International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 10 180 - 183

Method of Building Parallel Data Mining Algorithms by Thread-Safe Functional Blocks

Karshiev Zaynidin Abduvaliyevich
Department of Computer Systems
Samarkand branch of TUIT
Samarkand, Uzbekistan
zaynidin@gmail.com

Kupriyanov Mikhail Stepanovich
Department of Computer Science and Engineering
Saint-Petersburg Electrotechnical University
Saint-Petersburg, Russian Federation
mskupriyanov@mail.ru

***Abstract*—**In this article described method of constructing parallel data mining algorithms based on representation of an algorithm in the form of hierarchy of nested functional blocks. Offered method allows to build types of parallel structures of data mining algorithms with dispatcher and without, with parallelization by task and by data.

Keywords- *Data mining, parallel algorithms, data mining in distributed data, data parallelization, task parallelization.*

*****

I. INTRODUCTION

In accordance with the formal model [1] data mining algorithm can be decomposed into separate thread-safe functional blocks. Each block should receive the input of the processed data, the algorithm settings and a model of knowledge created in the early stages. This realizes the following principle: data and settings cannot be changed; only model of knowledge can be changed. Changing model of knowledge can be seen as the establishment within a block of a new model of knowledge based on the parameters passed to it: the dataset, the algorithm settings, and initial model of knowledge. Thus, each block is a "monolith" in terms of the basic algorithm. In fact, it is a mini-algorithm.

II. MAIN PART

When splitting algorithm into blocks necessary to observe the following requirements:

- algorithm is an ordered sequence of functional blocks that meet the conditions described below;
- each block can be either an operation performed on a model or a some structural element characteristic of data mining algorithms;
- block that implements an operation, should change the model of knowledge, applied to the input so that it remains an integral (i.e., must be performed by all the restrictions imposed on the model);
- inside the functional block should not be an external function call, all the work must be carried out on the basis of: data set, settings, and initial model of knowledge transferred to the functional block.

The functional block can contain executable code (indivisible operation), and the sequence of the other functional blocks. In general block can contain more than one sequence.

As a result of data mining algorithms can be represented as a hierarchy of nested functional blocks (Figure 1). At the same time algorithm itself, as a single functional block and is thread-safe.

It is important to note that if a functional block includes a sequence of nested thread-safe blocks, the entire block as a whole will be thread-safe. For example, the entire block b_i is thread-safe, (Figure 1), and not part of the sequence of its constituent blocks.

Let's consider the functional blocks of which data mining algorithms can consist [2].

Linear block performs some operation on the knowledge model: adding or deleting or changing. It does not contain nested blocks.

Cycle block contains a nested sequence of blocks and determines the operations performed during the initialization cycle, before and after each iteration, as well as the condition of exiting the cycle. In a software implementation, these functions must be performed in the following order:

- cycle initialization;
- check of a condition of loop termination;
- function executed before iteration;
- execution of nested sequence of blocks;
- function executed after iteration.

Decision block contains a checked condition and two sequences of blocks. By results of check of a condition this or that nested sequence of blocks is executed.

As it was marked above, one of the main advantages of the division of data mining algorithm into thread-safe functional blocks is that possibility of their parallel execution. For this purpose, functional blocks providing parallelization of data mining algorithms must be added. Let’s consider what blocks

are necessary for implementing various types of parallel execution.

Parallel block may contain one or more sequences of blocks. In case of parallelization by data it contains one sequence of blocks which must be cloned on several parallel processors. In case of parallelization by task it contains several sequences of blocks (one for each branch of algorithm). In addition this block can contain a nested sequence of blocks for scheduling of parallel blocks. The unified interface of this block is to describe the performance of split() and join() operation. This execution order should be as follows:

- execution operations of separation into parallel blocks (split);
- launch of the controller block (if any);
- launch of the parallel blocks;
- execution of join operations (join).

