Finding Fraud in Health Insurance Data with Two-Layer Outlier Detection Approach

Department of Computer Science, VU University Amsterdam
{rmkonijn,wojtek}@few.vu.nl

Abstract. Conventional techniques for detecting outliers address the problem of finding isolated observations that significantly differ from other observations that are stored in a database. For example, in the context of health insurance, one might be interested in finding unusual claims concerning prescribed medicines. Each claim record may contain information on the prescribed drug (its code), volume (e.g., the number of pills and their weight), dosing and the price. Finding outliers in such data can be used for identifying fraud. However, when searching for fraud, it is more important to analyse data not on the level of single records, but on the level of single patients, pharmacies or GP's.

In this paper we present a novel approach for finding outliers in such hierarchical data. Our method uses standard techniques for measuring outlierness of single records and then aggregates these measurements to detect outliers in entities that are higher in the hierarchy. We applied this method to a set of about 40 million records from a health insurance company to identify suspicious pharmacies.

1 Introduction

The inspiration for this paper comes from a real life fraud detection problem in health insurance, in the pharmacy domain. The goal of fraud detection in this context is to identify the most suspicious pharmacies that could possibly be involved in fraudulent activities, rather than identifying single claims that are suspicious. The main reason for not focusing on single outliers is that recovering money from single claims is costly, and that it can harm the relationship between an insurance company and the involved pharmacy, especially in the case of false positives. On the other hand, if the insurance company can detect substantial fraud linked to multiple claims of the same pharmacy, this business relationship is no longer so important and a vigorous money recovery action can follow.

In contrast to typical approaches for finding single outliers, [6], we propose a novel method for finding *groups* of outlying records that belong to the same class. Our method was successfully applied to a large set of health insurance claims, helping to identify several pharmacies involved in fraudulent behaviour.

Our method for detecting group outliers works in two stages. In the first stage we calculate outlier scores of single records. We use here classical methods for outlier detection that are based on distance measures, [2], or density estimation, [5].

Next, we calculate a statistic to measure the *outlierness* of each groups of records, where groups form logical entities. In our case, each entity is formed by all claims

A. Cuzzocrea and U. Dayal (Eds.): DaWaK 2011, LNCS 6862, pp. 394–405, 2011.

related to a pharmacy, or a combination of a pharmacy and a type of medication. We propose four different statistics that are used to define the final outlier score of these entities: (1) a rank-based statistic, (2) a weighted rank-based statistic, (3) a statistic based on the binomial distribution, and (4) a statistic that is based on the mean of the outlier score. These statistics can be applied in different situations to different outlier scores.

The statistics can be computed over different segments of the data to obtain the final score. Extra information about outlying entities can be obtained by constructing, for each entity, a so-called *fraud set*: a set of suspicious claims from a given entity. A fraud set is a minimal set of outlying records that should be removed from the whole set in order to make it "normal" again. Another, very useful instrument for displaying fraud evidence is a *fraud scatter plot*. Each point on such a plot represents a single entity; the x and y coordinates of a point are, respectively, the outlier score of the corresponding fraud set and the total amount of money involved in this fraud set, *fraud amount*. The fraud scatter plot can be used by fraud investigators to decide whether they should investigate the most likely fraud cases, or to focus on cases that are less suspicious, but involve high amounts of money.

Our paper is organized as follows. We start with a brief overview of related work. Then we present two approaches for calculating outlier scores of single records: distance-based and density-based. In Section 4 we explain four methods for aggregating individual scores, a procedure for identifying fraud sets, and a method for visualizing results with help of the fraud scatter plot. Results of our experiments are presented in Section 5, while the last section contains conclusions and some recommendations for further research.

2 Related Work

There is a lot of literature about methods for detecting single outliers in data. They are extensively presented in general survey articles on outlier detection techniques: [7], [1], and [6].

