# Outlier Detection Using Clustering Methods: a data cleaning application

Antonio Loureiro[1], Luis Torgo[2], and Carlos Soares[2]

[1] LIACC
[2] LIACC-FEP
University of Porto, R. Campo Alegre, 823, 4150 Porto, Portugal
{ltorgo,csoares}@liacc.up.pt,
Home page: http://www.liacc.up.pt/~{ltorgo,csoares}

**Abstract.** This paper describes a methodology for the application of hierarchical clustering methods to the task of outlier detection. The methodology is tested on the problem of cleaning Official Statistics data. The goal is to detect erroneous foreign trade transactions in data collected by the Portuguese Institute of Statistics (INE). These transactions are a minority, but still they have an important impact on the statistics produced by the institute. The task of detecting these rare errors is a manual, time-consuming task. Our methodology is able to save a large amount of time by selecting a small subset of suspicious transactions for manual inspection, which, nevertheless, includes most of the erroneous transactions. In this study we compare several alternative hierarchical clustering methodologies for this task. The results we have obtained confirm the validity of the use of hierarchical clustering techniques for this task. Moreover, our results when compared to previous approaches to the same data, clearly outperform them, identifying the same level of erroneous transactions with significantly less manual inspection.

## 1 Introduction

This paper addresses 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 are most likely to contain errors. The selected transactions will then be manually analyzed by specialized staff and corrected if an error really exists. The goal is to detect most of errors using the less possible manual work. Previous work on this problem has compared outlier detection methods, a decision tree induction algorithm and a clustering method [11]. The results obtained with the latter did not achieve the minimum goals that were established by the domain experts, and, thus, the approach was dropped. Here, we investigate more thoroughly the use of clustering methods to address this problem.

We describe an approach that uses hierarchical clustering methods to find clusters with few transactions that are expected to contain observations that are significantly different from the vast majority of the transactions. The key idea of our method consists in selecting the transactions allocated to small clusters

for manual inspection. We have tried this general framework with several hierarchical clustering variants. The main conclusion from an extended experimental comparison of these variants is that hierarchies using the Canberra distance function generally achieve better results according to the performance criteria of this application.

We have also compared our methodology with the results obtained in [11]. This comparison showed that our methodology improves previous results by keeping similar number of erroneous transactions identified with significantly less manual effort required.

Section 2 describes the problem being tackled in more detail as well as the results obtained previously [11]. We then describe our proposal in Section 3. Section 4 presents the experimental evaluation of our method and discusses the results we have obtained. In Section 5 we relate our work with others and finally we present the main conclusion of this paper in Section 6.

## 2 Background

In this section we describe the general background, including the problem (Section 2.1) and previous results (Section 2.2), that provide the motivation for this work.

### 2.1 Foreign Trade Transactions

Transactions made by Portuguese companies with organizations from other EU countries are declared to the Portuguese Institute of Statistics (INE) using the INTRASTAT form. Using this form companies provide information about each transaction, namely:

- 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 labeled "NC", below the row with the column names.[3] Errors often occur in the process of filling forms. For

[3] Note that, in 1998, the Portuguese currency was the *escudo*, PTE.

Trade with EU countries - Detailed Declaration
IMPORT (1998)

| O R | F L U | N LOTE | N FORM | N OPERATOR | M O N | N TRA | CNT | WEIGHT (KG) | COST (kPTE) | COST/WEIGHT (PTE/KG) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 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 were modified to preserve confidentiality.

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 in Euro instead of kEuro. 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. The corresponding forms were analyzed and it was concluded that the second transaction 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. Note that computational efficiency is not important because the automatic system will hardly take longer than half the time the human expert does.

### 2.2 Previous Results

In [11], a first approach to this problem only took the Cost/Weight attribute into account. The data used in that study contained transactions of 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 [8] and Fisher's clustering algorithm [5]. The third one, Knorr & Ng's cell-based algorithm [7], is an outlier detection algorithm which, despite being a multivariate method, was used only on the Cost/Weight attribute. The last is C5.0 [9], a multivariate technique for the induction of decision trees.

Although C5.0 is not an outlier detection method, it obtained the best results. This was achieved with an appropriate transformation of the variables and by assigning different costs to different errors. As a result, 92% of the errors were detected by analyzing just 52% of the transactions. However, taking advantge of the fact that C5.0 can output the probability of each case being an outlier, we can order the transaction by this probability. Using this procedure it is possible to obtain a result of 90% / 40%.

