<INSERT TITLE HERE>

Data Mining and Machine Learning / Intelligent Systems, Interaction and Multimedia Seminar

José Pedro Marques, Tiago Pereira, André Maia

Faculdade de Engenharia da Universidade do Porto

Abstract. Data Mining is a very broad area, with several algorithms applicable to the same problem. The purpose of this paper is to show a framework that given a specific problem as input, solves it using the best possible algorithm. To reach this goal, a distributed multi-agent system that tries to negotiate the best approach to each problem will be implemented, and this system will be tested with the data from the Portuguese Institute of Statistics, on an error detection problem.

1 Introduction

The purpose of this article is to present the framework that will be developed, as well as all the protocols that will be implemented. To best demonstrate the inner workings of the framework, a specific problem was chosen.

This article will include a description of the problem and the data available (Section X). This is followed by a description of the chosen algorithms. Then the framework's architecture is explained (Section X), and how each agent was implemented (Section X), followed by the negotiation process (Section X). Finally, the results are shown and a small conclusion is made (Section X)and possible future improvements are discussed (Section X)

2 The Problem

2.1 Description

The information from each transaction a Portuguese company makes with an EU country reaches the Portuguese Institute of Statistics(INE) through the INTRASTAT form. In this form, the company provides information about the transaction type (import/export), the item id, the weight of goods traded, the total cost, etc. Then this data is manually inserted into a database at INE. Figure 1 presents an excerpt of the data (é preciso perceber os dados).

As in all manual processes, both the form filling and the insertion in the database are error prone, which could lead to incorrect/inconsistent data. For instance, an incorrectly introduced item id will associate a transaction with the wrong item. While some errors may be irrelevant to some statistics, some errors can influence them greatly.

Therefore, when all of the transactions relative to a month have been entered into the database, they are manually verified with the aim of detecting and correcting as many errors as possible. In this search, the experts try to detect unusual values in the values that describe the transactions. Given that the number of transactions declared monthly is in the order of tens of thousands, this is a very costly process.

The idea then is to automatically select a subset of the transactions that includes almost all the errors that the experts would detect by looking at all the transactions.

2.2 Data

2.3 Chosen Algorithms

2.3.1 Clustering for Outlier Detection This technique is not usually used for outlier detection, but it can be used for that task. Every value assigned to a small cluster containing significantly fewer points than the others, is considered an outlier.

2.3.2 Decision Trees for Outlier Detection Algorithms based on decision trees, learn from a set of pre-classified examples, and build a model of the regularities found used to classify new cases. This kind of technique is better suited for the detection of outliers, where an isolated branch is classified has an outlier. This is an algorithm that as the advantage that it doesn't require the need to define the number of clusters, and it is not scalable, this is, he's complexaty over time is he's numbers of objects (O(n2)). Although the interpretation of results can be subjective.

References