INSTITUT FÜR INFORMATIK

Evaluation of Impact of Data Quality on Clustering with Syntactic Cluster Validity Methods

Elena Sivogolovko

Bericht Nr. 1107
August 2011
ISSN 2192-6247

CHRISTIAN-ALBRECHTS-UNIVERSITÄT

ZU KIEL

Institut für Informatik der
Christian-Albrechts-Universität zu Kiel
Olshausenstr. 40
D – 24098 Kiel

Evaluation of Impact of Data Quality on Clustering with Syntactic Cluster Validity Methods

Elena Sivogolovko

Bericht Nr. 1107
August 2011
ISSN 2192-6247

e-mail: efecca@gmail.com

Abstract

Relationship between Clustering and Data Quality is not completely studied. It is usually assumed that input dataset does not contain any errors or contains some "noise", and this concept of "noise" is not related to any Data Quality concept. In this research we focus on the four most commonly used data quality dimensions, namely accuracy, completeness, consistency and timeliness. Using definitions and constructs of these data quality dimensions, we evaluate the impact of data quality on clustering outcomes. Four different clustering algorithms and five real datasets were selected to show the interaction between data quality and cluster validity.

1 Introduction

The purpose of this research is to evaluate the impact of data quality on the outcomes of clustering. To do so, metrics for data quality and different data quality levels are defined based on prior work and applied to analyze the effects and interactions of this factor on the outcomes of clustering for a real-world datasets.

The main goal of clustering is to organize a collection of data items into clusters, such that items within a cluster are more "similar" to each other than they are to items in the other clusters. This notion of similarity can be expressed in very different ways, according to the purpose of the study, to domain-specific assumptions and to prior knowledge of the problem. Clustering is considered as unsupervised learning. because it is usually performed when no information is available concerning the membership of data items to predefined classes. Clustering is widely used in computer vision, pattern recognition, information retrieval, economics, bioinformatics and many other areas. A lot of different methods were developed in different communities for different clustering tasks.

The procedure of evaluating clustering results is called cluster validity. Three different approach are defined for cluster quality estimation: external, internal and relative. First of them uses some predefined knowledge, such as class labels, for structure quality evaluation. Second approach evaluates clustering results in terms or quantities that involve the vectors of dataset themselves. The main idea of third approach is the evaluation of clustering structure by comparing it to other clustering structures,resulting by the same

algorithms but with different input parameters or by the different algorithms. All these approaches uses syntactic quality definition. In internal cluster validity approach good cluster structure should be the same as predefined class structure. In external and relative approaches good cluster structure should have compact and separate clusters. Therefore there is no approach that consider semantic quality of a cluster structure. We can define four different aspect, which can influence the quality of obtained cluster structure quality: 1) input data structure 2) input data quality 3) clustering algorithm and 4) cluster validity measure. In our work we focused mostly on the second factor. While there are no universally agreed-upon definitions of data quality, there is no dispute about the importance it has and that the consequences when it is poor can be large. Data quality is often conceptualized in terms of dimensions, our study we considered four of these dimensions: accuracy, completeness, consistency, and timeliness, which are the most commonly used.

In clustering (and in Data Mining in general) all data quality errors are described by general term "noise". Misspelling attribute values, incomplete data, data in the wrong format, data with broken referral integrity, etc. — all of them are defined as "noise". Therefore the data quality dimensions do not have direct analogs among particular noise types. Data Mining and Data Quality areas are somewhat separated in this position. If data analyst has a dataset and some assumptions about its quality in terms of Data Quality what should he takes into account to choose a clustering algorithm? We suppose that low quality of the input data leads low cluster quality, so we formulate our general hypothesis, which should be use for each dataset and clustering algorithms as

Hypothesis: Each level of data quality has significant influence on the outcomes of clustering.

This can be illustrated by cluster validity indecies. In our research we tried to estimate the correlation between input data quality and clustering quality and give some recommendations on the choice of clustering algorithm.

