Christian Bayer and Robin Seidel*

Comparing distance measures on assessed medical device incident data using Average Silhouette Width

https://doi.org/10.1515/cdbme-2018-0126

Abstract: Many machine learning algorithms depend on the choice of an appropriate similarity or distance measure. Comparing such measures in different domains and on diversely structured data is common, but often performed in regards of an algorithm to cluster or classify the data. In this study, data assessed by experts is analyzed instead. The data is taken from the database of the Federal Institute for Drugs and Medical Devices (BfArM) and represents free text incident reports. The Average Silhouette Width, a cluster density measure, is used to compare the distance measures’ ability to discriminate the data according to the experts’ assessments. The Euclidean distance and four distance measures derived from the Jaccard similarity, the Simple Matching similarity, the Cosine similarity and the Yule similarity are compared on four subsets of this database. The results show, that a better data preprocessing is necessary, possibly due to boilerplate texts being used to write incident reports. These results will also provide the basis to compare improvements by different methods of data preprocessing in the future.

Keywords: Average Silhouette Width, Machine Learning, Distance Measures, Regulatory Affairs, Text Categorization

1 Introduction

The Federal Institute for Drugs and Medical Devices (BfArM) is the main competent authority for risk assessment according to the German Safety Plan for Medical Devices (MPSV, [2]). The MPSV also states the definition of incidents that have to be reported by manufacturers and professional users of medical devices. When an incident report reaches BfArM, it is assessed and stored within the competent authority’s database on incident reports. The risk, associated with a medical device which was involved in an incident, is scientifically evaluated on the basis of the information provided, mainly as free text, and often at least of one subset of all incidents already stored in the database. Such subsets, varying in size and structure, are created and analyzed on demand.

The number of incident reports reaching BfArM has been rising over the last years [1]. Because of this increase in incident reports, methods for handling large amounts of data and especially such methods, that accelerate processes for identifying and understanding coherences in free text reports, are of scientific interest.

In the recently seceded project "OntoPMS", funded by the German Federal Ministry of Education and Research (BMBF) within the program KMU-Innovativ/IKT (01IS15056G), BfArM has been working with a team of experts, to build an ontological search tool [19]. This study addresses the great potential that lies in machine learning algorithms like clustering and classification algorithms for text analysis ([6, 15]).

A wide variety of similarity and distance measures exist and lots of comparisons have already been made [3, 4, 8, 9, 12, 13, 16, 18]. [16] has recently demonstrated the importance of choosing an appropriate measure.

This new comparison study differs from many other comparisons of similarity and distance measures, because data which has been assessed by experts are used instead of clusters or classes computed by algorithms.

2 Challenges

Four challenges have to be overcome, before machine learning algorithms will be able to provide meaningful outcomes.

The first challenge is to decide for a proper distance measure. As the data is already assessed and therefore clusters or classes are known, distance measures are compared in regards to mimic this pre-classification. This comparison may have to be made for each subset from the database individually or based upon properties like its size.

The second challenge is to find a good way of preprocessing the texts, provided by manufacturers in incident reports.

The third challenge is to deal with a multi-label text categorization problem [15]. A new scheme for assessing device problems, based on Device Problem Scheme of the Food and Drug Administration (FDA) [10], has been established at BfArM in 2017. According to this scheme, an expert assess-
ment of an incident report may comprise more than one device problem.

The fourth challenge is to handle the hierarchical structure of this scheme, which is why classes are not pairwise disjointed [15].

This study addresses the first challenge and provides the basis for the second challenge. To avoid the third and fourth challenge in this early state, every code is generalized to the corresponding top level code of the hierarchy. This might lead to heterogeneous classes, because for example the distinct device problems "Failure To Pace Or Properly Pace" and "Power Source Issue" will both be seen as their common generalization "Electrical Issue". Also, only incident reports assessed with only one device problem after generalization are included in this analysis, thus omitting the multi-label categorization.

3 Methods

3.1 Preprocessing texts

As most common similarity or distance measures are unable to compare texts itself, the data has to be preprocessed. As for any other computation in this study, R (Version 3.3) is used for this task. Specifically, the package tm and its functions new(), removeWords() and stemDoc() are used, to translate the free texts into (typically sparse) term-document-matrices, exclude stopwords and stem the remaining words [5]. Any term-document-matrix created that way consists of a number of variables that might increase but never decrease, when a new observation is included.