The method described in this paper can be categorised as unsupervised outlier detection. Existing methods for unsupervised outlier detection (a missing label problem) can be split into the following categories: statistical methods and distance-based methods, with the later containing the sub-categories of depth-based methods, density-based methods, and clustering-based methods.

Depth-based methods measure the distance from a point to the center of the data. Points that have the highest distance are considered outliers. There are several definitions of depth, for example the Mahalanobis Depth, which is equal to the distance to the Mahalanobis distance to the mean of the data. Because outliers have a big impact on the location of the mean and the covariance matrix estimate, a robust estimate of these statistics can be used, [12]. The main disadvantage of depth-based methods is their inability of handling clusters in data – these methods assume that the data form a single cluster.

Distance-based methods require a distance measure to determine the distance between two instances. The main idea is that the distance between outlying instances and their neighbors is bigger than the distance between normal instances and their neighbors, [8].

Distance-based methods compare distances with respect to the whole dataset. Outlier score measures that are based on the distances between a specific point and points in its local neighborhood are called density-based methods. Examples are the Local Outlier Factor (LOF), [5], the Connectivity-based Outlier Factor (COF), [13], or the Multi-granularity Deviation Factor (MDEF), [10].

In the statistical community some methods have been investigated to detect multiple outliers at once. Based on a model fitted on the data, outliers are observations that deviate from the model, or that would deviate if the model were fitted without the observation (so-called a *deletion diagnostic*). There are two related issues, called *masking* and *swamping*, that have been investigated in [4]. Masking takes place in a situation when an outlier would not be revealed by calculating a single deletion diagnostic measure for each observation, but it would be detected by a multiple deletion diagnostic. The opposite situation, swamping, occurs when a pair of observations is declared anomalous only because one of the two is extreme: the bigger deviating observation *swamps* the smaller one.

To our best knowledge, the problem of finding a group of outliers that belong to the same entity (such as a pharmacy) has not been addressed yet in the existing literature.

3 Outlier Score for Single Records

In this section we present in more depth two approaches for calculating outlier scores for single records: distance-based and density-based. We start with some definitions and notations. Let D denote a set of n objects (called records or points) and let d denote a distance measure between these objects. The $k\text{-}distance$ of p, denoted as $k\text{-}distance(p)$, is the distance of p to its k-th nearest neighbor. The $k\text{-}distance\ neighborhood$ of p contains every object whose distance from p is not greater than the $k\text{-}distance(p)$. These objects are called the k-th nearest neighbors of p and are denoted by $S_k(p)$.

Distance-Based Scores. Distance-based methods are based on the proximity of points to each other according to a distance measure. There are several definitions possible that can be used to identify outliers. These definitions are usually based on the concept of the k-nearest neighbor, [11]. An object p is called a (k, n) outlier if no more than n-1 other points in the dataset have a higher value of the $k\text{-}distance$ than the point p itself. Note that this is a binary score: the top n points with the highest values of $k\text{-}distance$ are declared as an outlier, while all other observations are considered normal. Another definition is given in [8], who defines a $DB(perc, distance)$ outlier as follows: an object p in a dataset D is a $DB(perc, distance)$ outlier if at least fraction $perc$ of the objects in D lies further from p than $distance$. In other words, $distance$ can be seen as a radius around p, and if the percentage of points within this radius is smaller than $(1 - perc)$, p is declared anomalous. A yet another definition, [2], assigns a weight to each point p, which is defined as the sum of the $k\text{-}distance$ of all points within the $k\text{-}distance\ neighborhood$ of p. Outliers are those points that have the biggest weight. There are some small differences between the three definitions given above. The first definition by [11] does not provide a ranking of the outliers. For the definition by [8] it may be hard to set the parameters appropriately. The definition of [2] overcomes

these problems, but is computationally expensive. We used in our experiments this later definition of the scoring function.

Density-Based Scores. Another scoring function that we used in our experiments is a modification of the well-known LOF score, [5], which is based on the idea of the probabilistic distance that is described in [9].

