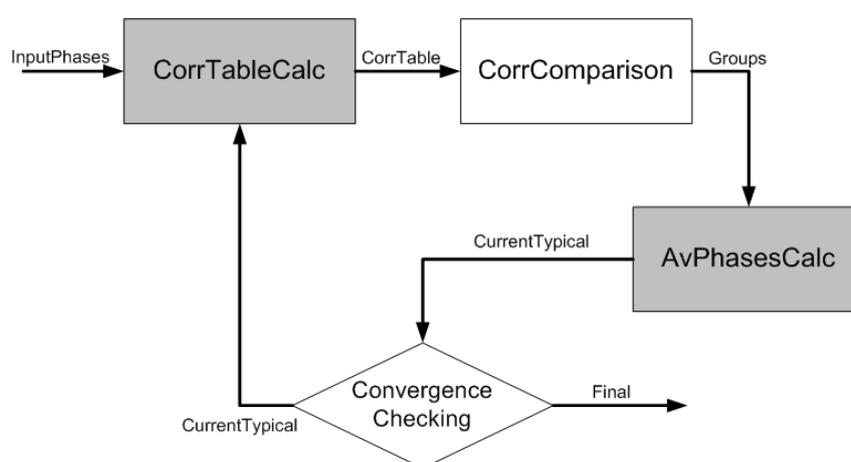


Рис. 3. Распараллеливание этапа расчета фазовой группировки

Таким образом, на каждой итерации расчет таблицы корреляции может быть распределен между доступными вычислительными мощностями, затем, после сбора полученные данных в единую таблицу, на основе анализа итоговой таблицы возможно параллельное вычисление типовых фаз.

Список литературы

1. Кластеризация // MachineLearning.ru URL: <http://www.machinelearning.ru/wiki/index.php?title=Кластеризация> (дата обращения: 26.02.2016).
2. Обзор алгоритмов кластеризации данных // Хабрахабр URL: <https://habrahabr.ru/post/101338/> (дата обращения: 26.02.2016).
3. Волков Ю.В., Тартаковский В.А. Региональная группировка данных, полученных на метеостанциях Евразии // Энергетика В.А. Эффективность, надежность, безопасность. – Томск: ООО «СКАН», 2013. – С. 177–178.
4. Вакман Д.Е., Вайнштейн Л.А. Амплитуда, фазы частота – основные понятия теории колебаний // Успехи физических наук. – 1977. – № 123. – Вып. 4. – С. 657.