Sender block includes nested block executing data preparation for sending. Unified interface should describe data sending operation. Sequence of execution shall be following:

- execution of the nested block;

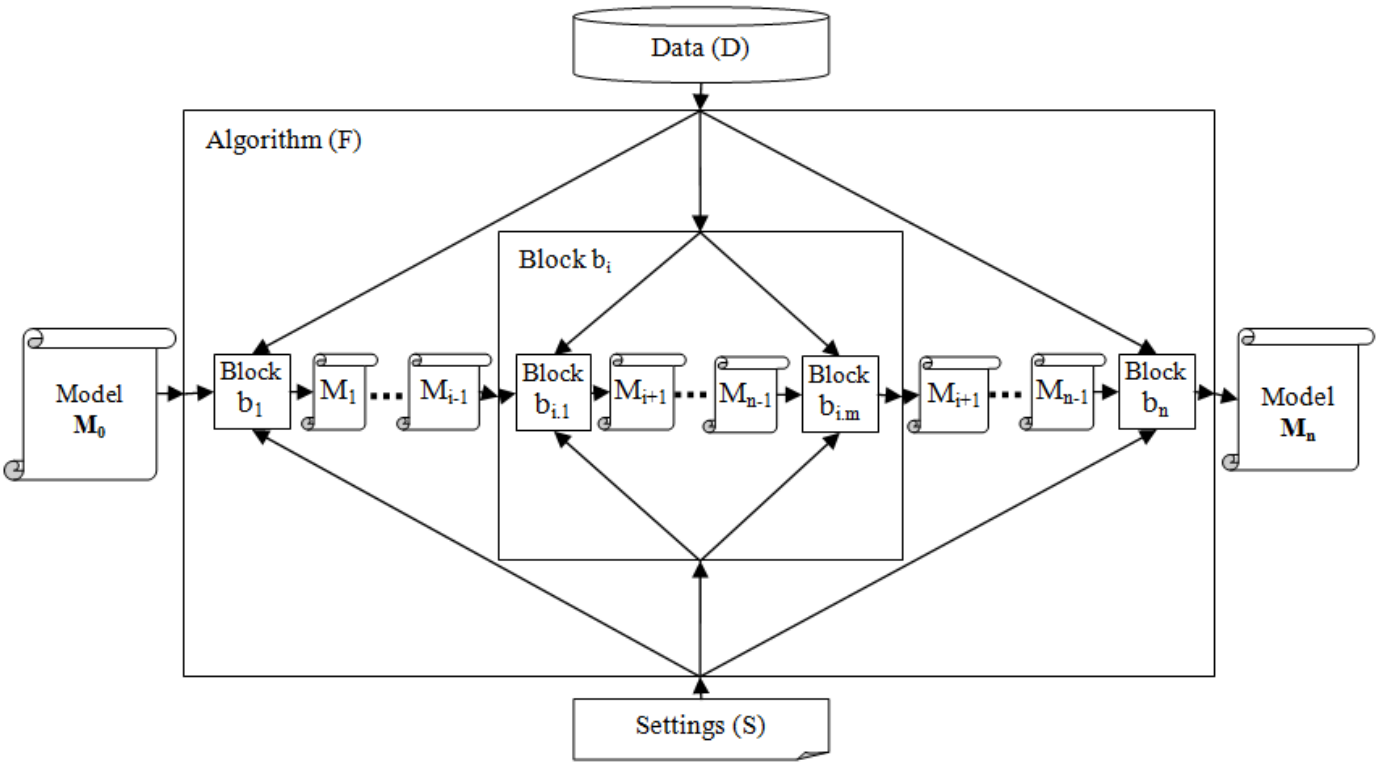

Figure 1. Block structure of data mining algorithm

- execution of sending operation.

Receiving block - includes the nested block executing data handling after their receiving. Unified interface should describe the operation of obtaining data. Sequence of execution shall be following:

- receiving operation execution;
- execution of the nested block.

The main objective of the formation of the hierarchical structure of data mining algorithms presented in Figure 1, is a decomposition algorithm thread-safe functional blocks. Extreme cases of such decomposition are:

- monolithic algorithm – when decomposed data mining algorithm on the composite thread-safe functional blocks failed and algorithm is represented one thread-safe functional block;
- loosely coupled algorithm – when each operation algorithm can be represented by a separate thread-safe functional block.

For execution of decomposition of an algorithm of data mining on thread-safe functional blocks, it is necessary to perform the following operations:

1. decompose data mining algorithm on elementary operations;
2. group into functional blocks of the actions executing the discrete change in the model of knowledge;
3. select structural elements and nested functional blocks in their sequence;
4. add functional blocks structure for parallel execution.

In the first step of the algorithm structure must be decomposed into elementary operations performed on the knowledge model: add, delete and change the individual elements of knowledge model. At the same time in structure of an algorithm the main controlling operators shall be obviously selected: cycles, conditional branches, etc.

Decomposition must be executed taking into account the following conditions:

- after each operation saved the integrity of the model of knowledge;
- the number of variables used in the different operations should be minimized and localized (variables to be used for successive operations only).

On this step allowed existence of the elementary operations which don't changing model of knowledge.