2 Background

2.1 Data Quality

Data Quality is a complex multidimensional concept. Most commonly it is considered as the measure how well the data fits to their intended use. Different numbers of data quality dimensions were suggested by previous research[4], in this research we considered four frequently used dimensions: accuracy, completeness, consistency and timeliness. Three quality levels are usually defined for each data quality dimension: low, medium and high. In this case the main question is: What does it means "low" or "high" quality for a given dimension? In general, the answer depends on the data area. For example, for social network database completeness or accuracy are not very important: almost all attributes can be missed or misspelled. But for a financial database accuracy and completeness are extremely important: even a small percentage of wrong or missing values is unacceptable. In this work we focused on medium level of data quality importance.

2.1.1 Accuracy

Accuracy has a particularly wide range of definitions with similar semantics. These definitions include the following: "the recorded value is in conformity with the actual value" [5], "agreement with either an attribute of a real-world entity, a value stored in another database, or the results of an arithmetic computation" [6] and "the closeness of the value in our database to the true value" [7]. Accuracy has also been defined to encompass groups of other dimensions such as completeness, consistency, or timeliness. Accuracy is difficult to measure since very often the real-world value is not known [8], [9], but with predefined data, where the "true" value is known, one can use some external metric for accuracy as for example

$$F_A = 1 - \frac{V_T}{N}$$

where V_T is the number of elements in a dataset having incorrect values and N is the total number of elements. According to previous research, a high level of accuracy can be defined as 92-100% of correct values, a medium level as 88-92% and a low level as 80-88% [27],[28], [29].

2.1.2 Completeness

Completeness can be defined in two different ways.

1. as data having all values recorded. [11]
2. as measure of how completely the target domain is represented in our database. [7]

In first case completeness can be measured as the ratio of the number of tuples with null values to the total number of tuples. In second case some special proxies should be used for completeness measurement. Also there is a flip side of completeness – the dataset should not contain entries for things which do not exist. Commonly used completeness metric is

$$F_C = 1 - \frac{M_T}{N}$$

where M_T is the number of elements having null values in fields. Prior research provides wide range of discussed completeness rates: 80-100% [29], 75-95% [30], 40-100% [32]. We used 84-100% complete instances as high completeness level, 67-84% as medium level and 50-67% as low level.

2.1.3 Consistency

As accuracy, consistency has a large variety of definitions, most of them refer to uniformity. Consistency can be considered as "the representation of the data value is the same in all cases" and as "format and definitional uniformity within and across all comparable datasets." [10] Consistency "ensures that the there no conflicts within or between data sets" [7]. Also consistency can be defined with respect to referential integrity [12], [13], [14]. The simplest consistency metric can be defined as follows

$$F_{Con} = 1 - \frac{R_T}{N}$$

where R_T is the number of tuples with violations of referential integrity. Previous research does not clarify what "high" or "low" consistency is. Blake at al. used the same values as for accuracy in order to represent consistency levels [29]. We suggest that in our case it is more appropriate to use the same values as for completeness.

2.1.4 Timeliness

Timeliness is extremely different from other data quality dimensions. It is critical for attributes that change over time, while it is irrelevant for attributes that are fixed and do not change. Two different notions are usually used for timeliness definition: volatility and currency. Volatility refers to the time period between real-world change and next change which makes original data invalid [15], [16]. Currency "refers to the age of data units used to produce the information products" [9]. The recommended timeliness metric is

$$F_T = \max(0, 1 - \frac{Currency}{Volatility})$$

As for consistency, there is only one prior work, where values for different timeliness levels was suggested [29]. We used the following ranges: 80-100% actual instances for high level, 60-80% for medium level and 40 - 60% for low level.

2.2 Cluster validity

In our work we used external cluster validity metrics for clustering quality estimation. In external approach results of clustering algorithm are evaluated according to some predefined knowledge about the cluster structure in the data set. In our case we used labeled real datasets for experiments, so we could use external approach as well. The following four validity indecies were used for experiments evaluation: Rand statistic, Jaccard coefficient, Folkes and Mallows index and F-measure. For the first three indicies the following concepts should be considered with respect to every pair of elements (x_j, x_i) in the dataset :

1. SS: if both elements belong to the same cluster and to the same class in predefined dataset structure.
2. SD: if elements belong to the same cluster and to different classes
3. DS: if elements belong to different classes and to the same class.
4. DD: if both points belong to different clusters and to different groups.

We defined SS,SD,DS and DD as the numbers of SS, SD, DS and DD pairs respectively and $M = \frac{N(N-1)}{2}$ is the number of all pairs in the data set.