For the explanation of the LOF Score we first need some definitions. Using the same notation as before, the reachability distance of an object p with respect to an object o is defined as:

$$reachDist_k(p,o) = max(k{-}distance(p), distance(o,p)) \tag{1}$$

This distance measure is used to correct for statistical deviations.

The density of each point is called the local reachability density of an object p. It is calculated as follows:

$$lrd_k(p) = (\frac{\sum_{o \in S_k(p)} reachDist_k(p,o)}{|S_k(p)|})^{-1} \tag{2}$$

In other words, the density of p is the average reachability distance from its k-$distance\ neighborhood$ to the point itself. For sparse regions, the value for lrd will be low, for dense regions it will be high.

Finally, the local outlier factor of an object p is defined as:

$$LOF_k(p) = \frac{\sum_{o \in S_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|S_k(p)|} \tag{3}$$

In other words, for an object p we compare its own density $lrd_k(p)$ with the density of the points in its k-$distance\ neighborhood$. If the densities are approximately equal, the LOF score will be close to one, if the density of p is relatively low, the LOF score will be high.

In our experiments we used a modified version of the LOF score, because it turned out to work better than other methods in detecting single outliers. We used the probabilistic distance, as defined in [9], to determine the *reachability distance*:

$$probReachDist_k(p,o) = max(pdist_{k,\varphi}(p), distance(o,p)), \tag{4}$$

where $pdist_{k,\varphi}(p)$ denotes the probabilistic distance of p to its k neighborhood, as measured within the radius φ, i.e., the minimum distance for which φk neighbors of p are covered.

The formulas for calculating the local reachability density and the LOF score remain the same. Note that using the probabilistic distance can also be seen as using two parameters: k_1 to determine a context set S which is used to compare densities, and $k_2 = \varphi k_1$ to calculate the distances between points and eventually their densities.

4 Statistics per Entity

In this section we address our main problem: detection of groups of outliers that belong to the same entity. The proposed approach for this problem involves two steps: (1)

calculation of outlier scores of all records, and (2) calculation of a statistic to measure the outlierness of each entity. In this section we present four different statistics for measuring the entity outlierness.

Each of these statistics is used to quantify the difference between two samples: the set of scores of records belonging to the entity and the set of scores of records that do not belong to the entity. Most outlier measures do not have a direct probabilistic interpretation. Also the range of scores strongly depends on the data set, or even on the scaling of the data set. For some outlier measures only the rank is important, while for others we are mainly interested in relative values. Furthermore, different kinds of fraud are possible. In the case of pharmacies, all fraud can be committed in a single claim or with charges concerning a single patient, but the fraud can also be distributed over many charges, charging just a little more per claim. This is why different statistics are needed. We present four different statistics that can be used under different circumstances: a rank based statistic, a weighted rank based statistic, a statistic based on the binomial distribution, and a statistic that is the standardised residual. The binomial outlier score is different from the other three statistics because of the fact that it does not take the ordering of the outlier scores into account. This statistic works well in combination with single scores that provide a list of top-n outliers, or that provide a binary outlier score. The other three statistics mainly differ in robustness against the outlier score values. The ordering from least robust to most robust is: 1) standardised residual 2) weighted rank outlier score and 3) rank-based outlier score. The positive aspect of using a robust score to aggregate per entity, is that it is not affected by a very high score of one single point thereby declaring the whole entity anomalous. On the other hand, such a single point with a very high score may also be very interesting, which would favor the use of a non-robust score.

Additionally, we describe how to incorporate the monetary value that is related to analysed records in the detection process, and demonstrate how to construct *fraud sets*. Finally, we show how a *fraud scatter plot* can be used to support decisions concerning further investigation of identified suspicious entities.

