Electrical Power and Energy Systems 73 (2015) 548–559

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

A parallel heuristic reduction based approach for distribution network fault diagnosis

Yinglong Ma *, Xiao Yu, Yuguang Niu

State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, North China Electric Power University, Beijing 102206, China

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form 30 April 2015
Accepted 9 May 2015
Available online 5 June 2015

Keywords:
Distribute network
Fault diagnosis
Rough set theory
Parallel heuristic reduction
MapReduce

ABSTRACT

For the large volume of data in power systems, existing approaches to rough sets reduction either run on a single machine, or are paralleledly achieved in an approximate manner. They also seldom consider rough sets based value reduction. These problems restrict them in applications of power systems. In order to accelerate attribute reduction and value reduction, and improve the efficiency of fault diagnosis analysis in power systems, in this paper, we present a parallel heuristic approach to exact attribute reduction and value reduction for fault diagnosis in distribution network. We obtain the diagnosis rules and diagnose and locate the faults in the distribution network. Our parallel algorithms have been implemented on the MapReduce platform. The experimental results show that our method can effectively improve reduction process and improve the accuracy of reduction results in dealing with a large volume of data sets.

Introduction

The fault diagnosis of distribution network mainly focuses on the alarm information of protective devices in different levels, the change of the state information and the characteristics of electrical measurement such as voltage or current. Possible fault locations and fault types can be inferred according to the logic of protective actions and the personal experience of operators. Plenty of approaches can be applied to fault diagnosis [1,23], such as Rough set [4], Bayes Network [12], GA algorithm [13,24] and other particular approaches [39,40]. Despite that they have demonstrated their powerful potential in fault diagnosis, their efficacy for fault diagnosis remains to be further examined when the real-time information is incomplete, distorted or partially missing in the environment of massive data.

Rough set theory [2,3] is an effective tool to deal with imperfect information which may be imprecise, inconsistent or incomplete. Due to the stronger ability of fault tolerance and processing incomplete data, rough set based approaches have been widely used in fault diagnosis of power systems [5–8]. The key problem of rough sets based approaches is to consider attribution reduction, obtainment of decision rules, and intelligent computation for efficient data processing [9,10]. Unfortunately, attribution reduction greatly increases the computational complexity. On one hand, many approaches were proposed to obtain all attribute reducts of a given data set in the last years. A theoretic framework based on rough set theory called positive approximation [18], which is to accelerate a heuristic process for feature selection from incomplete data. In [11], two algorithms CEBARKNC and CEBARKCC were proposed for reducing the attributes of decision table based on conditional information entropy from the viewpoint of information theory. A discernibility matrix method was also proposed to obtain all attribute reducts of a given data set [19]. On the other hand, many heuristic attribution reduction approaches have been presented to efficiently obtain an attribute reduct because the above algorithms for attribution reduction are usually time-consuming to process large-scale data. A heuristic feature selection algorithm based on positive region reduction in incomplete decision tables was proposed [20], keeping the positive region of target decision unchanged. New information entropy was defined to measure the uncertainty of incomplete information for reducing redundant features by the corresponding conditional entropy [21]. A definition of combination entropy was proposed to measure the uncertainty of an incomplete information system for obtaining a feature subset [22]. However, most of the existing rough set tools only run on a single computer to deal with small data sets. Especially for the massive data in power systems, it greatly restricts the applications of rough sets. Due to the inefficiency of reduction algorithm, especially for large data sets, partition model

* Corresponding author. Tel.: +86 10 61772643.
E-mail address: yinglongma@gmail.com (Y. Ma).

http://dx.doi.org/10.1016/j.ijepes.2015.05.027
0142-0615/

of granular computing for attribute reduction algorithm is a challenging task.

Parallel reduction is a promising solution for accelerating the speed of data processing and analysis, and improving the efficiency of fault diagnosis. Many approaches for parallel attribute reduction have been proposed in the last decades [25,26,28,30–35,38]. Some approaches to parallel reduction were proposed by evaluating attribute significance [14], affinity propagation clustering [25], in/discernibility relation [31], discernibility matrix [33], and attribute frequency [26]. The origin decision table consisted of massive data is divided into several sub decision tables, and an optimal single candidate attribute is selected from all decision sub-tables once by comparing the number of records in the positive region in each decision sub-table. An efficient rough feature selection algorithm was presented with a multi-granulation view for attribution reduction [15]. It mainly calculates the reduction with a smaller granularity and then integrates the various reducts conducted from those smaller granules, resulting in a final reduct. Most of the above algorithms use significance functions are still computationally time-consuming to deal with large-scale data sets, and therefore can obtain approximate reducts instead of exact ones.