Validity indicies formulas are

$$Rand = \frac{SS + DD}{M}$$

$$Jaccard = \frac{SS}{SS + SD + DS}$$

$$FM = \sqrt{\frac{SS}{SS + SD} * \frac{SS}{SS + DS}}$$

High values of these indicies indicate the high similarity between obtained cluster structure and predefined classes.

There are several approaches for generalizing F-measure to the clustering case. We used the following one: for cluster c_i and class g_j let's consider $Precision(i, j) = \frac{n_{ij}}{n_i}$ and $Recall(i, j) = \frac{n_{ij}}{n_j}$ where n_{ij} is the number of objects of cluster c_i which belong to class g_j, n_i is the number of objects in c_i, and n_j is the number of objects in g_j. The F-measure of c_i and g_j is defined as

$$F1(i, j) = \frac{2 * Precision(i, j) * Recall(i, j)}{Precision(i, j) + Recall(i, j)}$$

Overall F-measure value is computed as the weighted average F-measures for each class

$$F1 = \sum_j \frac{n_j}{N} \max_i F1(i, j)$$

The better the clustering quality, the higher the F-measure.

3 Data Quality Model

For our experiments we generated 100 different variants of each dataset for each data quality level. For data quality level modeling we first calculated

$$M = M(dataset, quality\_dimension, level)$$

the number of elements, which "should be" wrong in given dataset. Percentage of correct and incorrect data for each data quality dimension and quality level are defined in 2.1. After that we constructed the wrong elements subset W by randomly choosing M elements from original dataset.

3.1 Accuracy modeling

According to 2.1.1, we considered the number of elements with correct values as the measure of accuracy. To model different accuracy levels, for each element $x \in W$ we changed random number of element attributes $0 \leq d' \leq d$ to random values from the range $\left[\overline{x[i]} - 3\sigma(i), \overline{x[i]} + 3\sigma(i)\right]$, where $\overline{x[i]}$ is the grand mean of i-th attribute and $\sigma(i)$ is the standard deviation of i-th attribute correspondingly. This range was constructed according to Chebyshev's inequality: $P(|X - \overline{x}| \geq k\sigma) \leq \frac{1}{k^2}$ with assumption that all attributes are independent random variables. So probability that attributes value do not lay in the constructed range is less than 11.2%.

3.2 Completeness modeling

We considered completeness as data having all values recorded 2.1.2 and to model this data quality dimension for each element $x \in W$ we replaced random number of its attributes with $NULL$ value. It should be noted, that Weka [] clustering implementation replaced all such values with grand mean of corresponding attribute.

3.3 Consistency modeling

As it is shown in 2.1.3 consistency is often defined with reference to uniformity or referential integrity. In general, in dataset with some relationships, different consistency levels can be simulated by random switching of references between dataset elements, but in our case all datasets do not have any relationships, therefore we used the following consistency modeling method: for each $x \in W, x \in C_i$ we replaced values of random number of attributes with values of the same attributes of element from some other class: $x[i] = y[i], y \in C_j, j \neq i$.

3.4 Timeliness modeling

The following notions, volatility and currency, are used to define timeliness 2.1.4. If you do not have different snapshots of the same dataset, you can not create the different levels of timeliness as it is: with different volatility and currency. If only one dataset snapshot is given, it means that currency is fixed, but volatility can be simulated by generating "new" data. To model

timeliness for each $x \in W$ we replaced all element attributes values with new ones from range $\left[\overline{x[i]^c} - 3\sigma(i)^c, \overline{x[i]^c} + 3\sigma(i)^c\right]$, where $\overline{x[i]^c}$ and $\sigma(i)^c$ is mean and standard deviation of i-th attribute in class c correspondingly.

4 Experimental design

4.1 Datasets

In our work we used five real datasets from UCI repository. General information about size, dimensions and number of classes in each dataset is listed in table 1. **Cardiotocography** dataset consists of 2126 fetal cardiotocograms

Table 1: Datasets information

Name	Instances	Attributes	Classes
Cardiotocography	2126	34	3 or 10
Image Segmentation	2310	19	7
Page Blocks Classification	5473	10	5
Pen-Based Rec. of Handw. Digits	10992	16	10
Wall-Following Robot Navigation	5456	24	4

(CTGs), which were automatically processed and the respective diagnostic features er measured. The CTGs were also classified by three expert obstetricians and a consensus classification label assigned to each of them. The dataset can be used either for 10-class or 3-class experiments and in our work we used 10 as the correct number of clusters.

