Informal identification of outliers in medical data

Jorma Laurikkala\textsuperscript{a}, Martti Juhola\textsuperscript{a} and Erna Kentala\textsuperscript{b}

\textsuperscript{a}Department of Computer Science, 33014 University of Tampere, Finland.
Email: jpl@cs.uta.fi
\textsuperscript{b}Department of Otorhinolaryngology, Helsinki University Hospital, Helsinki, Finland.

Abstract. Informal box plot identification of outliers in real-world medical data was studied. Box plots were used to detect univariate outliers directly whereas the box plotted Mahalanobis distances identified multivariate outliers. Vertigo and female urinary incontinence data were used in the tests. The removal of outliers increased the classification accuracy of discriminant analysis functions and nearest neighbour method. Outliers were also evaluated subjectively by expert physicians, who found most of the multivariate outliers to truly be outliers in their area. The experts sometimes disagreed with the method on univariate outliers. This happened, for example, in heterogeneous diagnostics groups where also extreme values are natural. The informal method may be used for straightforward identification of suspicious data or as a tool to collect abnormal cases for an in-depth analysis. Keywords: Outliers, identification, machine learning

1. Introduction

There are many definitions for outliers which differ in words \cite{1,2,3}. We use the one of Barnett and Lewis \cite{1 pp.4}, who defined the outlier in a set of data to be "an observation (or subsets of observations) which appears to be inconsistent with the remainder of that set of data". This type of observations is often a problem in statistical analysis. Outliers may lower the model fit and cause the model to be less robust in predictive analysis. To illustrate, consider fitting a linear regression equation to data shown in Figure 1 \cite{1 pp. 261-263}. The regression line runs nicely through the scatter plot of ages and salaries of electrical engineers (\textit{N}=55, United Kingdom, 1974), but the most extreme observation M has a surprisingly strong impact on the analysis. Dropping M and refitting the regression line elevates the goodness of fit statistics \( R^2 \) from 0.452 to 0.526.

There are various origins of outliers. Human error often produces unintentional outliers. Data entry may be incorrect and missing value codes are sometimes used as real data \cite{1,3}. Outliers are frequently generated as the result of the natural variation of population or process one can not control. These outliers are from the intended population, but their values are unusual in comparison to the normal values. It is also possible to have an outlier that is not a member of population due to a sampling error.

Outlier identification (and consequent removal or accommodation) is part of the data screening process which should be done routinely before statistical analyses \cite{2,3}. The simplest and the most researched case is the identification of univariate outliers, where the distribution of a single variable is examined \cite{1}. Extreme data values are obvious
outlier candidates. When the distribution is symmetric, we suspect that candidate outliers are the extremes of the left or right tail. Correspondingly, the identified outliers are referred to as the lower and upper univariate outliers. In a skewed distribution the suspect outliers are likely to be the extremes of the longer tail (see Figure 2). Multivariate outlier detection is more difficult, because the multivariate distribution has no tails [1,4]. Multivariate outliers, such as engineer M, are sometimes also univariate outliers. However, multivariate outliers are not necessarily univariate outliers, because unusual combinations of normal values may cause the case to be a multivariate outlier.

Figure 1: Ages and salaries of electrical engineers (UK, 1974).

Filtering examples prior the analysis seems to be a less studied area in the machine learning [5]. Majority of the machine learning methods deals with irrelevant examples within the algorithm itself (embedded approach) or apply some suitable method as a subroutine during the learning (wrapper approach) [5]. We research machine learning methods, such as genetic algorithms and decision trees, in the context of modelling the medical data and discovering knowledge from it. These methods seem to be quite robust and, therefore, they perform well with the data containing missing values, noise and outliers. For example, the female urinary incontinence and the vertigo data sets, which we have studied with different methods [6,7,8,9], are messy real-world data.

Identification of outliers has recently begun to interest us for two reasons. Firstly, we consider balancing the imbalanced class distribution [e.g. 10] by reducing the largest classes before analysis. Outliers of the major classes seem to be worthwhile candidates for removal. In this line of work, the outliers are treated as poor data which may be removed without further analysis. Secondly, during the enlargement of the vertigo data, we noticed that the outliers may give us some additional insight of data. Outliers are not outright dropped from data, instead they are presented to expert physicians for further consideration.
In this paper, we identify both univariate and multivariate outliers with box plots [11] which are an informal method for outlier detection. The functioning of the informal method was tested in this work with two medical data sets. The results were evaluated objectively by performing discriminant analysis and nearest neighbour classification for the reduced data. Subjective evaluation was done by the expert physicians who studied the outliers manually.