4.1 Statistics per Entity

In this section we introduce several statistics to calculate the "outlierness" of an entity with respect to all other entities. The common idea behind all these statistics involves measuring the difference between two sets of numbers: a set of scores of all records from one entity and a set of scores of all other records from remaining entities. More precisely, let us suppose that our dataset has n records. For each record, we calculate an outlier score, so we have in total n outlier scores. Let us consider a single entity that we want to compare to other entities. The set of n scores can be split into $X_1, \ldots, X_{n1}$ and $Y_1, \ldots, Y_{n2}$, where $X_1, \ldots, X_{n1}$ are the scores of records from the entity under consideration, and $Y_1, \ldots, Y_{n2}$ are the scores of the remaining records. Now our problem can be formulated as follows: how to measure the difference between X and Y? In our experiments we used the following four methods of comparing X to Y.

Wilcoxon Mann-Whitney test with single outlier score. The first method is based on the popular, non-parametric two-sample test, called the Mann-Whitney-Wilcoxon rank-sum test, [3]. It defines the outlierness score of an entity as the p-value that is returned

by the Mann-Whitney-Wilcoxon test when comparing values of sets X and Y to each other.

Weighted rank outlier score. The Mann-Whitney-Wilcoxon test uses only ranks of the scores and not their actual values. However, we can weight the ranks of elements in X and Y by their values: the bigger the outlier the bigger its impact on the final entity score. More precisely, we define:

$$Z_{ij} = \begin{cases} 0 \text{ if } X_i < Y_j \\ \frac{Y_j}{\sum_{k=1}^{n_2} Y_k} \text{ if } X_i > Y_j \end{cases} \tag{5}$$

and

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} Z_{ij} \tag{6}$$

For large n we can assume U to be normally distributed and parameters of this distribution can be calculated from the vector of the partial sums of the sorted vector Y. Given these parameters, one can easily find the corresponding p-value.

Binomial outlier score. The calculation of this score starts with calculating the sets of scores X and Y, as described earlier. Then both sets are combined and sorted. The top p percentage of scores are viewed as outliers, where p is a pre-specified parameter. Under this definition of an outlier, the number of outliers that belong to the set X follows a binomial distribution with expected value $n_1 p$ and variance $n_1 p(1-p)$. The outlier score of X (relative to p) is now defined as $1 - cpdf(binomial(n_1, p), k)$, where k is the number of observed outliers in X, i.e., the mass of the right tail of the binomial distribution with the parameters n_1 and p that starts at k.

The value of parameter p is used to determine the percentage of records that are viewed as outliers. It can be set in different ways. In some cases the value of p is determined by a domain expert. The choice of p can also be based on the probability distribution of the outlier score function. One can approximate this distribution by using a histogram with two bins: one bin for 'low' outlier scores, and another one for 'high' outlier scores. The observations that fall into the bin of 'high' outlier scores are labeled as outliers, so p is the splitting point between the two bins. We estimate p by minimising the Kolmogorov-Smirnov distance, [3], between the distribution of the outlier score and the 'approximate' two-bin distribution. Another possibility is to use a heuristic that is based on the parameter p: for example, take the maximum outlier score per entity for a range of values for p. The disadvantage of this approach is that the final outlier entity score cannot be interpreted as a probability anymore.

Standardized residual of outlier score. This measure of entity outlierness is defined in terms of the average deviation from the mean of the outlier scores that belong to the given entity. The average standardized residual should follow a normal distribution. The corresponding p-value – the mass of the tail on the right from the observed value – is the outlier score.

4.2 Identifying Fraud Sets

Each approach described above uses a statistic U to describe the deviation of an entity. Because U follows a normal distribution, we can easily test the hypothesis that the observed value for U is equal to $E(U)$ with significance level α. This hypothesis will be rejected for the most outlying entities with the highest value of U. Suppose this hypothesis is rejected for an entity with a set of observations X. We define a *fraud set* for an entity X as the minimal set of records that should be removed from X in order to make the null hypothesis that the observed value of $U(X)$ is equal to $E(U(X))$ plausible (i.e., not to be rejected at a given significance level). Because the observations that should be removed are the ones with the highest outlier score, the fraud set is also the set of observations that should be investigated first, when checking if the entity is really outlying.