Image segmentation dataset consists of 7 classes. Each class contains 330 elements. The instances were selected randomly from a database of 7 outdoor images. The images were hand-segmented to create a classification for every pixel. Classes labels are brickface, sky, foliage, cement, window, path, grass.

Page Blocks Classification dataset contains information about blocks of the page layout of a document that has been detected by a segmentation process. This is an essential step in document analysis in order to separate text from graphic areas. Indeed, the five dataset classes are: text, horizontal line, picture, vertical line and graphics. In contrast to previous two datasets page blocks data is biased. Text class contains 4913 elements that is 89.8% of all dataset. Other classes contain 329, 28, 88, 115 elements respectively.

Pen-Based Recognition of Handwritten Digits dataset consists of 10 classes. Each class contains ≈ 1100 elements. It was created by collecting 250 handwriting digit samples from 44 writers. These writers was asked to write 250 digits in random order inside boxes of 500 by 500 tablet pixel resolution. The spatial re-sampled digits was represented as a sequence of T points $(x_t, y_t)_{t=1}^{T}$, regularly spaced in arc length, as opposed to the input sequence, which is regularly spaced in time. Dataset authors said that $T = 8$ gave the best trade-off between accuracy and complexity.

Wall-Following Robot Navigation dataset were collected as the SCITOS G5 robot navigates through the room following the wall in a clockwise direction, for 4 rounds, using 24 ultrasound sensors arranged circularly around its "waist". The classes are: Move-Forward – 2205 elements (40.41%), Slight-Right-Turn – 826 elements (15.13%), Sharp-Right-Turn – 2097 elements (38.43%), Slight-Left-Turn – 328 elements (6.01%).

4.2 Clustering algorithms

For our experiments we chose Weka implementation of following algorithms.

1. K-Means [18], [19] is one of the most known clustering algorithms. It is successfully used in many scientific and industrial applications. It starts with random initial partition and keeps reassigning the dataset elements to clusters based on the similarity between the element and the cluster centroid. It stops when stable cluster structure is found or some condition (for example, maximum number of iteration) is reached. In Weka library, the following parameters are required for K-Means run: 1) suggested number of clusters 2) maximum number of iterations.

2. Farthest First [22], [23]. The fast simple approximate clustering algorithm. It constructs the centroid set according to farthest-first approach and after that assigns elements to the nearest centroid. The only parameter the algorithm requires is the number of clusters.

3. DBScan [24] requires two parameters: neighborhood radius — ϵ and the minimum number of points required to form a cluster – $minPts$. Its main notions can be listed as follows 1) An ϵ -neighborhood of element x is $N_\epsilon(x) = \{y \in X | d(x, y) \leq \epsilon\}$ 2) x is a core object if its

ϵ -neighborhood contains more that $minPts$ points. 3) y is density-reachable from a core object x if a finite sequence of core objects between x and y exists such that each next belongs to ϵ-neighborhood of its predecessor. 4) y is density-connective to x if both of them are density-reachable from a common core object. DBScan cluster is a density-connective component, which grows in any direction that density spreads.

4. XMeans [25] can be considered as K-Means extension which does not require the pre-defined number of clusters. X-Means starts with K-Means partition to user-defined minimum number of clusters and after that algorithm tries to split each obtained cluster into two parts. It measures special index for the whole cluster and for its parts and makes a decision about splitting in order to improve this index. The algorithm stops when either maximum number of clusters or the maximum number of iterations is reached. X-means requires the following parameters: 1)minimum number of clusters 2)maximum number of clusters 3)maximum number of iterations 4) maximum number of K-Means iterations (all splitting operations are performed by K-Means).

