Distributed Data Mining for Astrophysical Datasets

Sabine M. McConnell, David B. Skillicorn

School of Computing, Queen’s University, Kingston, Ontario, Canada

Abstract. Over the past decade, data mining has gained an important role in astronomical data analysis. Traditionally, such analysis is performed on data at a single location. However, one of the main motivational forces behind a virtual observatory is the distributed nature of both data and computational resources. Existing data-mining methods for distributed data are either communication-intensive or result in a loss of accuracy. In this paper, we introduce a general approach to supervised data mining that allows data to remain distributed, but still produces satisfactory results. We demonstrate by applying the approach to a number of astronomical datasets.

1. Introduction

Data mining, as the automated or semi-automated discovery of novel and potentially useful information, has become an important tool for astrophysical datasets over the past decade. This becomes evident very quickly when one considers the widespread range of astronomical applications, utilizing a variety of techniques. Examples for this include star/galaxy separation (Odewahn and Nielsen 1994), classification of stars (Weaver 2000), galaxies (Odewahn 1997) and planetary nebulae (Faúndez-Abans 1996), antimatter search in cosmic rays (Bellotti 1997), detection of expanding HI shells (Daigle et al. 2003) and selection of Quasar candidates (Padovani et al. 2004).

In a traditional setting, data-mining algorithms are applied to data at a single location. When data is collected in a distributed way, this means that, at some point, the data must be transferred to a single computer. This enables conventional data-mining algorithms to be applied, but is costly in terms of communication, and storage facilities at the central site. Processing power and memory size limits at such a central site will limit the sophistication of algorithms, and the size of the datasets that can reasonably be processed. It also requires each data source to reveal its raw data completely, which may not always be appropriate.

One of the major design goals of the virtual observatory is to achieve universal access to geographically distributed datasets, so new data-mining approaches that can handle distributed data are required. At present, few approaches that can achieve these goals are known.

In this paper, we introduce a method for distributed data mining that both reduces the time needed to produce results while maintaining or even increasing the accuracy compared to the same algorithm applied to centralized data at a
single site. We demonstrate the effectiveness of this method using a variety of astronomical datasets.

2. Partitions of astronomical datasets

An astronomical dataset typically contains information about a set of objects, such as galaxies or stars, and a set of attributes or measurements about these objects. Because of the way data is collected, these datasets are partitioned by objects (horizontal partitioning) and by attributes (vertical partitioning).

Horizontal partitioning stems from a variety of factors such as the physical location of the collection instruments (Canadian Galactic Plane Survey vs. Southern Galactic Plane Survey, for example), differing target objects (Condon 1987 and Condon 1990), as well as detection limits (fainter objects cannot be detected by less sensitive instruments). Vertical partitioning stems from the fact that different surveys measure different parameters, for example spectra at different wavelengths. A more subtle form of vertical partitioning occurs because of the variation in resolution caused by technological advances as instruments become more sensitive.

Astrophysical data mining is therefore more complex than in many other domains where data may be partitioned vertically or horizontally, but typically not both. Mining astronomical datasets obtained from ground- or space-based surveys is therefore at one extreme of a spectrum of complexity.

3. Related Work

For horizontally distributed data, where information about different objects appears in different partitions, a number of approaches are known, typically producing similar accuracy to the equivalent centralized algorithm. These may require only minimal communication, but they often require some assumption that the data in different partitions is similar, so that local models are good approximations to the final, global model (Skillcorn 1999). This assumption may not be valid for astrophysical data.

Another successful approach for data partitioned by objects is based on ensembles, that is collections of classifiers. In ensemble techniques, predictors are built from local datasets and then combined, most often through a simple voting scheme, to produce a final classification for previously unseen data.

For vertically partitioned data, two approaches, Meta-Learning (Chan 1993) and Collective Data Mining (CDM) (Kargupta 1999) have been developed. Both techniques are expensive, and the accuracy achieved through CDM, which represents the function to be learned with an approximation as the sum of a set of basis functions together with their corresponding coefficients, is always lower than that achieved on a centralized dataset.

4. A Lightweight Distributed Data Mining Approach

Our approach to building a predictor for data that is both horizontally and vertically partitioned is to build a predictor locally on each partition, send these
predictors to a central site and combine their predictions by voting. Hence this is an ensemble technique extended to a more-general form of data partitioning. This is a simple idea, but still outperforms much more sophisticated and expensive alternatives, as we show below.