4.3 The Fraud Scatter Plot

Another very useful instrument for displaying fraud evidence is a *fraud scatter plot*: a graph of fraud amount versus outlier score of all records in the fraud set. Here, the *fraud amount* is defined as the total amount of money that is involved in the observations that are in the fraud set. The fraud scatter plot can be used by fraud investigators to decide whether they should investigate the most likely fraud cases, or to focus on cases that are less suspicious, but involve high amounts of money.

More precisely, for an arbitrary significance level α, the fraud scatter plot contains points, one per entity, with their x-coordinates being the outlierness score of an entity (we use the z-score of the observed value of U), and the y-coordinate being the the fraud amount.

4.4 Aggregation of Scores for Data Segments

In many applications financial transactions can be split into a number of segments. For example, claims can be organized into categories that are determined by the type of medicine involved, and patients allocated to segment per disease type. Each of the four statistics described earlier can be calculated for each segment and the resulting scores aggregated on the level of single entities. For the Wilcoxon-Mann-Whitney test and the statistic based on the binomial distribution, a normal approximation can be obtained (the other two statistics are already normally distributed). Let s_i be the z-score of an entity per segment, and S be $\frac{\sum_{i=1}^{n} s_i}{n}$, where n is the number of segments. The final outlier score of an entity is defined as $\Phi(S)$, where Φ is the cumulative probability density function (cpdf) of the standard normal distribution. This aggregation is not needed if there are no subsegments in the data.

5 Results and Analysis

Now we will describe some results that we obtained when applying our method to a relatively big set of 40 million records related to claims submitted by pharmacies. Each record contained information about the pharmacy (pharmacy ID), the prescribed medicine (type, subtype, product ID), cost, dosage, et cetera. In our research we have focused on three types of deviations: unusual prescriptions, errors that seem to be typos, and unusual number of "expensive" patients.

5.1 Strange Behavior in Prescribing Drugs

A common type of fraud in health insurance is called *unbundling*: a practice of breaking what should be a single charge into many smaller charges. The 'standard' formula for a single claim is $price = c + p * n$, where $price$ is the claim amount, c is a constant charge per claim, n is the number of units, and p is the price per unit. A pharmacy can commit unbundling fraud by splitting the charge into two or more charges, thereby earning the constant amount c twice (or more times). Two other common types of fraud are: delivering more units than stated on the prescription (and thus increasing turnover), and charging money for drugs that have never been delivered.

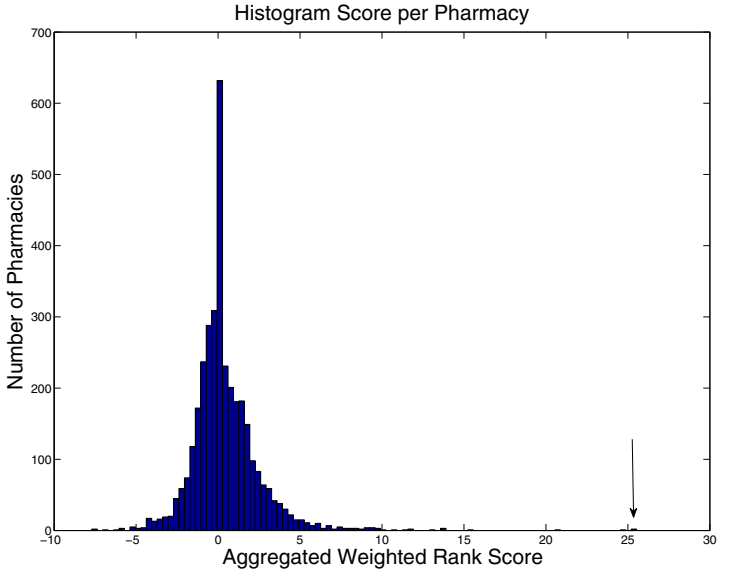

Fig. 1. Histogram of the Weighted Rank Score Statistic. There are two observations with a score higher than 25. We can also see that the distribution of the statistic is not completely normal due to some outliers and due to a peak around zero (these are pharmacies with very few claims).