5 Experiments

5.1 Clustering parameters tuning

We used predefined number of clusters where it was required. We tuned other important parameters in the naive way: we run algorithms with different parameter values, measured the validity indecies and chose the best parameters set for each algorithm. The OPTICS algorithm was used for DBScan $minPts$ value estimation. Final parameters are listed in table 2. As we have mentioned above, we used weka implementation for all algorithms. If parameter is not mentioned in the table, that means that we used default Weka value for it. (Except the maximum number of clusters in XMeans algorithm, which was set to 100)

Table 2: Clustering parameters

	K-Means (num. of clst.)	Farthest First (num. of clst.)	DBScan (ϵ, minPts)	XMeans (num. of iter.)
Cardiotocography	10	10	0.72688, 6	2
Image Segmentation	7	7	0.16735, 6	3
Page Blocks	5	5	0.18681, 6	2
Handw. Digits	10	10	0.35917, 6	2
Wall-Following	4	4	0.78, 6	1

5.2 "Ideal" case

In first stage of our experiments we considered all our datasets as datasets without any errors. We run clustering on them (with parameters defined in previous section) and measured the clustering quality. We performed 100 runs with different random seed for algorithms which use random initialization (namely KMeans, XMeans and Farthest First) and single run for DBScan. As it shown in table 3 two datasets - Cardiotocography and Page Blocks - have good cluster structure, Image Segmentation and Pen Digits can be clustered with difficulties and Wall Following data has bad cluster structure or the algorithms we used can not find it.

5.3 Results

After that we performed clustering on datasets with errors (they were generated according to the model described in 3) and calculated cluster validity values for each obtained cluster structure. An Unpaired Wilcoxon test was used to evaluate the influence of data quality on each clustering algorithm outcome. We considered each data quality dimension independently. We compared the cluster validity of structure obtained from the data with high quality level with the validity of structures obtained from medium quality level and low quality level correspondingly. We chose Wilcoxon test because we did not know real quality values distribution, therefore we could not use ANOVA or t-test, which require Normal date distribution. We also used the difference between samples means as the measure of quality changes, in order to distinguish the case when changes are statistically significant and have great absolute value and the case when changes are statistically significant but their absolute value is small. We consider changes as small if the

Table 3: "ideal" results

		Rand	Jaccard	Folkes-Mallows	F1
Cardiotocography	K-Means	0.937	0.651	0.786	0.805
	Farthest First	0.901	0.522	0.685	0.755
	DBScan	0.997	0.977	0.989	0.988
	XMeans	0.923	0.612	0.757	0.758
Image Segmentation	K-Means	0.862	0.388	0.561	0.643
	Farthest First	0.705	0.274	0.470	0.538
	DBScan	0.81	0.288	0.454	0.581
	XMeans	0.869	0.331	0.500	0.600
Page Blocks	K-Means	0.440	0.331	0.558	0.593
	Farthest First	0.828	0.824	0.908	0.870
	DBScan	0.833	0.828	0.909	0.866
	XMeans	0.356	0.219	0.454	0.473
Pen Digits	K-Means	0.913	0.432	0.605	0.715
	Farthest First	0.834	0.243	0.404	0.449
	DBScan	0.776	0.258	0.465	0.566
	XMeans	0.880	0.361	0.541	0.646
Wall-Following	K-Means	0.600	0.207	0.345	0.423
	Farthest First	0.530	0.260	0.422	0.462
	DBScan	0.45	0.266	0.439	0.445
	XMeans	0.599	0.208	0.347	0.423

difference between sample means is less than 0.04. (In general this approach is uninformative, but we supposed that in most cases it should provide some additional information)

Detailed experiments results are shown in the Appendix. Generalized results are presented in table 4. We used the following generalization method: for each algorithm, data quality dimension and quality level the most frequently obtained statistical value was shown. Our first observation is that data quality influence depends on dataset properties. If input dataset has bad cluster structure, low data quality will have small influence on cluster quality. If dataset has good cluster structure, low data quality will have significant negative effect. Nevertheless we can formulate the following recommendations for clustering algorithms usage in different situations. Accuracy highly depends on data structure. DBScan is usually considered as stable

Table 4: Statictical results. Legend: ■ — significant negative impact, ▼ — significant but small negative impact, □ — unsignificant changes,▲ — significant but small positive impact, ♦ — significant positive impact

