A Data Clustering Approach for Automated Optical Inspection of Metal Work Pieces

Ruth Tesfaye Zibello, Stephan Trahasch, Tobias Lauer
Department of Electrical Engineering and Information Technology
Offenburg University of Applied Sciences
Offenburg, Germany
e-mail: {ruth.zibello, stephan.trahasch, tobias.lauer}@hs-offenburg.de

***Abstract*—This paper describes the use of the single-linkage hierarchical clustering method in outlier detection for manufactured metal work pieces. The main goal of the study is to group defects that occur 5 mm into a work piece from the edge, i.e., the border of the metal work piece. The goal is to remove defects outside the area of interest as outliers. According to the assumptions made for the performance criteria, the single-linkage method has achieved better results compared to other agglomeration methods.**

Keywords-Hierarchical clustering; Outliers; Single-linkage method.

I. INTRODUCTION

Manufacturing processes of metals that end up in different uses involve cutting and shaping of work pieces. During this process, the machine blades that cut or bend such pieces tend to become dull over time, resulting in certain defects, such as dents, scratches, impressions and the like on the work piece.

This work addresses the problem of grouping defects around the border of a metal work piece from the manufacturing process of car body parts. Hence, the objective is to use cluster-based outlier detection in order to realize the clusters that form around the border.

The outcome could help in deciding whether the work piece can be used as is, needs to be polished (reworked) or must be tossed. Furthermore, it can help determine at which point in time the cutting or bending blade requires sharpening or replacement; which falls in the category of *predictive maintenance*.

Cluster analysis, a subfield of unsupervised learning, is used to determine homogeneous subgroups within a larger group of observations. Hierarchical clustering is the approach used to obtain clusters of defects. Variant linkage metrics were sought out during the work. The single-linkage method has turned out to yield the best results.

The remainder of this paper is structured as follows: Section II describes related work and general background. Section III describes the approach and methodology followed during the study. Section IV describes and discusses the results. Finally, in Section V, conclusions and future outlooks are discussed.

II. BACKGROUND

Outliers are observations that deviate from the remainder set of data. Outliers and their detection have been studied in different domains for a variety of applications.

A. Related Work

Statistical methods, supervised and unsupervised algorithms are found in literature to conduct outlier detection. These algorithms are further subdivided into *z*-score, classification-based, cluster-based, distance-based etc. to be implemented on univariate or multivariate outlier detection problems.

In [1], distance-based and cluster-based outlier detection algorithms were proposed. The goal was to improve the quality of data preprocessing and capture the underlying patterns using an outlier score for outlier reduction. Distance-based approaches fetch the top *r* % (percentage recall) of the data based on (dis)similarity measures. While cluster-based approach considers clusters with minimum number of objects as outliers.

The *k*-means algorithm was used for clustering data and Euclidean distance of each object from its corresponding cluster centroid was recorded. Recorded objects were sorted according to their score and those falling below a certain score were eliminated. This work concludes that cluster-based outlier detection outweighs distance-based. It was conducted on three R built-in health care datasets.

In [2], outlier detection based on hierarchical clustering method was conducted to detect erroneous foreign trade transactions in data collected by the Portuguese Institute of Statistics (INE).This work involved statistical performance evaluation according to the criteria specified by the domain experts. Variants of linkage methods presented similar results, but the distance function had major impacts in fulfilling the criteria. The Canberra distance function with a threshold of 5 resulted in performance evaluations of less than 50% of transactions containing at least 90-99 % of the errors, which was better than the desirable target.

This paper reports on an experiment on synthetically generated data that resembles manufactured metal work pieces with defects, using cluster-based outlier detection with hierarchical clustering.

 ISBN: 978-1-61208-631-6

B. *Hierarchical Agglomerative Clustering*

Hierarchical clustering is one type of method that creates a sequence of nested partitions, i.e., a hierarchy of homogeneous groups (clusters). The clusters are visualized in a tree-like structure named dendrogram [3] [4].

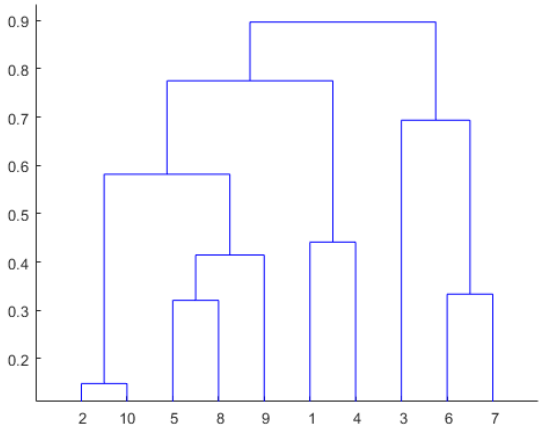

Figure 1. Example of a dendrogram.