The clustering approach based on Fisher's algorithm was selected because it finds the optimal partition for a given number of clusters of one variable. It was applied to all the transactions of an article, described by a single variable, Cost/Weight. The transactions assigned to a *small cluster*, *i.e.*, a cluster containing significantly fewer points than the others, were considered outliers. The distance function used was Euclidean and the number of clusters was $k = 6$. A small cluster was defined as a cluster with fewer points than half the average number of points in the $k$ clusters. The method was applied to data relative to two months and selected 49% of the transactions which included 75% of the errors. As the results did not accomplished the goals set by the domain experts, the method was abandoned.

## 3 Hierarchical Clustering for Outlier Detection

We describe an outlier detection methodology which is based on hierarchical clustering methods. The use of this particular type of clustering methods is motivated by the unbalanced distribution of outliers versus "normal" cases in these data sets. By definition, outliers are rare occurencies, thus representing a very low fraction of the total examples. In almost all attempts to create the initial clusters, non-hierarchical clustering methods would spread the outliers across all clusters. Given that most of those methods strongly depend on the initialization of the clusters, we expect this to be a rather unstable approach. Therefore, we use hierarchical clustering methods, which are not dependent on the initialization of the clusters.

### 3.1 Proposed Methodology

The key idea of our proposal is to use the size of the resulting clusters as indicators of the presence of outliers. The basic assumption is that outlier observations, being observations with unusual values, will be distant (in terms of the metric

used for clustering) from the "normal" and more frequent observations, and therefore will be isolated in smaller clusters.

The following algorithm outlines the main steps of our proposal for identifying outliers in a dataset using a hierarchical clustering method.

**Algorithm 1** *FindOutliers*

INPUT:
$DATA$, a dataset with $k$ variables and $n$ observations;
a distance function $d$;
a hierarchical algorithm $h$;
$nc$ a number of clusters to use (entailing a level of cut of the hierarchy);
a threshold $t$ for the size of small clusters.

OUTPUT:
$Out$, a set of outlier observations.

$Out \leftarrow \phi$
Obtain the distance matrix $D$ by applying the distance function $d$
to the observations in $DATA$
Use algorithm $h$ to grow an hierarchy using the distance matrix $D$
Cut the hierarchy at the level $l$ that leads to $nc$ clusters
FOR each resulting cluster $c$ DO
IF sizeof($c$) $< t$ THEN
$Out \leftarrow Out \cup \{obs \in c\}$

This algorithm has several parameters that need to be specified. In this paper we try several possible values of these settings and compare their performance experimentally as described in Section 4.2.

One of those parameters, the number of clusters, $nc$, is particularly important for the task we are addressing here, outlier detection. In clustering applications, the decision of the level of cut in a hierarchy is usually guided by some kind of intra-cluster similarity and/or inter-cluster dissimilarity. Our objective here is different. We want to ensure that outliers are isolated in small clusters. The level of cut must be set such that a reasonable number of clusters is obtained. If the number of clusters is small, particularly in large datasets, it will most probably lead to the outliers not being isolated in separate clusters and thus being included in clusters with "normal" observations. This would imply that lots of "normal" observations must be inspected and would increase the cost of the solution (*i.e.* too many false positives are selected). On the other hand a large number of clusters, particularly on smaller data sets, could lead to selecting several clusters containing only "normal" observations, thus once again increasing the number of false outliers (false positives). In our work we have set the number of clusters using the following formula,

$$nc = \max(2, n/10) \tag{1}$$

where $n$ is the number of observations of the data set.

This is an heuristic formulation motivated by trying to make the number of clusters depend on the size of the data set.

### 3.2 Application to the INTRASTAT data

The INTRASTAT data set has some particularities that lead to a few modifications to the methodology outlined in the previous section. These modifications aim at incorporating the experts' domain knowledge so that the methodology better meets their requirements.

According to INE's expertise the items should be inspected separately due to the rather diverse products that may be at state. As such we have applied our algorithm to the set of transactions of each item in turn. Since the number of transactions for each item varies considerably the level of cut of the obtained hierarchies (*c.f.* Equation (1)) also varies. Namely, the hierarchies of items with more transactions are cut in a level leading to a larger number of clusters.

Our methodology is designed for multivariate analysis. However, following a suggestion of domain experts we have focused our study in a single variable, $Cost/Weight$. Domain experts give particular attention to this variable as they believe it is the most efficient variable for detecting the errors.

Following experts' advise we have considered all items with few transactions (10 according to the experts) for manual inspection.

In order to evaluate the validity of the resulting methodology we have taken advantage of the fact that the data set given to us had some information concerning erroneous transactions. In effect, all transactions that were inspected by the experts and were found to contain errors, were labeled as such. Taking advantage of this information we were able to evaluate the performance of our methodology in selecting these erroneous transactions for manual inspection. The experts were particularly interested in two measures of performance: Recall and Percentage of Selected Transactions.