First we split the data into different segments, one segment per drug type. For each segment we use the following variables for each patient X: X_1: the total number of units (pills) used by a patient within a year, X_2: the total claim amount of a patient within a year, X_3: the number of claims, X_4: $\sum_{i=1}^{n} a_i$, where n is the number of claims, and a_i is 1 if the patient visited his family doctor within two weeks before the claim, and zero otherwise.

An outlier in these dimensions indicates strange or fraudulent claim behavior. We calculate an outlier score for single observations first. For this application we use the modified LOF Score, as described in Section 3.

For each medicine type we calculated the weighted rank score, where each patient is assigned to one or more pharmacies. We aggregated all these scores by summing them up and then standardizing them. The final scores of all pharmacies are shown in Fig. 1.

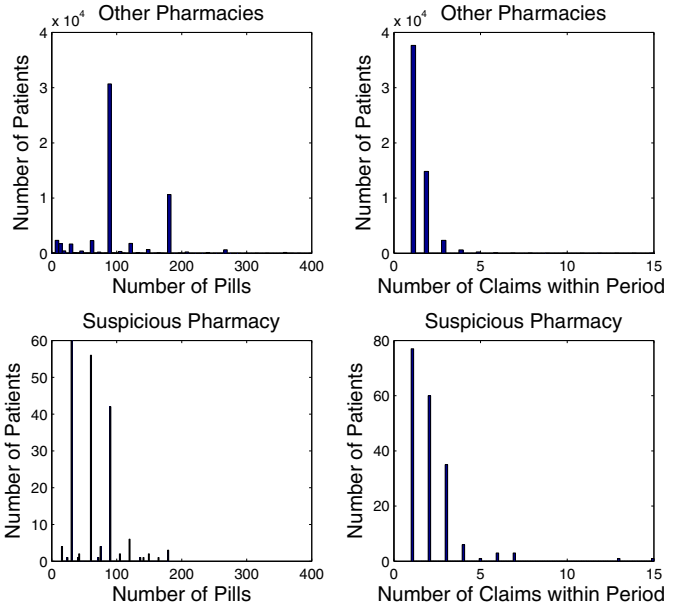

Fig. 2. Histograms of the 'number of pills prescribed' and the 'number of claims' for the drug type Aspirin. The two histograms below show the distribution of patients of the outlying pharmacy. The upper two histograms show the distribution of the other pharmacies. From these graphs it can be concluded that these distributions are different. The number of pills is much lower than expected, while the number of claims is higher: this is a signal for *unbundling* fraud.

We compared data from the two most outlying pharmacies with data from the remaining pharmacies. For the top outlying pharmacy the distributions of variables X_1 (the number of pills) and X_3 for the drug type 'Aspirin' are given in Figure 2.

5.2 Finding Typos

Sometimes pharmacies make mistakes when entering the number of units that is prescribed, thereby 'accidentally' overcharging. We calculated the following z-scores:

- X_1: the standardized claim amount,
- X_2: the claim amount, standardized at the drug type level,
- X_3: the claim amount divided by the total costs of the patient within a year, grouped at the drug type level and then standardized,
- X_4: the claim amount divided by the total costs made by the patient on the same drug, standardized.

Typos will score high within all dimensions, therefore we used as an outlier score for a single record the smallest one: $score = min(X_1, \ldots, X_4)$.