Algorithm	Quality Level	Accuracy	Completeness	Consistency	Timeliness
K-Means	Meduim	▼	■	▼	▼
	Low	▼	■	▼	□
FarthestFirst	Meduim	□	■ - ▼	■	▼
	Low	■	▼	■	□
DBScan	Meduim	▼	■	■	▼
	Low	▼	■	■	■
XMeans	Meduim	▼	▼	■	▼ - ▲
	Low	▼	■	■	♦

for accuracy changes, but we should say that in some cases accuracy errors can cause extra-"merge" situation when two classes merge because an error element becomes a "bridge" between them. That's why DBScan has some benefits only if clear outliers are considered as accuracy errors. In most cases difference between DBScan validity on data with high accuracy and on data with low accuracy is not very big, but difference between validities on high and ideal data accuracy sometimes is huge. KMeans and XMeans algorithms with accepted accuracy quality levels in most cases produce small significant error. Completeness depends on data structure too, but for datasets with bad structure FarthestFirst and XMeans can be mentioned as the most stable. Consistency in most cases has the most negative effect. According to results presented in table 4 KMeans seems to work better with inconsistent data, but this idea is supported only by Rand metric, according to other three validity indicies low data consistency also has big and significant negative effect on KMeans results. Therefore we should say that all algorithms work badly with inconsistent data. Partitioning algorithms work better with timeliness changes than DBScan. If significant data changes are expected, the $minPts$ parameter should be re-estimated before DBScan run. In case of timeliness changes the most stable algorithm is XMeans, but we should mention, that in some situations even the worst DBScan partition can be better than XMeans one.

We also should mention that our observations depend on Weka clustering algorithms implementation and our data quality modeling process. That's

why it can be considered as recommendations only.

6 Conclusion

In this research we studied the influence of different data quality dimensions on clustering outcomes. We considered four data quality dimensions: accuracy, completeness, consistency and timeliness. We demonstrated that first three of them have significant negative effect on clustering algorithms results and the last one, timeliness, has significant negative impact if algorithm uses some knowledge which was obtained from "ideal" data and can be easily changed when new elements are added in a dataset(like $minPts$ in DBScan). We constructed the relationship between Data Quality concepts and clustering concepts and made some recommendations on usage of different clustering algorithms with respect to expected data quality level.

We considered data quality dimensions independently from each other, studying the impact of several dimensions simultaneously can be a good point for future work, because in real-word application there are many datasets with different data quality errors at the same time. Also we used only four clustering algorithms and some other widely-used clustering approaches, for example, hierarchical one, were not considered in our work and impact of the data quality to these approaches can be studied in the future. Also there is no research about influence the data quality on fuzzy clustering, which is important in different scientific and industrial areas.

References

[1] R. Blake, P. Mangiameli. The Effects and Interactions of Data Quality and Problem Complexity on Classification. ACM Journal of Data and Information Quality, Vol. 2, No. 2, Article 8, February 2011.

[2] N. Grira, M. Crucianu, N. Boujemaa. Unsupervised and Semi-supervised Clustering: a Brief Survey. In Proc. 'A Review of Machine Learning Techniques for Processing Multimedia Content', Report of the MUSCLE European Network of Excellence, 2005.

[3] M. Halkidi, Y. Batistakis, M. Vazirgiannis. On clustering validation techniques. Intelligent Information Systems Journal 17(2-3), 107-145, 2001.

[4] S. A. Knight, J. Burn. "Developing a framework for assessing information quality on the World Wide Web". Informing Science, Vol. 8, pp. 159-172, 2005.

[5] D.P. Ballou, H.L. Pazer. Modeling data and process quality in multi-input, multi-output information systems. Manag. Sci. 31, pp 150-162, 1985.

[6] B.D. Klein, D.L. Goodhue, G.B. Davis. Can humans detect errors in data? Impact of base rates, incentives, and goals. MIS Quart. 21, pp 169-194, 1997.

[7] T. Dasu, T. Johnson. Exploratory Data Mining and Data Cleaning. Wiley, p. 105, 2003.

[8] G. Shankaranarayanan, Y. Cai. Supporting data quality management in decision-making. Decis. Support Syst. 42, pp. 302-317, 2006.

[9] D.P. Ballou, R.Y. Wang, H. Pazer, G.K. Tayi. Modeling information manufacturing systems to determine information product quality, Management Science, 44 (4), pp. 462-484, 1998.

