Information Sciences 332 (2016) 56–71

Contents lists available at ScienceDirect

Information Sciences

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

Social choice in distributed classification tasks: Dealing with vertically partitioned data

Mariana Recamonde-Mendoza*, Ana L. C. Bazzan

PPGC, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 21 January 2014
Revised 8 September 2015
Accepted 2 November 2015
Available online 6 November 2015

Keywords:
Distributed data mining
Vertical partition
Heterogeneous sources
Social choice theory

ABSTRACT

In many situations, a centralized, conventional classification task can not be performed because the data is not available in a central facility. In such cases, we are dealing with distributed data mining problems, in which local models must be individually built and later combined into a consensus, global model. In this paper, we are particularly interested in distributed classification tasks with vertically partitioned data, i.e., when features are distributed among several sources. This restriction implies a challenging scenario given that the development of an accurate model usually requires access to all the features that are relevant for classification. To deal with such a situation, we propose an agent-based classification system, in which the preference orderings of each agent regarding the probability of an instance to belong to the target class are aggregated by means of social choice functions. We employ this method to classify microRNA target genes, an important bioinformatics problem, showing that the predictions derived from the social choice tend to outperform local models in this application. This performance gain is accompanied by other interesting advantages: the aggregation methods herein proposed are extremely simple, do not require transfer of large volumes of data, do not assume an offline training process or parameters setup, and preserve data privacy.

1. Introduction

A quite common assumption in classification tasks is that the data set for model development is centralized and fully accessible by the classifier [1,60]. In this case, there are several well-established classification algorithms that are able to provide accurate models. A survey can be found in [24]. Nonetheless, in many situations data centralization may be impracticable or undesirable due to context-specific constraints, e.g., storage and computing costs, communication overhead and privacy or intellectual property concerns, resulting in a distributed data mining (DDM) problem [45]. Examples of scenarios where these restrictions may arise are biomedical research, fraud detection in financial organizations, and calendar management by software assistants, in which ethical, legal or privacy issues prevent data sharing, thereby inducing a physical distribution of data. Under these constraints, data mining requires distributed data analysis, with minimal data communication among sources [27].

Generally, DDM is performed by generating local models based on distributed data analyses and then adopting a strategy to combine them into a composite, global model [60]. In a typical DDM setting, data sets are partitioned primarily in one of two ways.

On the one hand, the sources may be homogeneous, in the sense that each site contains exactly the same features set, i.e., the same description, for different instances across similar domains. This is referred to as *horizontally partitioned data* (Fig. 1a). As

* Corresponding author. Tel.: +55 5185790829.
E-mail addresses: mari.mendoza@gmail.com, mrmendoza@inf.ufrgs.br (M. Recamonde-Mendoza), bazzan@inf.ufrgs.br (A.L.C. Bazzan).

http://dx.doi.org/10.1016/j.ins.2015.11.001
0020-0255/

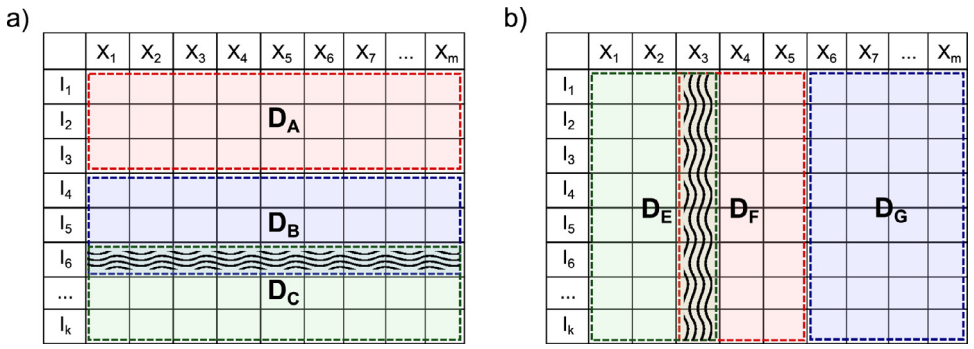

Fig. 1. In DDM, data sets may be partitioned either (a) horizontally, in which homogeneous sources D_A, D_B and D_C carry the same type of information (features set) about different instances, or (b) vertically, in which the features set is distributed among heterogeneous sources D_E, D_F and D_G. Note that overlaps may occur in both situations (hachured area).