Because this score is really designed for detecting 'top-n' outliers, the binomial statistic to aggregate the outlier scores per pharmacy seems to be the most appropriate.

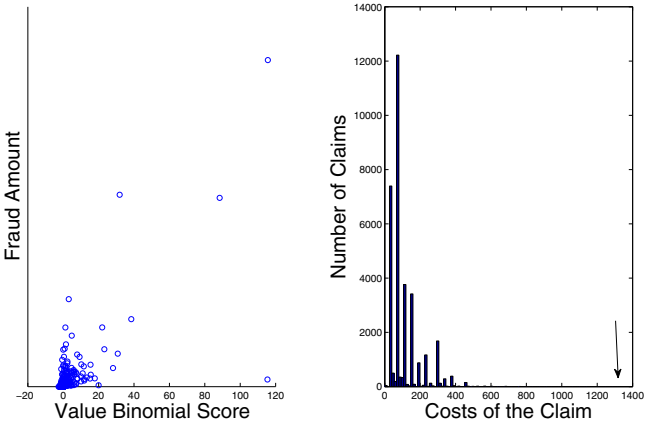

Fig. 3. The graph on the left shows the fraud scatter plot. The score according to the binomial statistic is plotted against the money that is involved in the claims. The most interesting pharmacy is the one in the upper right corner: it has a large deviation and a high fraud amount. The histogram on the right shows an example of an outlier of this pharmacy for the 'Glucose Test strip'. The outlying claim made by this pharmacy is the claim of about 1300 Euros. The pharmacy is outlying because it has many of such claims.

The value of the parameter p can be found after a few trial-and-error attempts followed by a manual inspection of found outliers. The fraud scatter plot and an example outlier are displayed in figure 3.

5.3 Patients with High Claim Costs

Pharmacies may also be delivering more units than stated on the prescription (and thus charging more money). The difference with a typo is that this time the claim amounts are not extremely high, but just a little higher than normal.

To discover this type of outliers we split the data into segments, using one drug type per segment. For each segment we defined the following two variables:

X_1: the claim amount,

X_2: the claim amount divided by the total costs per patient within a year.

We standardized both dimensions. Because we were interested in global outliers we used as an entity outlier score the mean distance to the k nearest neighbors, [2]. Next, we calculated the deviation from the mean statistic per patient per drug type, and aggregated the scores by summing them. Finally, we used the binomial outlier score to aggregate patient scores on the level of pharmacies. We estimated the parameter p by approximating the density of the score per patient by a histogram of two bins, see Figure 4. By inspecting the fraud scatter plot we could conclude that the most interesting outlier is the pharmacy with the highest amount of fraud. For this pharmacy we plotted some of its outlying claims within the drug type 'Erythropoietin' (a.k.a. Epo), see Figure 5. It is evident that those claims of this pharmacy are outliers, because of the high amounts per claim.

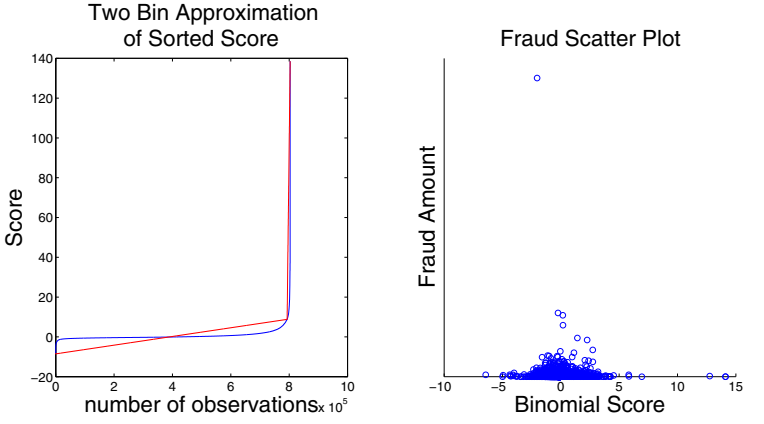

Fig. 4. The graph on the left shows how the cumulative distribution function of the LOF scores is approximated by a histogram of two bins. The bin sizes of this histogram are determined by minimizing the Kolmogorov-Smirnov Distance between the cdf and a function with two linear components. The graph on the right shows the fraud scatter plot. The x-axis shows the deviation from the expected value of the statistic and the y-axis shows the amount of money that is involved within the outlying transactions.

