Information Sciences 000 (2016) 1–21

Contents lists available at ScienceDirect

Information Sciences

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

Complexity-based parallel rule induction for multiclass classification

Shahrokh Asadi [a], Jamal Shahrabi [b,*]

[a] Faculty of Engineering, Farabi Campus, University of Tehran, Tehran, Iran
[b] Department of Industrial Engineering, Amirkabir University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 26 September 2015
Revised 27 October 2016
Accepted 31 October 2016
Available online xxx

Keywords:
Rule induction
RIPPER
Multiclass classification
Genetic algorithm (GA)

ABSTRACT

To classify multiclass classification problems in RIPPER, classes are first sorted according to the increasing order of their prior probabilities. The rules for each class are then learned in that order. This learning process has two major shortcomings: (1) the order of class learning has a significant impact on the outcome of the classification, such that different permutations of classes perform differently; (2) the order in which the rules are learned is very important because the first rule to be fired determines the class of the instance. However, the correct class could be identified by another rule further down the list that is ignored and thus never examined. This paper offers two contributions that extend RIPPER to multiclass classification problems and address the mentioned shortcomings. The first issue is resolved by giving all of the classes an equal opportunity for rule extraction, and the class complexity is calculated using the description length. In the second execution of the algorithm, to learn each new rule, the complexity of the rules in each class is computed such that the class with the lowest remaining complexity is selected to determine the next rule. This algorithm is known as Complexity-based Parallel Rule Learning (CPRL), which can overcome the problem from the order of classes in rule learning. Furthermore, a Genetic Algorithm (GA) is developed to find the near optimal order of the rules with Evolutionary Reordering of the rules in the Decision list (ERD), which mitigates the second problem. Experimental results on 20 data sets demonstrate that the proposed algorithm outperforms RIPPER and can be considered to be a promising alternative for multiclass classification problems.

1. Introduction and literature review

Classification is among the most studied problems in Data Mining (DM) [24]. Learning rules from instances are the basis for rule-based classification, which facilitates a deep understanding of the problem in which the novel pattern is discovered. Taking a DM point of view, rule-based classification is preferred to many classification techniques [20]. Given a set of classified instances, the goal is to explore a logical description that correctly classifies unseen instances. Knowledge is represented as multiple *If – Then* rules in a rule set. Such rules state that the presence of one or more features implies or predicts the consequent. A typical rule has the following form:

Rule : **If** f_1 and f_2 and . . . f_n **Then** class

* Corresponding author.
E-mail addresses: s.asadi520@gmail.com (S. Asadi), jamalshahrabi@gmail.com (J. Shahrabi).

http://dx.doi.org/10.1016/j.ins.2016.10.047
0020-0255/

Please cite this article as: S. Asadi, J. Shahrabi, Complexity-based parallel rule induction for multiclass classification, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.10.047

where f_i, $i \varepsilon \{1, 2, \ldots, n\}$ is the feature that leads to the prediction of the consequent. The process of inducting rules based on instances is called inductive learning [3,19,35,37].

A large number of inductive algorithms leverage a separate-and-conquer approach, wherein a single rule is generated in each iteration to explain a portion of the training data. The covered instances are then eliminated (separate), and the next rule is learned using the remaining data (conquer) [20]. The purpose of the learning process is to cover all of (or a maximum number of) the positive instances while having no (or a minimum number of) negative instances. A new instance is positive provided that it leads to at least one rule being fired; otherwise, the instance is classified as negative. Separate-and-conquer algorithms for rule learning involve a variety of search methods that ultimately generate rule sets (i.e., unordered rule sets) or decision lists (i.e., ordered rule sets). Several such methods are listed in Table 1.