Recall ($\%R$) can be informaly defined in the context of this domain as the proportion of erroneous transactions (as labeled by the experts) that are selected by our models for manual inspection. Ideally, we would like our models to select for manual inspection a set of transactions that included all the transactions that were previously labeled by the experts as errors. INE experts mentioned the value of 90% as the target for this performance measure.

Regarding the percentage of selected transactions ($\%S$) this is the proportion of all the transactions that are selected for manual inspection by the models. This statistic measures the savings of the methodology: the lower this value the more manual effort is saved. INE experts mention 50% as a desirable target for this statistic.

## 4 Experimental Results

In this section we perform an experimental evaluation of our method. Firstly, we present the results of a comparative evaluation between several hierarchical

clustering variants. Secondly, we compare our approach to the previous studies that addressed this data set.

In this comparative studies we have used data from the INTRASTAT database regarding the exportation transactions that were processed in January, February, March, May, June, August, September and October of 1998.

### 4.1 Experimental methodology

Given that INE processes the data in a monthly basis we have decided to use this very same logic in our tests. This methodology will also enable us to compare our results with the results obtained in [11], where the same strategy was followed. The experimental methodology is better described by the following algorithm:

**Algorithm 2**

$TotTrans \leftarrow \|TESTDATA\|$
$TotErrors \leftarrow \|\{tr \in TESTDATA : \text{label}(tr) = error\}\|$
$TotInsp \leftarrow 0$
$TotRight \leftarrow 0$
For each month $m$ Do
    For each item $i$ Do
        $D \leftarrow$ Transactions of month $m$ of item $i$
        If $\|D\| < 10$ Then
            $Insp_{m,i} \leftarrow D$
        Else
            $Insp_{m,i} \leftarrow FindOutliers(D)$
        $TotInsp \leftarrow TotInsp + \|Insp_{m,i}\|$
        $TotRight \leftarrow TotRight + \|\{tr \in Insp_{m,i} : \text{label}(tr) = error\}\|$
$\%R \leftarrow TotRight/TotErrors$
$\%S \leftarrow TotInsp/TotTrans$

This algorithm calculates the value of the two evaluation statistics, $\%R$ and $\%S$, for the testing period, provided we specify the hierarchical clustering parameters necessary to run the function $FindOutliers$ (*c.f.* Algorithm 1).

### 4.2 Comparing different variants of hierarchical clustering

Our first objective in the experimental evaluation of our proposal was to check the impact of the hierarchical clustering parameters on the evaluation statistics. With that goal in mind we have run Algorithm 2 for different clustering setups.

Regarding distance functions we have experimented with Euclidean and Canberra functions. Other alternatives like Manhattan and Maximum distance functions would produce similar results to these distances in a univariate setting like our INTRASTAT data set.

With respect to clustering algorithms we have used three alternatives all available in the R statistical environment [10].[4] We have used the *hclust* algo-

[4] http://www.r-project.org

rithm from package *mva*, and algorithms *agnes* [6] and *diana* [6] from package *cluster*. *Hclust* and the *agnes* are agglomerative algorithms (*i.e.* bottom-up). *Hclust* was tried with seven different agglomeration techniques: average, centroid, complete, mcquitty, median, single and ward. *Agnes* was used with five agglomeration techniques: average, complete, single, ward and weighted. The *diana* algorithm is divisive (*i.e.* top-down) and at each step it divides the cluster with largest dissimilarity between any two of its observations.

All these variants of distances and algorithm variants make up a total of 26 different alternatives that were compared on the same data.

Having defined the 26 different hierarchies that we will compare we still need to define another important parameter of Algorithm 1, the size of the clusters considered as containing outliers (parameter $t$). In the present comparative study we have tried the values 5 and 10.