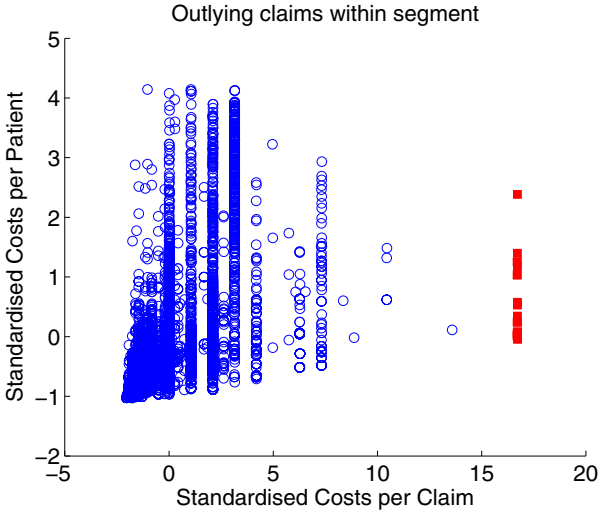

Fig. 5. Scatter plot of some of the outliers of the suspicious pharmacy for the drug 'Erythropoietin' (Epo). The outliers are the red squares on the right. The claims represented by the red squares are all delivered by the same pharmacy.

6 Conclusions and Further Research

We presented a novel approach for finding outlying entities in hierarchical data. Our method uses standard techniques for measuring outlierness of single records and then aggregates these measurements to detect outliers in entities that are higher in the hierarchy. Our approach turned out to work very well in a practical setup, where many fraud cases were detected relatively fast.

Further research will address the issue of adding apriori information about entities (such as pharmacies, hospitals or physicians) into the model. For example, it is well known that some pharmacies (e.g., internet or mail-order pharmacies) exhibit different claim patterns than conventional ones. Discovery and incorporation of this type of information into our method is a challenging research problem.

References

1. Agyemang, A., Barker: A comprehensive survey of numeric and symbolic outlier mining techniques. Intelligent Data Analysis 10(6/2006), 521–538 (2005)
2. Angiulli, F., Pizzuti, C.: Fast outlier detection in high dimensional spaces. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 43–78. Springer, Heidelberg (2002)
3. Bain, Engelhardt: Introduction to Probability and Mathematical Statistics. Duxbury Press, Boston (1992)
4. Barnett, V., Lewis, T.: Outliers in Statistical Data. John Wiley and Sons, Chichester (1994)
5. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. SIGMOD Rec. 29(2), 93–104 (2000)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys (CSUR) 41, 15:1–15:58 (2009)
7. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. 22(2), 85–126 (2004)
8. Knorr, E.M., Ng, R.T.: Algorithms for mining distance-based outliers in large datasets. In: Unknown, pp. 392–403 (1998)
9. Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A.: Loop: local outlier probabilities. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 1649–1652. ACM, New York (2009)
10. Papadimitriou, S., Kitagawa, H., Gibbons, P.B., Faloutsos, C.: Loci: Fast outlier detection using the local correlation integral. In: International Conference on Data Engineering, p. 315 (2003)
11. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. SIGMOD Rec. 29, 427–438 (2000)
12. Rousseeuw, P.J., Driessen, K.V.: A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223 (1999)
13. Tang, J., Chen, Z., Chee Fu, A.W., Cheung, D.: A robust outlier detection scheme for large data sets. In: 6th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, pp. 6–8 (2001)