Machine Learning and Statistics to Detect Errors in Forms: Competition or Cooperation?

Carlos Soares[1], Pavel Brazdil[1], and Cláudia Pinto[2]

[1] LIACC/Faculty of Economics, University of Porto, R. Campo Alegre 823, 4150-180 Porto, Portugal, {csoares,pbrazdil}@liacc.up.pt
[2] INE/DRN, Edíficio Scala, R. de Vilar 235, 9º andar, 4050, Porto, Portugal claudia.pinto@ine.pt

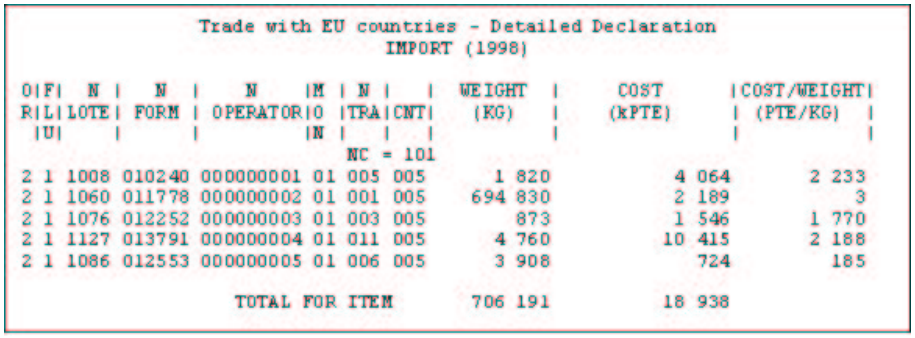
Abstract. We address the problem of detecting errors in foreign trade forms, which are submitted by companies to the Portuguese Institute Statistics (INE). In previous work, we have compared statistical techniques for outlier detection with an inductive learning method, with the latter obtaining the best results. Here, we present more recent results on that problem. We also hypothesize that the combination of outlier detection methods with inductive learning algorithms might be a better approach to dealing with this problem and we propose ways of doing so.

1 Introduction

This paper is concerned with the problem of detecting errors in foreign trade data collected by the Portuguese Institute of Statistics (INE). The objective is to identify the transactions that most likely contain an error. These will then be manually analyzed by specialized staff and corrected if an error really exists. Previous work on this problem has compared statistical methods for outlier detection with C5.0, a decision tree induction algorithm [1]. The latter technique obtained the best results, achieving the minimum goals that were established by the domain experts. This approach, however, only addressed part of the problem. As will be explained in Section 2, errors may be detected by looking into several fields while, in the work mentioned and which will be summarized in Section 3, only one field was taken into account. Here we extend the work to deal with all the relevant fields. We address this problem using a combination of models [2], which is currently a very popular approach in Machine Learning and Data Mining (Section 4). The model combination approach is then proposed as a way of enabling "cooperation" rather than "competition", as was done in [1], between learning and outlier detection methods (Section 5).

2 The Problem

The transactions made by portuguese companies with organizations from other EU countries are communicated to the Portuguese Institute of Statistics using the INTRASTAT form. In this form the company provides information about the transaction, namely:



Trade with EU countries - Detailed Declaration
IMPORT (1998)

O|F| N	N	N	N	M	N		WEIGHT	COST	COST/WEIGHT
R|L|LOTE	FORM	OPERATOR	O	TRA	CNT		(KG)	(kPTE)	(PTE/KG)
|U|			N						
			NC = 101						
2 1 1008	010240	000000001	01	005	005		1 820	4 064	2 233
2 1 1060	011778	000000002	01	001	005		694 830	2 189	3
2 1 1076	012252	000000003	01	003	005		873	1 546	1 770
2 1 1127	013791	000000004	01	011	005		4 760	10 415	2 188
2 1 1086	012553	000000005	01	006	005		3 908	724	185
		TOTAL FOR ITEM					706 191	18 938	

Fig. 1. An excerpt of the INTRASTAT database. The data presented was modified to preserve confidentiality.

- item id
- weight of the traded goods
- total cost
- type (import/export)
- source, indicating whether the form was submitted using the digital or paper versions of the form
- form id
- company id
- stock number
- month
- destination or source country, depending on whether the type is export or import, respectively.

At INE, the data are inserted into a database. Figure 1 presents an excerpt of a report produced with data concerning import transactions for a month in 1998 of item with id 101, as indicated by the field labelled "NC", below the row with the column names.

Given that both form filling and its transcription to the database are manual processes, errors often occur. For instance, an incorrectly introduced item id will associate a transaction with the wrong item. Another common mistake is caused by the use of incorrect units like, for instance, declaring the cost as PTE instead of kPTE. Some of these errors have no effect on the final statistics while others can affect them significantly.