The results of this study will make it possible to compare different procedures for further preprocessing like weighting variables or reducing their number.

3.2 Average Silhouette Width

In [13], two cluster validation techniques have been used to compare the performances of distance measures in a clustering task. One of these techniques, the Average Silhouette Width (ASW) [14], is used on data, assessed by experts, in this study. Therefore, this comparison is independent of an algorithm, the Average Silhouette Width of an incident report may comprise more than one device problem.

To define the ASW for distance measures, a definition of the silhouette of a single object is needed. Let \( x \) be the name of the object (incident report) and let \( A \) be its cluster (device problem). Then \( i(x) \) shall be the mean distance between \( x \) and any other object of \( A \). For any other cluster \( C \), the mean distance between \( x \) and any object of \( C \) has to be computed. This distance will be called \( d(x, C) \), where \( d() \) is a distance measure. Let \( e(x) \) be \( \min_C d(x, C) \). If at least one of the values \( i(x) \) and \( e(x) \) is larger than 0, the silhouette \( s \) of object \( x \) is defined as follows:

\[
s(i) = \frac{e(x) - i(x)}{\max\{e(x), i(x)\}}.
\]

Otherwise, if \( i(x) = e(x) = 0 \), then \( s(x) = 0 \). It is simple to see, that the Average Silhouette Width ranges from \(-1\) to \(+1\), where larger numbers indicate stronger structured clusters.

3.3 Distance measures for comparisons

An overview of the distance measures compared in this study is provided in Table 1. All measures except the Euclidean distance are defined as similarities in the R-library proxy. These five measures are capable of comparing two preprocessed incident reports \( x \) and \( y \).

Let \( a \) be the number of words which occur at least once in \( x \) and \( y \) (positive co-occurrence), \( b \) the number of words, that occur at least once in \( x \), but not in \( y \), \( c \) the same as \( b \) with roles of \( x \) and \( y \) exchanged, and \( d \) the number of words within the preprocessed bag of words, that do not occur in either \( x \) and \( y \) (negative co-occurrence). The values \( a, b, c \) and \( d \) will be used to define binary measures, which do not take into account, whether a word is mentioned more than once.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>( d_{Jacc}(x, y) = 1 - \frac{a}{a+b} )</td>
</tr>
<tr>
<td>Simple Matching</td>
<td>( d_{Simple}(x, y) = 1 - \frac{a}{a+b+c+d} )</td>
</tr>
<tr>
<td>Yule</td>
<td>( d_{Yule}(x, y) = 1 - \frac{a}{a+d} )</td>
</tr>
<tr>
<td>Cosine</td>
<td>( d_{Cosine}(x, y) = 1 - \frac{a+b}{\sqrt{</td>
</tr>
<tr>
<td>Euclidean</td>
<td>( d_{Euclidean}(x, y) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - y_i)^2} )</td>
</tr>
</tbody>
</table>

It should be noted, that if and only if negative co-occurrences have influence on a distance of two observations, the distance between these objects depends on other observations in the subset.
The Simple Matching similarity [17], the Yule similarity [20] and therefore their corresponding distances, \( d_{\text{Sim}} \) and \( d_{\text{Yul}} \), are influenced by negative co-occurrences. Therefore, they can be expected to perform worse than the other three distances, at least, when the size of the subset increases.

In detail, \( d_{\text{Jac}} \) as corresponding distance to the Jaccard similarity [7] is expected to outperform \( d_{\text{Sim}} \) at least on larger term-document-matrices. As the \( d_{\text{Eucl}} \) would resemble \( d_{\text{Cos}} \), if every row of the term-document-matrices would be normalized [6], it can be expected, that \( d_{\text{Eucl}} \) will perform worse than \( d_{\text{Cos}} \). Negative co-occurrences influence \( d_{\text{Yul}} \), but as their number is multiplied with the number of positive co-occurrences instead of added to them, it is unclear, how \( d_{\text{Yul}} \) will perform, but it can be expected it to not perform as well as \( d_{\text{Jac}} \) and \( d_{\text{Cos}} \) on large subsets.

