### A DETAIL INVESTIGATION ON THE MEDICAL DATABASES BY IMPLEMENTING VARIOUS METHODS

#### ABSTRACT

Outlier detection is presented in detail in chapter 1. The finding of outliers for high-dimensional datasets is a challenging data mining task. Different perspectives can be used to define the notion of outliers. Hawkins et al., 2002, defines an outlier as "an observation which deviates so much from other observations as to create suspicions that it was generated by a different mechanism". While 'Barnett and Lewis, 1994' define it as "An outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that dataset".

#### METHODS

Previously, the outlier detection issues were resolved by the creation of a data model based on probability and also making use of mathematical methods of applied statistics and probability theory. A probabilistic model can be either a priori given or automatically constructed by the given data. Any data object whether belonging to probabilistic model or any other distribution law is decided from the construction of the probabilistic model and in any case that it doesn't suit that model, it is measured as an outlier. Standard probability distributions and combinations of the same are used to create probabilistic models and at times, they may include unknown parameters projected while data mining. Algorithms for approximating probability distributions by empirical data exist along with a priori given probability distributions.

The condition when it comes to the regression model is much more complicated. Here, certain outlying points will influence the regression much more than certain others. In massive volumes of highly multidimensional data (which makes the task challenging), it is of great importance to detect outliers while data mining. The challenging nature is due to the fact that crucial outliers may hide in one dimensional data vision. This makes the detection of one dimensional outliers based on scanning one field or variable or attribute at a time ineffective. There are three fundamental approaches to the problem of outlier detection:

1. **Type 1** – The outliers are estimated without any previous knowledge of the data. This is essentially a learning approach analogous to unsupervised clustering. The approach processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers.

2. **Type 2** – Represents normality as well as abnormality. This approach requires pre-labelled data, tagged as normal or abnormal.

3. **Type 3** – Models only normality (or in a few cases models, only abnormality). It may be considered semi-supervised as although the normal class is taught, the algorithm learns to recognize abnormality as well.

Outlier detection methods can be divided into uni variate and multi variate methods.

#### DATABASES BY IMPLEMENTING VARIOUS METHODS

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Medical dataset</th>
<th>No. of instances</th>
<th>No. of attributes</th>
<th>No. of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bupa Liver disorders</td>
<td>345</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Parkinson</td>
<td>198</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Statlog Heart</td>
<td>279</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Thyroid</td>
<td>216</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Hibermann</td>
<td>306</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

### KEY WORDS:

- Outlier detection
- Medical databases
- Data mining
- Unsupervised clustering
- Normality
- Abnormality

---

**Submitted**: 24th August, 2019  
**Accepted**: 19th September, 2019  
**Publication**: 15th December, 2019

---

www.worldwidejournals.com
The coefficient of determination $R^2$ is an essential tool to assess the model fit. The regular $R^2$ always increases with increased number of factors while the adjusted $R^2$ considers the model complexity. A good model should maximise the adjusted $R^2$ i.e. a measure of the precision of well predicted future outcomes. Adjusted $R^2$ is an $R^2$ that adjusts for a number of explanatory terms in a model; which increases an improvement in the model occurs due to the introduction of a new term. There exists a chance of the adjusted $R^2$ to be negative and hence would be less than or equal to $R^2$. An F-test is any statistical test in which the test statistic has the F-distribution under the null hypothesis. The F-test in one-way analysis of variance is used to assess whether the expected values of a quantitative variable within several pre-defined groups, differ from each other. The alpha value arising from a test gives the p-value. “Degrees of freedom” is an integer value measuring the extent to which an experimental design imposes constraints upon the pattern of the mean values of data from various meaningful subsets of data. Lower p-value than the substantial level of test signifies the importance of the model. Residual is defined as the error predicted from the difference between the predicted value and the actual value. The kurtosis is observed to be of sub Gaussian type.

**Regression Assessment Parameters Used data set:**

**RESULTS DATA SET SIZE: 768**

**TREE**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Value</th>
<th>21 cases</th>
<th>avg(CLASS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI &lt; 28.8500 then avg(CLASS) = 0.0701</td>
<td>(std-dev = 0.2851)</td>
<td>197 examples (30.54%)</td>
<td></td>
</tr>
<tr>
<td>BMI &gt;= 28.8500</td>
<td>avg(CLASS) = 0.5882 (std-dev = 0.4613)</td>
<td>17 examples (3.31%)</td>
<td></td>
</tr>
<tr>
<td>AGE &lt; 29.5000</td>
<td>AVG &lt; 29.5000 then avg(CLASS) = 0.4737 (std-dev = 0.5016)</td>
<td>114 examples (22.18%)</td>
<td></td>
</tr>
<tr>
<td>PG &gt; 144.5000</td>
<td>(std-dev = 0.2851)</td>
<td>197 examples (30.54%)</td>
<td></td>
</tr>
<tr>
<td>BMI &gt; 28.8500</td>
<td>AVG &lt; 29.5000 then avg(CLASS) = 0.4737 (std-dev = 0.5016)</td>
<td>114 examples (22.18%)</td>
<td></td>
</tr>
</tbody>
</table>

The dataset is split into growing and pruning sets by the regression algorithm. A two-step algorithm was used wherein a maximal tree fitting the possible growing set was built in the first step and nested sub-trees were tested as per the cost complexity principle. The optimal tree was selected on the pruning set and the simplest sub-tree with performance close to the optimal tree was selected on the growing set.

**CONCLUSION**

Routine health check-ups is a common practice among adults in most developed countries. Therefore, lesser expensive precautionary measures in case of detection of any disease in its early stages of development gives a patient a better chance at survival than detection of the same at a later stage. Clinical databases with patient information are essential to medical researchers and doctors. In fact, the study with medical data by using the DM techniques is virtually an unexplored frontier which needs extraordinary attention. It can be suggested that: (i) The anomalous outlier conducts facilitate the survey of valuable information buried in their domain. This in turn assists the decision makers in their functioning. (ii) The present experimental results can be used by the
medical doctors to sensibly predict tools from the vast medical database.

(iii) Some of the most promising areas would be patient symptoms, diagnoses and behaviours as well as the thorough understanding of their complex relationships.

For the experiment to be executed on outlier detection, five medical datasets namely viz., liver (345,7), Parkinson(198,24), Heart(270,14), Thyroid(216,6), Haberman (306,4) as instances and attributes are utilized respectively. From the present statistical analysis, it is found that 78 outliers for liver, 67 for Parkinson, 82 for heart, 110 for Thyroid and 61 for Haberman medical datasets are detected.

REFERENCES