The number of transactions declared monthly is in the order of tens of thousands. 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 on a few attributes. One of these attributes is Cost/Weight, which represents the cost per kilo and is calculated using the values in the Weight and Cost columns. In Figure 1 we can see that the values for Cost/Weight in the

second and last transactions are much lower than in the others. If we analyze the corresponding forms we can conclude that the second is, in fact, wrong, due to the weight being given in grams rather than kilos, while the last one is correct.

Our goal was to reduce the time spent on this task by automatically selecting a subset of the transactions that includes almost all the errors that the experts would detect by looking at all the transactions. To be acceptable by the experts, the system should select less than 50% of the transactions containing at least 90% of the errors. Additionally, the experts would like to be able to understand the decisions made by the system, and thus, it should provide an interpretable model. Note that efficiency is not important because the automatic system will hardly take longer than half the time the human expert does.

The automation of this task is not easy. Firstly, although it is not shown in Figure 1 for the sake of simplicity, quantity and cost may be declared in two fields each. In same goods the quantity may be measured in a way other than weight, which is called *supplementary units*. For instance, in a transaction of cars, the supplementary units are the number of units. A transaction may also have two values, the one which is declared, the *invoice value*, and a *statistical value*, calculated using an appropriate formulae. Therefore, we may have up to four combinations of Cost/Quantity to analyze. This adds more complexity to the problem since, on one hand, different combinations are the best error predictors for different groups of transactions. On the other hand, none of the combinations is present in all the transactions. Therefore, a complete solution to the problem requires all the combinations to be treated.

Furthermore, the error in the example shown above was easy to detect, but many of the errors spotted by the experts are not so obvious and their experience is essential to detect them. This information cannot be used directly in this work because we have no indication of which errors are "obvious" and which ones are not. Other difficulties are the small proportion of errors relative to the number of transactions (<0.5%), the patterns of normal transactions, which differ from item to item, and items with very few transactions.

3 Previous Results

In [1], a first approach to this problem only took the Cost/Weight attribute into account. The data concerned transactions from five months in 1998. It was provided in the form of two files per month, one with the transactions before being analyzed and corrected by the experts, and the other after that process. A considerable amount of time was spent preparing the data, for instance, to eliminate transactions that existed in one of the files but not in the other.

Four very different methods were applied. Two come from statistics and are univariate techniques: box plot [3] and Fisher's clustering algorithm [4]. The third one, Knorr & Ng's cell-based algorithm [5], is an outlier detection algorithm which, despite being a multivariate method, was used only on the Cost/Weight attribute. The last is C5.0 [6], a multivariate technique from the field of machine learning, based on induction of decision trees.

Decision tree induction is not an outlier detection technique and so we need to help the system to learn the notion of outlier. We created a new attribute that captures the notion of distance between points, i.e. different values of the Cost/Weight variable. Thus, NCost/WeightDistance represents the distance between a point (a value of Cost/Weight) and the average, in number of standard deviations:

$$\text{NCost/WeightDistance} = \frac{Cost/Weight - \overline{Cost/Weight}}{\sigma_{Cost/Weight}}$$

Note that for each transaction we use the average and standard deviation of the corresponding item. Finally, we added the attributes average and standard-deviation of cost weight and number of transactions, all calculated for each item id. From the set of original attributes, we selected only Cost, Weight and Cost/Weight. All the others are not expected to be predictive in their original form and, thus, were discarded. Following advice from the domain experts, import and export transactions were handled separately because their errors are quite different.

Given that C5.0 provides a confidence score associated with each prediction, it enables a compromise between the manual effort required and the number of errors detected to be made. This can be achieved by analyzing the positive predictions (i.e. the transactions that are identified as potential errors) in descending order of confidence and stopping when enough transactions have been analyzed. The results obtained based on this approach indicate that it is possible to detect 90% of the errors by analyzing just 40% of the transactions. Therefore, we were able to detect the required number of errors with less 10% of transactions than the domain experts would be willing to analyze.

Furthermore, the top nodes of the decision tree induced by C5.0 contain knowledge which makes sense to the domain experts. The first node, as would be expected, does outlier detection using the NCost/WeightDistance attribute. The second node, identifies small values of Cost/Weight. The experts confirmed that they pay more attention to smaller values of this attribute, because when an error generates a small value of Cost/Weight it usually affects the statistics more significantly than if it generates a large value. Finally, the third node recommends full manual verification of items with few transactions.

4 Dealing With All Cost/Quantity Fields

As mentioned in Section 2, a complete solution to the problem requires that all four combinations of the Cost/Quantity attributes be handled and not just one, like was done in the work described in the previous section. There are a number of ways to do that. We have opted to follow a model combination approach, which is currently very popular in Machine Learning and Data Mining [2]. We create a different data set for each problem (i.e. each combination), containing the attributes corresponding to the particular Cost/Quantity combination, the target attribute and other general attributes, like item id, etc, and then we

induce a model from each of those data sets. The next issue is how to combine the prediction of each of these models. Again there are several alternatives. We have opted for a voting scheme, which is also a common approach [7]. However, rather than using a uniform vote, where the most frequently voted class wins for each transaction, we have used weighted voting. The weight of each prediction is the corresponding confidence which enables us to explore the strength of each prediction [7].