2. Methods

Test of discordancy, formal or informal, is needed to declare extreme values as outliers. Formal testing requires a test statistic, which usually assumes some well-behaving distribution, on basis of which the extremes are possibly declared outliers. Most of the test statistics, for example many Dixon-type tests, are designed to identify a single univariate outlier or an outlier pair using a normal distribution [1]. Unfortunately, the medical data sets we study are problematic in the statistical point of view. The data sets may be mixed, i.e. they contain both quantitative and qualitative variables, and the distribution of the continuous variables is frequently skewed or non-normal. Application of various test statistics would require identification of the distributions, transformations and possibly estimation of distribution parameters. For large data sets this process would be very difficult and tedious. Considering the practical aims of our research, we decided to test discordancy informally using regular box plots [11].

2.1 Box plot outlier identification

The box plot is a well-known simple display of the five-number summary (lower extreme, lower quartile, median, upper quartile, upper extreme) [11]. Box plots are most suitable for exploring both symmetric and skewed quantitative data, but they can also identify infrequent values from categorical data. Unlike in the quick box plot, the extremes of the box plots are not the smallest and largest data values, but the most extreme data values that are not extreme enough to be considered outliers [11]. Figure 3 shows a box plot for the salaries of the electrical engineers discussed earlier.

Figure 2: A histogram for annual salary.  
Figure 3: A box plot for annual salary.
The thresholds for lower and upper outliers are defined as follows: lower threshold = lower quartile - step and upper threshold = upper quartile + step. Step is 1.5 times the interquartile range (upper quartile - lower quartile) which contains 50% of the data. Value \( x \) is a lower outlier, if \( x < \) lower threshold and an upper outlier, if \( x > \) upper threshold. Box plot identifies engineer M's salary as an upper univariate outlier (see Figure 3).

2.2 Multivariate outliers

Methods for identifying univariate outliers are based on unarguable order of data values. For example, in the box plot method salaries are sorted in ascending order and, on the basis of the order, extremes, quartiles and outliers can be found. There is no unambiguous total ordering for \( N \) multivariate observations, but different sub-orderings have been suggested [1,12], of which the reduced sub-ordering is the most often used in the outlier study [1].

Reduced sub-ordering is established in two phases [1,12]. Firstly, a set of scalars \( R = \{ r_i \} \) \((i=1,...,N)\) is produced by transforming each multivariate observation \( \mathbf{x}_i \) into a scalar \( r_i \). Then, \( R \) is sorted to produce the actual ordering of the multivariate data. The transformation is often done with a distance metric [12] and, therefore, the extremes are those multivariate observations associated with the largest values in \( R \).

We used in this study sub-ordering based on Mahalanobis distance which is of the generalised distance metric type [1,4,12]

\[
r^2 = (\mathbf{x}_i-\alpha)\Gamma^{-1}(\mathbf{x}_i-\alpha),
\]

where \( \alpha \) and \( \Gamma \) are problem specific. Mahalanobis distance is obtained by selecting \( \alpha \) and \( \Gamma \) to be population mean \( \mu \) and covariance \( \Sigma \). Usually the population values are unknown and they are estimated with sample mean vector \( \mathbf{m} \) and sample covariance matrix \( \mathbf{S} \)

\[
r^2 = (\mathbf{x}_i-\mathbf{m})\mathbf{S}^{-1}(\mathbf{x}_i-\mathbf{m}).
\]

Gamma-type probability plots are recommended to be used with generalised distances in outlier detection [1 pp. 274-275,4]. However, we applied box plots, as in univariate outlier identification, because we can not assume that multivariate observations come from a normal distribution.