[10] D.P. Ballou, H.L. Pazer. Modeling completeness versus consistency tradeoffs in information decision contexts. IEEE Trans. Knowl. Data Engin. vol. 15, pp 240-243, 2003.

[11] P. Gomes, J. Farinha, M.J. Trigueiros. A data quality metamodel extension to CWM. In Proc. of the 4th Asia-Pacific Conference on Conceptual Modeling. pp. 17-26, 1997.

[12] Y.W. Lee, L.L. Pipino, J.D. Funk, R.Y. Wang. Journey to Data Quality. The MIT Press,2006.

[13] A.F. Karr, A.P. Sanil,D.L. Banks. Data quality: A statistical perspective. Statist. Method. vol. 3, 137-173, 2006.

[14] C. Ordonez, J. Garcia-Garcia. Referential integrity quality metrics. Decis. Support Syst. vol. 44, pp. 495-508,2008.

[15] L.L. Pipino, Y.W. Lee, R.Y. Wang. Data quality assessment. Comm. ACM 45, pp. 211-218, 2002.

[16] Y. Wand, R.Y. Wang. Anchoring data quality dimensions in ontological foundations. Comm. ACM vol. 39, pp. 86-95, 1996.

[17] R.C. Dubes, A.K. Jain. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey 07632, 1988.

[18] J. Hartigan. Clustering Algorithms. John Wiley & Sons, New York, NY, 1975.

[19] J. Hartigan, M. Wong. Algorithm AS136: A k-means clustering algorithm. Applied Statistics, vol. 28, pp. 100-108, 1979.

[20] D. Fisher (1987). Knowledge acquisition via incremental conceptual clustering. Machine Learning. 2(2):139-172.

[21] J. H. Gennari, P. Langley, D. Fisher (1990). Models of incremental concept formation. Artificial Intelligence. 40:11-61.

[22] Hochbaum, Shmoys. A best possible heuristic for the k-center problem. Mathematics of Operations Research. vol. 10(2), pp. 180-184, 1985.

[23] Sanjoy Dasgupta: Performance Guarantees for Hierarchical Clustering. In: 15th Annual Conference on Computational Learning Theory, 351-363, 2002.

[24] M. Ester, H-P Kriegel, J. Sander, X Xu. . A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. of the 2nd ACM SIGKDD, pp 226-231, 1996.

[25] Dan Pelleg, Andrew W. Moore: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: Seventeenth International Conference on Machine Learning, 727-734, 2000.

[26] P. Berkhin. Survey Of Clustering Data Mining Techniques, 2002.

[27] C. Fisher, E. Lauria, C. Matheus. In search of an accuracy metric. In Proceedings of the 12th Int. Conf. on Information Quality, 2007.

[28] B.D. Klein, D.L. Goodhue, G.B. Davis. Can humans detect errors in data? Impact of base rates, incentives, and goals. MIS Quart. 21, pp. 169-194, 1997.

[29] R. Blake, P. Mangiameli. The Effects and Interactions of Data Quality and Problem Complexity on Classification. Journal of Data and Information Quality (JDIQ), vol. 2(2), Feb. 2011.

[30] A. Parssian,S. Sarkar,V.S. Jacob. Assessing data quality for information products: Impact of selection, projection, and cartesian product. Manag. Sci. vol. 50, pp. 967–982, 2004.

[31] M. Ge, M. Helfret. A framework to assess decision quality using information quality dimensions. In Proc. of the Int. Conf. on Information Quality, 2006.

[32] M.A. Goncalves, B.L. Moreira, E. A. Fox, L.T. Watson. "What is a good digital library?" – A quality model for digital libraries. Information Processing & Management, vol. 43(5), Sept. 2007, pp 1416-1437.