The hierarchy ranges from the lowest level (leaf), i.e., each observation in its own cluster, to the highest level (root) consisting of all observations in one cluster. There are two approaches of applying hierarchical clustering: agglomerative and divisive.

Agglomerative clustering works in a bottom-up manner where each object is initially considered as a single-element cluster. Similar pairs of clusters are merged repeatedly until all points are grouped into one root cluster.

The counterpart approach is divisive clustering. It uses a top-down approach; starting from a root cluster and recursively splitting a heterogeneous cluster until each observation is in its own cluster.

Merging and splitting of clusters is performed based on the (dis)similarity measures. The default measure is the Euclidean distance between two observations, whereas the measure between each cluster of observations requires cluster agglomeration (linkage) methods [4].

- Complete: uses the maximum (largest) value of the dissimilarities to link clusters
- Single: opposite to complete, smallest (minimum) value is considered
- Average: as the name describes, it takes the average value of the distance
- Centroid: computes dissimilarity between the centroids of the clusters
- Ward's minimum variance: minimizes the total within cluster variance.

Figure 1 illustrates an example of a dendrogram, where the x-axis shows the observations of the data and the y-axis represents the cophenetic distance (distance between merging/splitting clusters).

Hierarchical clustering method would be more stable approach rather than partitioning clustering techniques because it is not dependent on the initialization of the clusters. The commonly used agglomeration methods are complete, average and ward's minimum variance, these tend to produce balanced trees, whereas, single and centroid tend to produce unbalanced and inversions of clusters respectively.

III. APPROACH

Data used in this work are synthetically generated images resembling a cut part of a car panel (metal work piece) with size 800 mm × 100 mm containing random defects. For this analysis, 25,000 datasets have been generated.

Records in the dataset represent defects that occur during the cutting of the work piece in the production line. Each dataset contains 5 variables and observations in thousands or minimum of hundreds.

Each dataset has the following variables and they are described as follows:

- ID: represents the *i*-th work piece image (0-24999)
- X: horizontal axis of the work piece in mm (0-799)
- Y: Vertical axis of the work piece in mm (0-99)
- D: depth of the defects in µm (micrometers)
- C: category (types) of defects as numbers (1 = dent, 2 = scratch, 3 = pinhole)

Table I shows some rows of data for a randomly chosen workpiece as an example. The variables of interest for clustering are the locations (X and Y) and depth of the defect (D). Data preparation was done by scaling the variables of interest, as they were measured in different units. The variables ID and C were removed as they had no influence in the formation of clusters.

In our approach, agglomerative clustering applying the single linkage method based on Euclidean distance was used to conduct the formation of clusters.

Clusters are identified either by cutting the hierarchy of the resulting dendrogram at a certain height or specified by a domain expert with a predetermined number.

Since each dataset had different records of defects, for the present work the goal was to make the resulting clusters be dependent on the number of observations (*n*).

The following formula obtained from [2] has been used for the number of clusters.

$$n_c = \max\left(2, \frac{n}{10}\right) \quad (1)$$

It influences the formation of clusters to be dependent on the number of observations within each dataset.

TABLE I: DATA SAMPLE FOR WORKPIECE 13800

ID	X	Y	D	C
13800	0	0	15	3
13800	1	0	15	1
13800	7	35	33	2
13800	40	88	9	3

 ISBN: 978-1-61208-631-6

As performance criteria, assumptions of the production line were made. The first 2000 work pieces will not have any defects as the blade would be new and sharp. Therefore, the first 2000 generated records of data shall just have some generated noises which represent defects with minimum depth. These defects are of no influence in creating a cluster.

In order to satisfy the assumption taken and also realize which synthetic data gave reasonable results, clustering tendency was assessed. Hopkins statistics, a statistical clustering tendency method, was used for assessing whether the data contained inherent grouping structure or random noises [4].

The result value of a Hopkins statistic is a probability which indicates whether the given data $\mathcal{D}$) has non-random or uniformly distributed structure. The following formula shows how clustering tendency using this statistical method is obtained.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \quad (2)$$

Hopkins statistical probability (H) is the mean of the nearest neighbor distance in a simulated dataset (random $\mathcal{D}$) divided by the sum of the mean nearest neighbor distances in the real ($\mathcal{D}$) and across the simulated dataset. If the value of H > 0.5, then it is concluded that the dataset $\mathcal{D}$ has meaningful clusters [4].

As per the assumptions made for the performance criteria the first 2000 datasets had H < 0.5 and did not contain inherent groups, therefore, no clustering technique was applied to these datasets.

The remaining datasets had H > 0.5. Hence, the next step was to apply hierarchical clustering using the single linkage method based on Euclidean distance and remove outliers.

The approach used in [2] has been adapted to cluster and remove defects outside the border as outliers. The key idea was to use the size of the resulting clusters as indicators of the presence of outliers. In the case of this work, outliers would be those clusters with a number of elements less than some threshold τ.