3. Materials and experimental setup

Outliers were searched from two medical data sets. The female urinary incontinence data (see Table 1) was collected retrospectively in the Department of Obstetrics and Gynaecology of Kuopio University Hospital, Finland [6]. The examples are described with 16 variables of which 7 are binary and 9 quantitative. Two variables (uroflowmetry and cystoscopy) were dropped from the analysis, because they had extremely high missing value rates. The vertigo data (see Table 2) was collected in the vestibular unit of the Helsinki University Central Hospital, Finland. The patients, referred to the vestibular laboratory, filled out a questionnaire concerning their symptoms, earlier diseases, accidents, use of medicine, tobacco and alcohol [9]. The information was stored in the
patient database of the expert system ONE [9]. The diagnoses were confirmed by an experienced specialist in the field of otoneurology. In this study, we focused on the six largest patient groups with vertigo and used the 38 most important variables of all the 170 available variables [7,8]. The subset of variables consisted of 16 quantitative variables, 10 ordinal variables and 12 nominal variables, 11 of which were dichotomous. The missing values were replaced in both data sets with means within diagnostic classes. Means were rounded for discrete variables.

Multivariate and univariate outliers were identified separately with box plots by each the diagnostic class, as usual [3]. Discriminant analysis and nearest neighbour classification ($k=1$) with the Euclidean distance metric were used for the objective evaluation. These methods were selected, because they are classical methods for classification in the areas of statistics and machine learning, respectively. Four versions of both the medical data sets were evaluated: the original data (version 0) and three reduced data sets (versions 1-3). Reduced sets were created by excluding 1) the multivariate outliers, 2) the multivariate outliers and 10% of the non-multivariate outlier examples with the most univariate outliers and 3) a random sample from the original data. The random removal was a baseline method where the number of removed examples was the same as the number of outliers excluded from the data set version 2.

The effect of removing the outliers was measured with the classification accuracy of the discriminant analysis functions and nearest neighbour classifier. Classification accuracy $A$ in per cents is $A = 100\% N_c / N$, where $N_c$ is the number of correctly classified examples and $N$ is the number of all examples in the data set. Since discriminant analysis was conducted with the whole data set, it gave indication how the descriptive analysis was affected. On the other hand, nearest neighbour method could be used to study the predictive analysis, because it was performed with 3x10-fold cross-validation.

4. Results

Tables 1 and 2 show the frequencies of the outliers identified from the female urinary incontinence and vertigo data sets, respectively. The tables also show the sizes of diagnostic groups and, for the comparison with the number of outliers, the absolute frequencies corresponding to the 10% portion of each diagnostic group ($N_{10\%}$). The outlier frequencies behaved as expected. The largest classes had the highest number of outliers and the overall number of outliers was reasonably small in both the data sets.

Table 1: Frequencies of the original female urinary incontinence data and its outliers by the diagnostic classes ($N_M = $ multivariate outliers, $N_U = $ univariate outliers).

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Original data</th>
<th>Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>$N_{10%}$</td>
</tr>
<tr>
<td>Stress</td>
<td>323</td>
<td>32</td>
</tr>
<tr>
<td>Mixed</td>
<td>140</td>
<td>14</td>
</tr>
<tr>
<td>Sensory urge</td>
<td>33</td>
<td>3</td>
</tr>
<tr>
<td>Motor urge</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Normal</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Sum</td>
<td>529</td>
<td>53</td>
</tr>
</tbody>
</table>
Table 2: Frequencies of the original vertigo data and its outliers by the diagnostic classes ($N_M = \text{ multivariate outliers, } N_U = \text{ univariate outliers}$).

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Original data</th>
<th>Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>$N_{10%}$</td>
</tr>
<tr>
<td>Vestibular schwannoma</td>
<td>128</td>
<td>13</td>
</tr>
<tr>
<td>Benign positional vertigo</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>Meniere's disease</td>
<td>243</td>
<td>24</td>
</tr>
<tr>
<td>Sudden deafness</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>Traumatic vertigo</td>
<td>53</td>
<td>5</td>
</tr>
<tr>
<td>Vestibular neuritis</td>
<td>60</td>
<td>6</td>
</tr>
<tr>
<td>Sum</td>
<td>564</td>
<td>56</td>
</tr>
</tbody>
</table>

The predictive accuracies of nearest neighbour method and the descriptive accuracies of the discriminant analysis functions are reported in Table 3. There is a clear improvement in the classification ability of both methods in the female urinary incontinence data. Also, removal of a randomly selected sample produced less accurate results, than excluding the identified outliers. The removal of outliers helped the classification of the vertigo data only slightly.