Fortunately, the MapReduce/Hadoop provides an ideal extension to distributed and parallel data processing [17]. The parallel methods based on the MapReduce [17] technique have been put forward to deal with the massive data [27,29,36,37]. Parallel attribute reduction by utilizing the MapReduce is a feasible solution towards massive data analysis and processing [30]. There are some most relevant researches to parallel rough sets based reduction [16,41]. Zhang et al. [16,41] proposed a parallel method based on MapReduce and rough set approximations for large-scale attribute reduction. On one hand, reduction processes in the work generate *approximate reducts* rather than exact reducts in nature, and therefore classifications based on the reduction method is not exact although their classification accuracy experiments demonstrate that their parallel methods are effective for selecting relevant features from very large data set. On the other hand, the authors failed to consider parallel value reduction, which is crucial to process numerical and missing data. In contrast, our parallel reduction method proposed in this paper will be *exact*, and consider the parallelization of *both attribute reduction and value reduction*. In additions, few researches for parallel reduction based on MapReduce can be found to address fault diagnosis based on massive data from power systems.

In this paper, we present a parallel approach to attribute reduction and value reduction for fault diagnosis in distribution network in order to accelerate attribute reduction and value reduction, and improve the efficiency of fault diagnosis analysis in power systems. Firstly, the initial decision table based on the historical fault information is constructed. Then, the parallel heuristic algorithms are proposed for attributes reduction and value reduction. And at last, the diagnosis rules can be obtained and the failure components can be located and diagnosed. Our parallel algorithms have all been implemented on Hadoop–MapReduce platform. By comparing with other algorithms in UCI data sets, the experimental results show that our method can effectively improve reduction efficiency and ensure the accuracy of reduction results in dealing with a large volume of data set.

This paper is organized as follows. Section 'Preliminaries' briefly gives a description on knowledge representation of Rough set theory and parallel reduction, as well as MapReduce. In Section 'Parallel reduction algorithm', as the core of this paper, parallel reduction algorithms which include attribute reduction and value reduction with MapReduce are represented in detail. System architecture and case analysis is presented in Section 'System architecture and case analysis'. Section 'Experimental evaluation' is the experiments and evaluation with other methods. Conclusion and future work is described in Section 'Conclusion'.

Preliminaries

MapReduce model

To solve the storage and computing problems of massive data, Google first provides a feasible solution for massive data mining – a large-scale distributed file system GFS (Google File System) which has the capacity of massive data storage and access, and a parallel programming model called MapReduce, which is a parallel programming technique derived from the functional programming concepts.

The MapReduce programming model is composed of a Map procedure and a Reduce procedure, in which the key-value pairs are taken as the basis data structure and the input and output data should be key-value pairs. The Map function mainly takes an input pair and procedures a set of intermediate key-value pairs for filtering and sorting. The MapReduce library groups together all intermediate values associated with the same intermediate key and transforms them to the Reduce function. The Reduce procedure accepts an intermediate key and a set of values for that key. It merges together these values to form a possibly smaller set of values. Typically, just zero or one output value is produced per Reduce invocation. Using MapReduce programming model, knowledge reduction for massive data can be realized.

Rough set representation of fault information

Rough set theory is an elegant and powerful methodology in extracting rules from decision tables or Pawlak information system. Rough set theory can not only keep the same capacity of classification by reducing the uncertain, imprecise data or those data lack of prior knowledge, but also get the corresponding sample sets that contain the characteristics of original data with minimum condition attributes. We use the Rough Set method for fault information representation.

Definition 1 (*Fault Decision Table*). A fault decision table can be described as $S = \langle U, A = C \cup D, V, f \rangle$, where U is a finite set of fault information records; $A = C \cup D$ is a non-empty and finite set of attributes, where C is the set of fault conditional attributes of system components, and D is a non-empty set of decision attributes, and $D \cap C = \emptyset$; V is the sets of attribute values; and $f = U \times A \to V$ is the fault information function which defines that an attribute value corresponds to a specific attribute of each fault record in U.

Definition 2 (*Indiscernibility relation*). Suppose that $S = \langle U, A = C \cup D, V, f \rangle$ is a fault decision table, and let $B \subseteq A$ be a subset of condition attributes. An associated equivalence relation with respect to B called the indiscernibility relation: $IND(B) = \{(x, y) | (x, y) \in U \times U, \forall b \in B(f(x) = f(y))\}$. It represents that two records x, y satisfying the relation $IND(B)$ are indiscernible by attributes from B. The equivalence relation $IND(B)$ partitions U into some equivalence classes denoted by $U/IND(B)$ or U/B.

Definition 3 (*Reduct and Core*). Given the set of records U and its attributes set A, let $B \subseteq A$, and B is called a reduct of the fault information table if and only if

(1) $IND(B) = IND(A)$;
(2) $\forall b \in B, IND(B) \neq IND(B \setminus \{b\})$