On the second step executed grouping of actions into functional blocks executing the discrete change of knowledge model. Grouping should be performed based on the set of rules:

- In the functional block the operations connected among themselves by the general local variables are selected. The basic blocks are in such operations change the state of knowledge model, to which are added prior and / or subsequent operations, do not change the model. Grouping operations do not change the model, with the adjacent basic operations carried out on the basis of their logical accessories. An indication of such a logical connection is the use of the intermediate (local) variables. These variables must be located within the generated functional block.
- Allocated functional block must be fully included in the control statement (condition, cycle, etc.). For example, one functional block should not be a running operation in a loop and is running cycle.

On the third step structural blocks are created. Control statements (loops, conditional branches, etc.) does not change the state of the model, but they must be separated into individual functional blocks. They should include functional blocks, which operations are performed within the control statements.

Further for creation of a parallel data mining algorithm the structural blocks providing parallel execution can be entered into the received structure:

- parallel block;
- sender block;
- receiver block.

Considering property of thread-safety of the received functional blocks, the parallel block can be added anywhere in the hierarchy. At the same time the functional blocks entering the parallel block as the nested sequence should be defined. The main condition of such entrance is that all of the functional blocks included in the parallel block must have the same level of hierarchy.

Parallel block can include also functional blocks of the most top level of hierarchy, i.e. in this case all algorithm will be executed parallely. At the same time the blocks which are under it will be enclosed in the parallel block.

Sender and receiver blocks need to be added together in parallel sequences of functional blocks, nested in a parallel block, if required information exchange between these sequences.

The method of creation of parallel data mining algorithms described above allows to receive different types of parallel structures of data mining algorithms:

- with parallelization by task;
- with parallelization by data;
- with interaction between parallel branches for distributed memory systems;
- without interaction between parallel branches for distributed memory systems;
- with manager;
- without manager.

To build a data mining algorithm with dispatcher should be allocated to the corresponding sequence of the parallel block and existing functional blocks of the algorithm, or add a new functional blocks that solve dispatching problems. In the absence of the dispatcher sequence of the functional blocks in the parallel block, parallel algorithm without dispatcher will be received.

When forming parallel block sequences of functional blocks of the same block (cloning) and the separation between the two sets of data structure parallel data mining algorithms will be obtained according to the parallelization. If the sequences of the parallel block will contain the different functional blocks processing the same data, the structure with parallelization by tasks will be received.

Structures of parallel data mining algorithms interact with and without, respectively, may be prepared by the addition or not of sender and receiver functional blocks.

As it was already marked earlier, the parallel block in general represents one functional block which accepts on an input a data set, execution settings and knowledge model, and returns the changed knowledge model. At the same time in the parallel block of the sequence of the functional blocks are executed parallely. In such sequences of run setting are copied directly, and operation with a data set and knowledge model can differ.

In this regard it is possible to select the following types of parallel data mining algorithms relatively parallely of the executed sequences of the functional blocks, built a knowledge model and the processed data set:

- single data source and single model (SDSM) – in this case, the parallel sequence of functional blocks work with one data source and one knowledge model;
- multiple data sources and single model (MDSM) – in this case, the parallel sequence of functional blocks work with multiple data sources and one knowledge model;
- single data source and multiple models (SDMM) – in this case, the parallel sequence of functional blocks work with one data source and multiple knowledge models;

- multiple data sources and multiple models (MDMM) – in this case, the parallel sequence of functional blocks work with multiple data sources and multiple knowledge models;

III. CONCLUSION

Thus, this paper proposes method of constructing parallel data mining algorithms based on representation of an algorithm in the form of hierarchy of nested functional blocks. To construct a parallel data mining algorithm the method assumes execution of the following steps:

- decompose data mining algorithm on elementary operations;
- group into functional blocks of the actions executing the discrete change in the model of knowledge;
- select structural elements and nested functional blocks in their sequence;
- add functional blocks structure for parallel execution.

REFERENCES

[1] I. Kholod, Z. Karshiyev, and A. Shorov, "The Formal Model of Data Mining Algorithms for Parelleize Algorithms" Advances in Intelligent Systems and Computing, Springer International Publishing Switzerland 2015, vol. 342, pp. 385–394, 2015.

[2] Kholod, I.I., Karshiyev, Z.A.: Parallelization of the algorithm Naïve Bayes on the basis of block structure. In: XV International Conference on Soft Computing and Measurements SCM'2012, vol. 1, pp. 182–185, Saint-Petersburg , 2012.