Given that we are now dealing with all four combinations of the Cost/Quantity attributes, the number of transactions is larger than it was in the data used in the previous section because some of the transactions did not have value for the corresponding Cost/Quantity attributes combination. Therefore, our goal is to investigate whether the results in this new setting are better or worse than the ones obtained before. The evaluation method used is the hold-month-out approach, an adaptation of the well-known hold-out approach. The model is induced using data from all but one month and evaluated on the latter. The results are slightly worse than the ones obtained above although well within the limits established by the domain experts: 47% of the transactions were selected, containing 91% of the errors. This performance reduction was expected because the combinations of Cost/Quantity attributes which are now handled have much less data and very few errors. Therefore, the corresponding models are expected to have less generalization power, thus, hurting the general performance of the system.

5 Combining Inductive Learning and Outlier Detection

As summarized in Section 3, a few methods for the task of detecting potential errors in foreign trade data have been compared before [1]. These methods included not only an inductive learning algorithm but also outlier detection methods. As an alternative to the competition approach followed in that work, we believe that a collaboration approach, combining these methods, could obtain better results. As mentioned above, the combination approach is quite popular in Machine Learning [2] and, to the best of our knowledge, has never been applied to the problem of outlier detection.

As in the previous section, combination of models could be made by voting. An alternative approach is inspired in stacked generalization [8], where the predictions of base models are fed to a second-level inductive algorithm. This algorithm generates a model that predicts the same target as the base models but using their outputs as inputs, i.e. attributes. Cascade generalization [7] is an extension of stacked generalization, which feeds the original attributes used by the base models to the second-level algorithm, together with the predictions of the base models.

We have investigated a few outlier detection methods to use at the base level. Two methods that have been traditionally used in statistics for outlier detection are discordancy tests and principal component analysis [9]. More recent approaches that have been developed in the context of data mining are distance-

based [5], density-based [10] and unsupervised [11, 12] methods. Some of the more recent methods have been developed with large amounts of data in mind, as in the present case. We will select some of these methods to be used in our work. With this approach, we expect to improve our previous results. That is, we expect to select less than 40% of the transactions containing more than 90% of the errors.

6 Conclusions

We have presented recent results on the problem of detecting errors in foreign trade data that is collected by the Portuguese Institute Statistics (INE). We have used the structure of the problem, which was ignored in previous work, to divide it into four different sub-problems. The models generated for each of those problems were combined using weighted voting. We have also proposed the combination of outlier detection methods with inductive learning techniques, and investigated some outlier methods that could be used for that purpose. We will implement and empirically evaluate this approach.

Acknowledgments The authors wish to thank the INTRASTAT team at INE, without whom this work would not have been possible: Armindo Carvalho, Vítor Cortez, Óscar Alves, Fernanda Sengo, Ilda Alves and Natércia Ferreira.

References

1. Soares, C., Brazdil, P., Costa, J., Cortez, V., Carvalho, A.: Error detection in foreign trade data using statistical and machine learning methods. In: Proceedings of the 3rd International Conference and Exhibition on the Practical Applications of Knowledge Discovery and Data Mining (PADD99). (1999)
2. Diettrich, T.: Ensemble learning. In Arbib, M., ed.: The Handbook of Brain Theory and Neural Networks, MIT Press (2002)
3. Milton, J., McTeer, P., Corbet, J.: Introduction to Statistics. McGraw-Hill (1997)
4. Fisher, W.: On grouping for maximum homogeneity. Journal of the American Statistical Association (**53**) 789–798
5. Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24th VLDB Conference. (1998)
6. Quinlan, R.: C5.0: An Informal Tutorial. RuleQuest. (1998) http://www.rulequest.com/see5-unix.html.
7. Gama, J., Brazdil, P.: Cascade generalization. Machine Learning **41** (2000) 315–343
8. Wolpert, D.: Stacked generalization. Neural Networks **5** (1992) 241–259
9. Barnett, V., Lewis, T.: Outliers in Statistical Data. John Wiley & Sons (1978)
10. Breunig, M., Kriegel, H.P., Ng, R., Sander, J.: LOF: Identifying density-based local outliers. In: Proceedings of the MOD 2000, ACM (2000)
11. Yamanishi, K., Takeuchi, J., Williams, G.: On-line usupervised outlier detection using finite mixtures with discounting learning algorithms. In: Proceedings of the KDD 2000, ACM (2000)

12. Lauer, M.: A mixture approach to novelty detection using training data with outliers. In Flach, P., de Raedt, L., eds.: Proceedings of the 12th European Conference on Machine Learning, Springer (2001) 300–311