Once a rule in a decision list is learned, all of its associated positive and negative instances are removed. The order of the rules in the decision list is important. Using a decision list to classify a new instance, if the rule does not match the instance, i.e., if the values of the attributes in the instance do not satisfy the conditions of the rule, then the next rule is examined. In the case where none of the rules in the decision list satisfy the new instance, the default class is used. The class with the largest number of instances in the training data is assigned as the default class [46]. In unordered rule sets, after a rule is learned, only the positive instances are removed from the training data. To determine the class of a new instance, the entire set of generated rules is scanned to find a rule that can cover the instance. An inference process, such as the majority vote of the fired rules, is used to classify the instance, even though compared to unordered rule sets, decision lists provide a simpler mechanism for classifying new instances. However, they generate rules that are more difficult to interpret [19,28]. In this paper, we focus on decision lists for learning rules whose general structure is as follows:

If (conditions 1) **Then** class 1
Else If (conditions 2) **Then** class 2
. . . **Else** default class

Repeated Incremental Pruning to produce Error Reduction (RIPPER) [11] is undoubtedly one of the best algorithms that implements a separate-and-conquer strategy to separate positive and negative instances. This algorithm, which extends its predecessors including Incremental Reduce Error Pruning (IREP) [21], is the first to effectively tackle the overfitting problem and is still considered state-of-the-art in inductive rule learning [20,28,44]. RIPPER is a powerful classification rule learning algorithm that can address a very common problem of rule-based classification problems: the so-called curse of dimensionality. This algorithm can generate simple and compact sets of classification rules [28].

There have been many attempts to improve RIPPER. One especially interesting example is FURIA (Fuzzy Unordered Rule Induction Algorithm) [28], which learns fuzzy rules instead of crisp rules and unordered rule sets instead of decision lists. Yildis [46] used mathematical programming to extend RIPPER for multiclass classification problems. Asadi and Shahrabi [5] developed an Ant Colony Optimization (ACO) algorithm to determine the order of rules learned by RIPPER. In another paper [6], the same authors propose to sort the classes based on the prior probabilities, and in contrast to RIPPER, the rules are learned in a parallel manner. A new pruning measure called LogLaplace was developed in their work.

Multiclass problems are abundant in real-life applications in medicine, genomics, bioinformatics, or computer vision [16,48]. Quite often, the increase in the number of classes comes with greater complexity in the estimated decision rules. This relationship potentially leads to overfitting and increasing the computational cost of the recognition system. In addition, the difficulties in classification could only be present for some classes while others are separated with minimal error.

In RIPPER, a multiclass classification problem is broken down into a sequence of several two-class problems. In the basic sequential rule induction algorithm, a specific class order is typically determined according to different heuristics, without being specifically defined. The learner usually receives the classes in increasing order of prior probabilities. Previous findings indicate that the order in which the classes are learned by the learner has a significant impact on the performance of the final rule set [8,10,46], which makes the permutations of learning the classes important. The quality and interestingness of the subsequent rules tend to decrease as the rules are learned [31,32]. This property is a major shortcoming of RIPPER in learning decision lists, and it is especially true for multiclass problems because each rule is dependent on previously learned and removed rules. This type of learning (i.e., sequential learning) has several disadvantages. Most importantly, the algorithm is very likely to be trapped in a local optimum, and the learning process could be biased toward a certain class or classes because the rules are not necessarily learned in the correct order. In addition, it could be no longer possible to find interesting rules from all of the classes. The first proposed algorithm in this paper improves RIPPER for multiclass classification problems by avoiding local optima, preventing biases toward a certain class or classes, and allowing interesting rules to be learned from all of the classes [6].

In the proposed algorithm, the classes are first sorted in increasing order of their prior probabilities, with rules being learned simultaneously for all of the classes according to their prior probabilities; the class complexities are computed in terms of their Description Length (DL). Next, the classes are once again sorted according to complexity to designate the least complex one as the default class. Then, the rules for all of the classes are learned simultaneously in increasing order of the class's remaining complexity. Each iteration of the algorithm involves a comparison of all of the classes, to select the class that has the minimum complexity, and then, a rule from that class is learned.

Evolutionary Algorithms (EAs) [15,45] have been successfully applied to rule induction [38]. EAs, especially Genetic Algorithms (GAs), are considered to be one of the most successful search techniques for complex problems and have proved to be an important technique for learning and knowledge extraction. Aguilar-Ruiz et al. [1] proposed a GA that is capable of

Please cite this article as: S. Asadi, J. Shahrabi, Complexity-based parallel rule induction for multiclass classification, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.10.047