For the evaluation presented here, we use a decision tree, which classifies the dataset by a hierarchical series of single-attribute relations (Quinlan 1986). This is a supervised data-mining technique in which the goal is to learn how to predict the target class of new data, given training data labelled with the target class. An example of such a classification task is galaxy classification: images of unclassified galaxies are categorized based upon a model derived from known galaxy classifications (Odewahn 1997). For our implementation we used J48, the decision tree implemented in the data mining package WEKA, which is available for download at http://www.cs.waikato.ac.nz/ml/weka/. It is important to note that our approach is not restricted to decision trees, but can use any weak learner.

An outline of the algorithm is given below.

1: create training set
2: create out-of-bag test set
3: create vertical partitions of the training set
4: for all vertical partitions do
5: create horizontal partitions for the training set
6: end for
7: for all partitioned datasets do
8: build a model using local data only
9: classify the out-of-bag test set
10: communicate the results to a centralized site
11: end for
12: Combine the local results using voting to obtain overall results

Training sets are created by resampling the data with replacement until a set of the original size of the dataset is reached. Typically about a third of the data will never have been sampled, and plays the role of a test set, called an out-of-bag test set. Both simple voting, and weighted voting using test accuracy are used.

This approach was evaluated on 5 astronomical datasets, two of which were taken from the SDSS Data Release 3. Table 1 shows the number of objects, the number of classes (distinct groups of objects) as well as the number of attributes (information about the objects) for each of the datasets.

<table>
<thead>
<tr>
<th>Name</th>
<th>Objects</th>
<th>Classes</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dwarfs</td>
<td>8119</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Galaxy/QSO</td>
<td>11427</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Galaxies</td>
<td>1151</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Galaxy/Star</td>
<td>4192</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>Variable Stars</td>
<td>10970</td>
<td>3</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1. Overview of dataset properties
5. Results and Discussion

Table 2 shows the prediction accuracies for a decision tree given the entire dataset, and the best and worst prediction accuracies for distributed predictors. Predictors were built for all combinations of partitions where the number of horizontal partitions ranges from 1 to 8; and the number of vertical partitions ranges from 1 to the number of attributes. Hence, in the hardest case, predictors were learned from single attributes, and only 1/8th of the objects.

<table>
<thead>
<tr>
<th>Name</th>
<th>centralized case</th>
<th>best distributed case</th>
<th>worst distributed case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dwarfs</td>
<td>99.69</td>
<td>98.65</td>
<td>96.61</td>
</tr>
<tr>
<td>Galaxy/QSO</td>
<td>80.98</td>
<td>80.96</td>
<td>75.11</td>
</tr>
<tr>
<td>Galaxies</td>
<td>77.14</td>
<td>78.06</td>
<td>71.60</td>
</tr>
<tr>
<td>Galaxy/Star</td>
<td>93.55</td>
<td>94.96</td>
<td>87.03</td>
</tr>
<tr>
<td>Variable Stars</td>
<td>99.97</td>
<td>99.95</td>
<td>90.03</td>
</tr>
</tbody>
</table>

Table 2. Prediction accuracies

From these results, it can be seen that, for each of the datasets under investigation, the achieved accuracy in the best distributed case is at least as good as that obtained from a centralized approach. For the two datasets which exhibited an increase in accuracy, the result is statistically significant. Even for the worst case, the Variable Stars dataset, the accuracy drop is only 10%, and occurs in the worst configuration when each vertical partition contains only
a single attribute. When the partitions contain 2 attributes, the prediction accuracy immediately increases to over 95%.

In general, other advantages of our approach are that it requires less time to build the prediction models, since each is learning from a smaller dataset; and the models are typically smaller than that learned from centralized data, even in aggregate.

6. Summary

Our lightweight distributed data-mining approach, in which local models are built from data partitioned both by attributes and objects and then combined through a simple voting scheme, can achieve performance comparable to, and occasionally better than, that of a centralized predictor for all of the datasets investigated here.

The advantages of our approach are: it is simple, it requires very little communication, it can use any weak learner as the underlying predictor, and it is typically faster than the equivalent centralized version. In addition, no transmission of raw data is necessary, so our approach is also suitable for situations where disclosure of data is not desired, such as for example an initial proprietary period after survey completion.

Given the success of data mining in astrophysical datasets, data mining will be an important component of the Virtual Observatory. Modelling from distributed data without collecting it in a single location requires data-mining approaches of the kind described in this paper. The existence of a simple and effective way to build predictors from such distributed data demonstrates ad-
ditional advantages of the Virtual Observatory over existing centralized data repositories.

References