The threshold used is a fixed number which can be replaced based on the assumptions or a domain expert user sees fit.

This method of outlier detection requires parameters to be specified. The main parameters are the number of clusters n_c and threshold τ. Table II shows the algorithm for outlier detection adapted from [1].

Once the final cluster(s) is/are obtained the further analysis is done in order to determine which cluster(s) has/have defects in the pre-classified range.

IV. RESULTS

The results shown in this section are for the example dataset described in Section 0 (TABLE).

Dataset 13800 contains data about defects present in the 13800th metal work piece manufactured. This dataset contains 1351 observations and 5 variables. The initial clustering determined by (2) ends up with 135 clusters.

TABLE II: ADAPTED ALGORITHM FOR OUTLIER DETECTION USING HIERARCHICAL CLUSTERING [1]

Input :

- Dataset (D) with *n* observations and *k* variables
- Standardize variables of interest (X,Y and D)
- Compute H (Hopkins Statistic)
- d: Euclidean distance function
- h: Hierarchical clustering (single-linkage method)
- $n_c = \max\left(2, \frac{n}{10}\right)$
- τ : Threshold = 50 (10 , 20 ….)

Output :

Out: set of outlier observations
If H < 0.5:
    Out → ϕ
If H > 0.5:
    Obtain d from scaled data
    Use algorithm h to grow hierarchy from d
    Group initial clusters with n_c
    For each resulting cluster c Do:
        IF sizeof(c) < τ THEN
        Out → Out U {obs ∈ c}

Figures 2 and 3 show the original dataset and the initial cluster for this specific dataset.

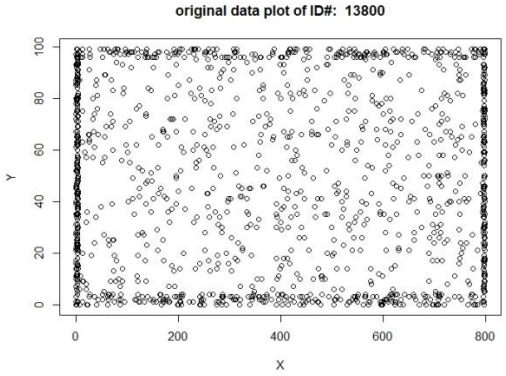

Figure 2. Original data for 13800 work piece.

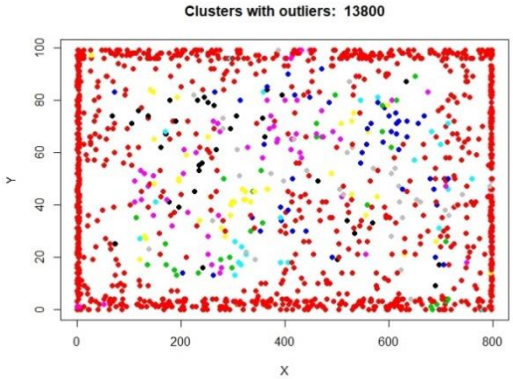

Figure 3. Initial n_c = 135 clusters for 13800 workpiece.

The threshold parameter for this work has been fixed to 50 indicating the minimum number of elements that a cluster should hold. The clusters with number of elements less than the threshold are considered as outliers and removed. The sample dataset ends up having one final cluster, as illustrated in Figure 4.

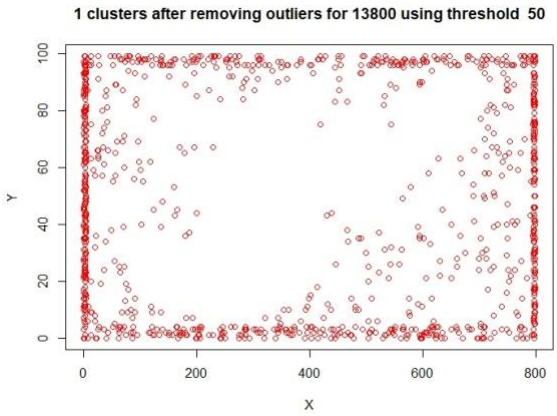

Figure 4. Final cluster for 13800 using single linkage.

Variants of agglomeration techniques were sought before deciding on the single linkage method. The results for the other agglomeration techniques did not fulfill our performance criteria, which aimed to find clusters around the border of the work piece. The other techniques either did not remove clusters, or removed clusters from around the border as a whole. The threshold used for the other agglomeration techniques is less than 50 because they produce balanced trees. The number of elements in all clusters was relatively similar. In contrast, the single linkage produced trees which are unbalanced, i.e., few clusters contained a large number of elements, whereas the rest had a small number of elements and were eliminated as outliers. Figure 5 shows the clusters resulting from different agglomeration techniques with different thresholds for sample work piece 13800.