Table 3: Accuracies (%) of the nearest neighbour method (NN) and discriminant analysis functions (DA) in different types of the female urinary incontinence and vertigo data.

<table>
<thead>
<tr>
<th>Versio of data</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incontinence</td>
</tr>
<tr>
<td></td>
<td>NN</td>
</tr>
<tr>
<td>Original</td>
<td>83.9</td>
</tr>
<tr>
<td>Multivariate outliers removed</td>
<td>85.6</td>
</tr>
<tr>
<td>Multi- and univariate outliers removed</td>
<td>87.2</td>
</tr>
<tr>
<td>Random sample removed</td>
<td>82.1</td>
</tr>
</tbody>
</table>

5. Discussion

Outlier identification was studied with informal box plot method using as the test material two real-world data sets. There were two motivations for the identification. Firstly, outliers can be considered to be suspicious data whose removal before applying inductive machine learning methods is reasonable. Especially, dropping of the outliers of the largest classes balances the class distribution and, consequently, makes the classification of the members of the smaller classes easier [10]. Secondly, the knowledge carried by the outliers may be valuable for the domain experts, who may gain additional insight into the data by examining them.

The suitability of the method for the straightforward data reduction was studied objectively with discriminant analysis and nearest neighbour method which are well-known classification methods in statistics and machine learning. Removal of the identified outliers from the female urinary incontinence data improved clearly the classification ability of discriminant analysis functions and nearest neighbour method.
The descriptive accuracy of the discriminant functions improved from 84.3% to 89.5% when both multivariate outliers and examples with the largest number of univariate outliers (14% of the data) were removed from the original data set. The prediction accuracy of the nearest neighbour classifier raised from 83.9% to 87.2%. However, the improvement was only marginal in the vertigo data set.

The most probable explanation for the differences is the characteristics of the data sets. All the female urinary incontinence data were collected retrospectively from the patient records [6], while the main body of vertigo data was obtained carefully in prospective fashion [9]. In addition, the vertiginous patients used in this study were selected to meet the definitions of the six diagnostic classes [9]. As a result, there were not much improvement left in the classification of this data. The earlier machine learning experiments [7,8] lend additional support for this conclusion. The best results were obtained with the decision trees [8] whose predictive classification accuracy ranged from 94% (Meniere's disease) to 100% (benign positional vertigo and vestibular neuritis).

The identified univariate and multivariate outliers were presented to the expert physicians, who evaluated the suspect data to decide whether it was truly outlying. In their opinion, most of the multivariate outliers where abnormal cases. For example, in one case the post voiding residuals and urgency score were abnormally high for a stress-incontinent woman. However, closer examination revealed that the diagnosis was correct, because she had to drink excessively due bowel problems. The experts sometimes disagreed with the box plot method on univariate outliers. The most frequent reason was the natural variation in diagnostic parameters between patients. The upper or lower outliers were extreme, but yet reasonable values for a parameter in a particular diagnostic class. This happened, especially in heterogeneous diagnostics groups, where also extreme values are natural. Vestibular schwannoma and Meniere's disease are examples of the heterogeneous diagnostic groups. Both diseases worsen during the time and the extreme values come often from the patients who have had these diseases for a long time.

The experimental results suggest that box plots can be used for data reduction, but the benefit obtained of excluding outliers is data set dependent. The method can be used only when the classification task is learnable, i.e. there is enough attributes to establish a reasonably well defined classification. This is a fundamental requirement for the machine learning methods, as well as, the multivariate statistical methods. The subjective evaluation by the experts gave controversial, but sound, results. There were real abnormalities in the multivariate outliers and the disagreement on the univariate outliers resulted from the method, which does not utilize any prior knowledge in the outlier identification.

The major limitation of this work is the use of informal box plot method. Discordancy test statistics identify outliers on the basis of the solid theory. Unfortunately, these methods make many assumptions which should be met in order the tests to be applicable. Therefore, one can also argue that a well-known and widely used informal method may be a more appropriate choice for large practical applications, where these assumption are not usually fulfilled. However, the future work should also address the formal test statistics. It would be also interesting to apply some type of heterogeneous distance function [13] for the multivariate outlier detection where we used Mahalanobis distance which works best with continuous quantitative data.
References