an example, consider the case of several weather stations registering the same features sets (e.g., temperature, humidity, wind speed) to characterize weather in a geographically distributed environment (different instances), whose data will be further used for meteorological predictions. Despite the physical distribution of information and the impossibility to communicate the raw data given its prohibitively large volume [34], classical learning algorithms are more easily adaptable to this scenario given that all local models are learned from data of comparable quality and deal with the same form of input (the set of features), which facilitates their comparison and, eventually, their combination into a global model [62].

On the other hand, data sources may be heterogeneous, which means that they carry different kinds of information, i.e., different features sets, related to the same set of instances. In this scenario, data distribution occurs by means of *vertically partitioned data* (Fig. 1b). For instance, biomedical applications often need to consult records distributed among several heterogeneous domains, such as clinical data, genotype data and medical imaging, to define a more accurate diagnostic for a single patient. Under this condition, the distributed nature of data is a more critical issue, because conventional classification algorithms are likely to fail in building a precise model given that the accurate prediction of the class of unlabeled instances usually requires access to all features that are relevant for their classification [40]. Moreover, the combination of locally developed models into a global model is not very straightforward because their performance can present a substantial variation for different parts of the input space and not every combination strategy can effectively deal with this situation [56].

As already noted in literature, problems related to DDM with homogeneous data sources have been widely studied and, in general, they are more easily treatable with existing classification algorithms [28,56,62]. Conversely, the accurate classification under distributed, vertically partitioned data remains an open and challenging problem in the field [40,60]. Hence, in this paper we are concerned with improving classification results in problems whose scenario implies a vertical partitioning of data among several heterogeneous sources. We assume that data centralization is neither possible nor desirable due to domain-specific constraints as aforementioned, and consequently a non-standard strategy must be applied to overcome this drawback.

A natural choice to deal with applications that require distributed problem solving are multiagent systems (MAS), which not only are inherently distributed systems, but also cope well with heterogeneous data. Indeed, there has been an increasing interest around the integration of agent technologies and data mining, as discussed in the recent publication by Cao and colleagues, and others [9,29,51]. The approach proposed in the current work meets this trend by developing an agent-based data mining system to address vertically partitioned data in DDM. More precisely, agents in our system encapsulate distinct machine learning (ML) algorithms and are responsible for two main tasks, (i) individually build local models based on their private information about the classification task and (ii) collaboratively work with other agents towards the construction of a global, consensus model. Despite eventual similarities between this methodology and the fundamental functioning of ensemble learning algorithms, we remark that multiagent systems and ensemble learning are two different, but equally suitable techniques to approach DDM problems, as reviewed in [14]. We briefly discuss this issue in Section 2.

Here, we focus on a specific research question related to agent-based DDM [7], namely how to integrate the knowledge unveiled by a group of agents into a globally coherent model. Typically, communication is a bottleneck in distributed applications because the cost of transferring large blocks of data may be too expensive and jeopardize the efficiency of the system [51]. Yet, neither simple combiners such as averaging, nor meta-learners, which usually require few data communication among sources, are considered suitable to deal with heterogeneous environments [56]. Simple combiners are vulnerable to the large discrepancies that may be observed among agents' performance over different parts of the input space. Meta-learners that assign different weights to agents according to the quality of their predictions often require off-line training, which may be unfeasible for a large and distributed data set, especially if the features are distributed.

In this paper, we investigate the viability of mechanisms inspired by the social choice theory as a strategy to derive a global model in DDM problems, while efficiently handling the fundamental trade-off between communication and accuracy. Our approach does not require frequent communication or transferring of large blocks of data among agents and is also independent of parameters setting and tuning. Still, empirical evaluation with a bioinformatics data set suggests that this apparently simple