The results of the 26 hierarchies for each of the two $t$ values, in terms of the two statistics relevant for the domain experts, $\%R$ and $\%S$, are show in Figure 2

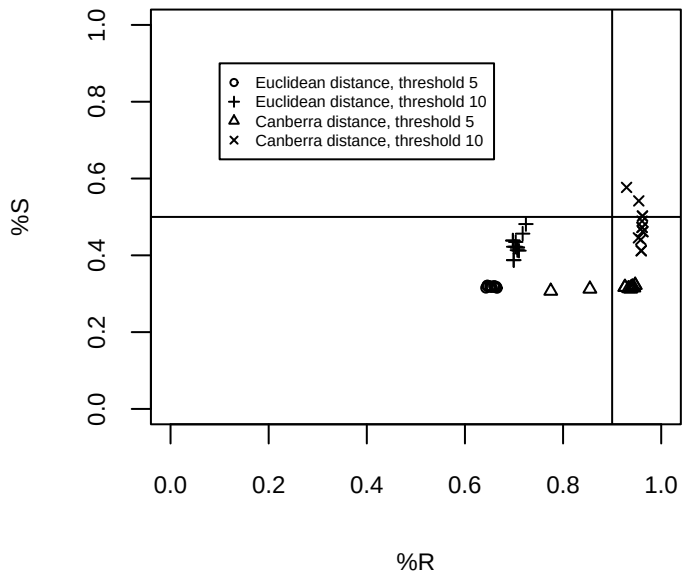


**Fig. 2.** The $\%R$ and $\%S$ of the different variants of hierarchical clustering methods.

The plot indicates that the distance function is the parameter that affects results the most. The variants that use Canberra distance systematically outperform the variants that used Euclidean distance. In fact this tendency is so strong that we could even talk about two groups of performance: the group formed by Euclidean-based variants that usually achieve recalls from 60% to 80%; and the Canberra-based variants that usually obtain recalls between 90% and 99%. The differences in the proportion of selected transactions are not so evident but Canberra-based alternatives usually select more observations. Canberra distance function gives more relevance to diferences between values with low absolute value. According to INE experts errors in transactions with low absolute value are particularly serious for them. This could explain the better score obtained by Canberra-based variants.

Regarding the issue of the cluster size threshold, $t$, the value of 5 generally produces slightly better results, particularly in the case of Caberra variants.

In general, the results obtained with Caberra variants with the cluster size of 5 are quite good, achieving the objectives of the domain experts of %R higher than 90% and %S lower than 50%.

### 4.3 Comparing our results with previous results on the same data

As we have mentioned in Section 2.2 the INTRASTAT data was already explored using several data mining techniques. The experimental setup and data used in this previous work [11] was the same as ours. In this section we compare our results with the results obtained in this previous attempt at detecting erroneous transactions.

With respect to our variants it would not be correct to select the best results for the comparison with the other methods. As such, we have calculated the average score in terms of %S and %R for each distance / cluster size threshold combination, leading to four representative scores of our methodology.

The results of this comparison are shown in Figure 3.

The result of the hierarchies using Canberra distance and cluster size of 5 are clearly better than all other results. This variant presents one of the highest mean %R and simultaneously one of the lowest %S.

This group of hierarchies is only outperformed in terms of %R - and by a close margin - by the hierarchies that used the same distance and value 10 for parameter $t$ and C5.0, but at the cost of much more manual inspection (higher %S values)

## 5 Related Work

Most existing work on detecting outliers comes from the field of statistics and it is usually based on distribution properties (*e.g.* [1]) of a single variable. These studies use standard distributions (*e.g.* Normal) to fit the data and detect outliers. This is one of the main drawbacks of these approaches as most data mining problems do not fit these standard distributions.

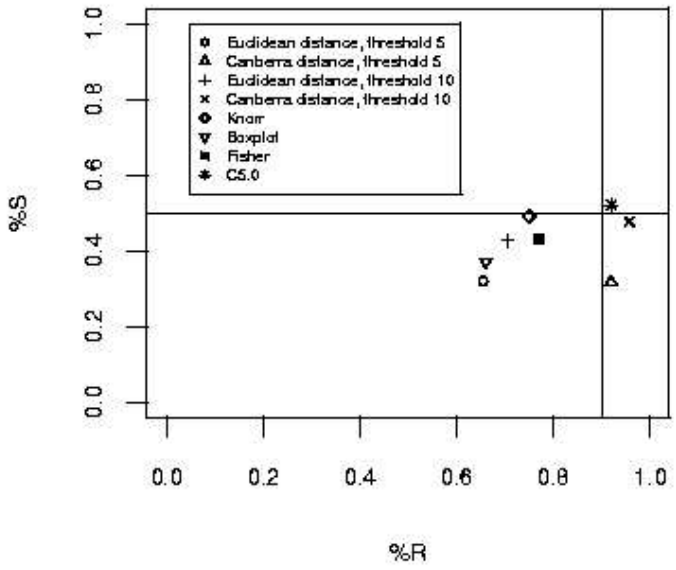


**Fig. 3.** The comparison with results of previous approaches to the INE problem.

The vast majority of work that mentions outliers does it with the purpose of developping modelling techiques that are robust to their presence. This means that outliers are not the focus of modelling, as in our work.