During "OntoPMS", a large number of incident reports had been analyzed manually and it became evident, that at least some manufacturers use specific words repeatedly and most likely even boilerplate texts when writing incident reports. It is expected that all five measures will negatively affected by boilerplate texts, which is explicitly not removed via preprocessing to show its influence in future analysis. The question arises, whether combining incident reports from more than one manufacturer in a single subset decreases the measures’ performances.

### 3.4 Subsets for comparison

Four realistic subsets of data, containing only incident reports assessed according to the current scheme, have been created from the database. They differ in number of incident reports included (few, 100 to 150, denoted by "r" or many, over 500, denoted by "R") as well as the number of manufactures responsible for reporting (one, denoted by "m", and at least three, denoted by "M"). See Figure 1 for an overview of the four subsets and how the incident reports are unevenly distributed among two or three classes in absolute and relative sizes.

### 4 Results

The results of this study (Table 2) can be viewed from different perspectives. According to [11], none of the distance measures found a substantial structure in any of the four subsets, documenting the need for better data preprocessing. Heterogeneous classes and only modest data preprocessing not addressing boilerplate texts adequate can be causative for the observed lack of structure. As clusters have not been generated using the same distance measure as for computing the ASW like in [11], it should not be expected, that expert assessed data will reveal a structure as clear as created by a clustering algorithm. In some seldom and extreme cases, even identical incident reports can be assessed differently, when additional information arrived i.e. by phone.

<table>
<thead>
<tr>
<th>Subsets</th>
<th>( rm )</th>
<th>( rM )</th>
<th>( Rm )</th>
<th>( RM )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>-0.07</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Simple Matching</td>
<td>-0.20</td>
<td>-0.15</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Yule</td>
<td>-0.38</td>
<td>-0.53</td>
<td>0.10</td>
<td>0.16</td>
</tr>
<tr>
<td>Cosine</td>
<td>-0.10</td>
<td>-0.01</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>Euclidean</td>
<td>-0.15</td>
<td>-0.08</td>
<td>0.03</td>
<td>0.02</td>
</tr>
</tbody>
</table>

The performances of all five measures increase on the two larger term-document-matrices. Neither \( d_{\text{Sim}} \) nor \( d_{\text{Yul}} \) are as sensitive to larger numbers of negative co-occurrences as foreseen. When it comes to discriminating observations of the larger subsets "Rm" and "RM" by device problem, \( d_{\text{Yul}} \) performs better than the four other distance measures, although performing worse than them on the smaller subsets "rm" and "rM". While \( d_{\text{Yul}} \) seems to be sensitive to the number of incident reports, \( d_{\text{Jac}} \) and \( d_{\text{Cos}} \) seem to be rather robust.

\( d_{\text{Jac}} \), which only outperforms \( d_{\text{Sim}} \) on the smaller subsets "rm" and "rM", is least affected by the different properties
of the four subsets. Also, $d_{\text{Cos}}$ and $d_{\text{Euc}}$ are at most moderately affected. As expected, $d_{\text{Cos}}$ performs better than $d_{\text{Euc}}$ on all four subsets.

Binary measures do not perform better or worse than non-binary measures in general on these four subsets.

No clear trend about the influence of boilerplate texts of different manufacturers can be seen, when looking at the performances of all measures on "rm" compared to "rM", and on "Rm" compared to "RM", respectively.

Although their performances are of low quality so far, $d_{\text{Jacc}}$ and $d_{\text{Cos}}$ seem to be least affected by the number of incident reports and the number of manufactures. $d_{\text{Yul}}$ could proof useful, when large subsets are created.

5 Conclusion and Outlook

In this study, a way of comparing distance measures on preclassified data is presented. The ASW provided useful information about how distance measures are able to discriminate previously assessed data.

The next step to improve the results will be to analyze different ways of data preprocessing to address the problem of boilerplate texts. This study provided results allowing comparing the influences of weighting procedures as well as procedures for reducing dimensions not only among each other, but whether and to what degree they lead to an overall improvement. As the data is assessed by experts, it is possible to compare different data driven weighting procedures with procedures, where terms are weighted by the experts assessing the incident reports.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent is not applicable. Ethical approval: The conducted research is not related to either human or animals use.

References


[2] Verordnung über die Erfassung, Bewertung und Abwehr von Risiken bei Medizinprodukten (Medizinprodukte-