After having the final clusters, the depth variable is classified into three ranges; 0-9 µm, 10-19 µm and ≥ 20 µm. Figure 6 shows a histogram that illustrates the number of observations in each range within a cluster. This can be helpful to determine whether or not the metal work piece could be of use.

Based on the result depicted above, it could be concluded that the 13800 metal workpiece is of no use as most of the defects have ≥ 20 µm.

It is also of particular interest to follow a series of consecutive work pieces in the production process, in order to see trends and possibly predict – and hence, avoid – machine failures. Figure 7 illustrates how the number of elements within each range of the final cluster(s) changes through time.

This information indicates the growth of defects with high depths in the production process through time which in turn could be used as an indicator when to replace or sharpen the blade in the cutting device. It can be concluded that after manufacturing about 10,000 work pieces, the blade requires sharpening or replacement.

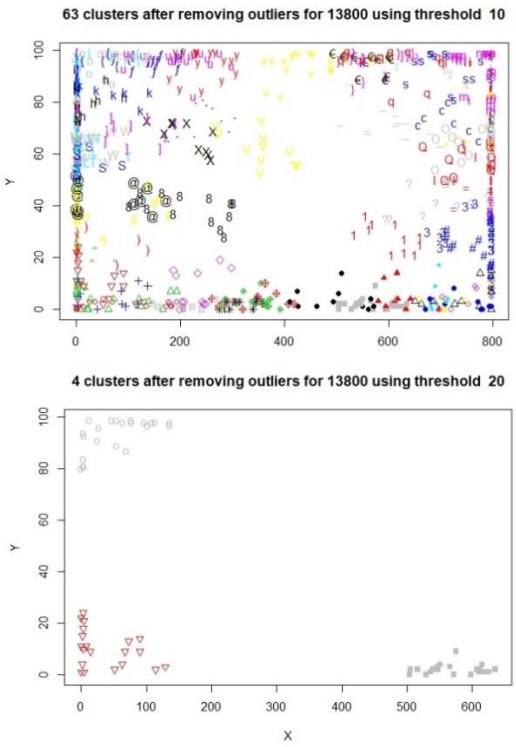

Figure 5. Sample results for 13800 workpiece with complete linkage method.

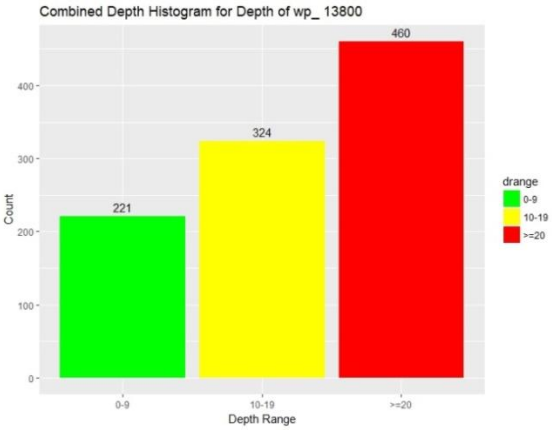

Figure 6. Combined depth histogram of all clusters in 13800.

Figure 7: Depth range over time of each cluster for all datasets.

V. CONCLUSION

In this work, hierarchical clustering using the single linkage method has been used to determine clusters of defects around the border of metal work pieces. Clusters with a number of elements less than a fixed threshold are removed as outliers.

Single linkage agglomeration makes up for an ideal choice to be used in outlier identification in comparison to other agglomeration techniques because it tends to produce unbalanced trees where observations are infused one at a time.

Prior to determining clusters within each dataset, the clustering tendency of each dataset is determined.

The results of this work indicate whether the manufactured workpiece could be of use, requires some polishing or is of no use at all. The indicator is for the sharpness of the blade that cuts or bends the workpieces, in order to have less or insignificant defects.

Future outlooks for this work could be to use the obtained results in a classification problem where all datasets contain labels of good, ok and bad work pieces. Based on their labels, work pieces could be intelligently classified.

ACKNOWLEDGMENT

The authors wish to thank Peter Strohm of Jedox AG for his support and valuable input.

REFERENCES

[1] A. Christy, G. Meera Gandhi, and S. Vaithyasubramanian, “Cluster Based Outlier Detection Algorithm For Healthcare Data,” *Procedia Computer Science*, no. 50, pp. 209–215, 2015.

[2] A. Loureiro, L. Torgo, and C. Soares, “Outlier Detection Using Clustering Methods: a data cleaning application,” Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector, 2004.

[3] M. J. Zaki and W. J. Meira, Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014.

[4] A. Kassambra, “STHDA: Statistical Tools For High-Throughput Data Analysis,” [Online]. Available from: http://www.sthda.com/english/wiki/print.php?id=234. [retrieved: March 2018]

 ISBN: 978-1-61208-631-6