Zhang and colleagues [15] developped a clustering method (BIRCH) able to handle outliers. Still, once again outlier identification was not the main purpose of this work.

Knorr and Ng [7] introduced the notion of distance-based outliers. This work is strongly related to ours and it was one of the methods compared to ours in Section 4.3.

Breunig and colleagues [2, 3] described an approach based on the same theoretical foundations as density-based clustering. They defined a notion of Local Outlier Factor (LOF) for each observation. The key idea of this work is that the notion of outlier should be "local" in the sense that the outlier degree of any observation should be determined by the clustering structure in a bounded neighborhood of the observation. These authors claim that this notion is able to detect a larger number of outlier types than other approaches.

Other works have looked at the problem of outliers from a supervised learning perspective (*e.g.* [11, 12]). Usually, the goal of these works is to classify a given

observation as being an outlier or a "normal" case. These approaches may suffer from problems of strong unbalanced class distributions which may deteriorate the performance of the models [13]. Moreover, in the INE application the labels are not completely trustable in particular the ones concerning the "normal" cases that may actually be not inspected transactions.

A related field of research is activity monitoring, namely with the purpose of detecting frauds. This task consists of monitoring a online data source in the search for unusual behavior (*e.g.* [4, 14]). In the case of INE it does not make sense to consider the data as an online source, as the procedure is to search for errors at the end of each month.

## 6 Conclusions

In this paper we have presented a method for the detection of outliers using an hierarchical clustering approach. This method uses the size of the resulting clusters as the key factor for identifying groups of observations that are distinct from the majority of the data.

We have applied our methodology to the task of detecting erroneous foreign trade transactions in data collected by the Portuguese Institute of Statistics (INE). The results of the application of our method to this problem clearly met the performance criteria outlined by the human experts. Moreover, our results outperform previous approaches to this same problem.

We have also presented the results of a comparison of several variants of hierarchical clustering algorithms in the context of our methodology. These experiments were designed with the goal of understanding the sensitivity of our method to changes in its various parameters. The results of this comparison indicate that in the INE problem the performance of the method is dependent on the used distance function. Hierarchies using the Canberra distance function clearly outperform other alternatives.

In the future we intend to explore several forms of automatically selecting the best parameters for our methodology. Namely, we will use a sliding window approach were data from previous months is used to select the best alternative, which wil then be used for identifying the erroneous transactions of the current month. Finally, we also plan to apply our method to other problems, namely to multivariate data.

## References


1. V. Barnett and T. Lewis. *Outliers in statistical data*. John Wiley, 1994.
2. M. M. Breunig, H. P. Kriegel, R. Ng, and J. Sander. Optics of: Identifying local outliers. *Lecture Notes in Computer Science*, 1704:262–270, 1999.
3. M. M. Breunig, H. P. Kriegel, R. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *Proceedings of ACM SIGMO 2000 International Conference on Management of Data*, 2000.

4. Tom Fawcett and Foster Provost. Activity monitoring: Noticing interesting changes in behavior. In Surajit Chaudhuri and David Madigan, editors, *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 53–62. ACM, 1999.
5. W.D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53:789–798.
6. L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.
7. Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of 24rd International Conference on Very Large Data Bases (VLDB 1998)*, pages 392–403. Morgan Kaufmann, San Francisco, CA, 1998.
8. J.S. Milton, P.M. McTeer, and J.J. Corbet. *Introduction to Statistics*. McGraw-Hill, 1997.
9. R. Quinlan. *C5.0: An Informal Tutorial*. RuleQuest, 1998. http://www.rulequest.com/see5-unix.html.
10. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. ISBN 3-900051-00-3.
11. C. Soares, P. Brazdil, J. Costa, V. Cortez, and A. Carvalho. 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.
12. L. Torgo and R. Ribeiro. Predicting outliers. In N. Lavrac, D. Gamberger, L. Todorovski, and H. Blockeel, editors, *Proceedings of Principles of Data Mining and Knowledge Discovery (PKDD'03)*, number LNAI in 2838, pages 447–458. Springer, 2003.
13. G. Weiss and F. Provost. The effect of class distribution on classifier learning: an empirical study. Technical Report ML-TR-44, Department of computer science, Rutgers University, 2001.
14. Gary Weiss and Haym Hirsh. Learning to predict extremely rare events. In *AAAI Workshop on Learning from Imbalanced Data Sets*, pages 64–68. Technical Report WS-00-05, AAAI Press, 2000.
15. T. Zhang, R. Ramakhrishnan, and M. Livny. Birch: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada*, pages 103–114. ACM Press, 1996.