Appendix A: Detailed Experimental results

In tables 5, 6, 7, 8 and 9 average cluster validity values for each dataset and different data quality levels and algorithms are presented. We used them for define the absolute changes value.

$$\Delta_M = |\overline{f_Q(S_{Dataset}(A, Q\_Dim, High))} - \overline{f_Q(S_{Dataset}(A, Q\_Dim, Medium))}|$$

$$\Delta_L = |\overline{f_Q(S_{Dataset}(A, Q\_Dim, High))} - \overline{f_Q(S_{Dataset}(A, Q\_Dim, Low))}|$$

where A is the clustering algorithm, Q_Dim is the data quality dimension, S is the obtained cluster structure and f_Q is the cluster validity index. For all used validity indecies $f_Q(S) \in [0, 1]$ and we considered the quality changes as high if $\Delta \geq 0.04$. According to absolute changes value and Wilcoxon statistical test results we constructed four tables (one table for each data quality dimension), which described data quality influence to clustering results. Table symbols legend looks like following:

■ — significant negative impact

▼ — significant but small negative impact

□ — unsignificant changes
▲ — significant but small positive impact
♦ — significant positive impact

Table 5: Cardiotocography: averages

			Rand	Jaccard	FM	F1
Accuracy	K-Means	High	0.91	0.53	0.69	0.76
		Medium	0.9	0.49	0.66	0.75
		Low	0.9	0.47	0.64	0.74
	Farthest First	High	0.58	0.23	0.41	0.44
		Medium	0.55	0.19	0.37	0.39
		Low	0.44	0.18	0.38	0.36
	DBScan	High	0.47	0.3	0.51	0.47
		Meduim	0.38	0.19	0.42	0.36
		Low	0.38	0.18	0.39	0.34
	XMeans	High	0.89	0.49	0.65	0.71
		Medium	0.88	0.45	0.62	0.7
		Low	0.88	0.43	0.6	0.69
Completeness	K-Means	High	0.93	0.6	0.75	0.78
		Medium	0.89	0.48	0.65	0.73
		Low	0.87	0.41	0.58	0.7
	Farthest First	High	0.89	0.49	0.66	0.74
		Medium	0.84	0.42	0.59	0.69
		Low	0.79	0.35	0.52	0.63
	DBScan	High	0.98	0.9	0.95	0.96
		Medium	0.95	0.73	0.84	0.91
		Low	0.92	0.6	0.75	0.87
	XMeans	High	0.91	0.57	0.72	0.74
		Medium	0.87	0.43	0.6	0.68
		Low	0.84	0.38	0.55	0.66
Consistency	K-Means	High	0.92	0.57	0.73	0.77
		Medium	0.9	0.47	0.64	0.72
		Low	0.87	0.39	0.56	0.68
	Farthest First	High	0.87	0.43	0.6	0.68
		Medium	0.79	0.31	0.48	0.58
		Low	0.76	0.25	0.4	0.52
	DBScan	High	0.98	0.87	0.93	0.96

		Medium	0.94	0.66	0.79	0.89
		Low	0.9	0.5	0.67	0.82
	XMeans	High	0.91	0.58	0.73	0.74
		Medium	0.87	0.43	0.6	0.67
		Low	0.85	0.38	0.55	0.65
Timeliness	K-Means	High	0.92	0.59	0.74	0.75
		Medium	0.9	0.56	0.71	0.73
		Low	0.91	0.57	0.72	0.73
	Farthest First	High	0.92	0.65	0.78	0.8
		Medium	0.95	0.73	0.84	0.85
		Low	0.93	0.66	0.79	0.82
	DBScan	High	0.98	0.88	0.94	0.97
		Medium	0.95	0.74	0.85	0.93
		Low	0.92	0.6	0.75	0.89
	XMeans	High	0.9	0.54	0.7	0.7
		Medium	0.92	0.63	0.76	0.74
		Low	0.91	0.61	0.75	0.74

Table 6: Image Segmentation: averages

			Rand	Jaccard	FM	F1
Accuracy	K-Means	High	0.86	0.38	0.56	0.65
		Medium	0.85	0.37	0.54	0.64
		Low	0.84	0.34	0.52	0.61
	Farthest First	High	0.5	0.2	0.42	0.41
		Medium	0.47	0.19	0.4	0.38
		Low	0.39	0.17	0.39	0.35
	DBScan	High	0.71	0.23	0.4	0.48
		Medium	0.7	0.22	0.39	0.46
		Low	0.7	0.21	0.37	0.46
	XMeans	High	0.87	0.33	0.5	0.6
		Medium	0.86	0.31	0.47	0.59
		Low	0.85	0.3	0.46	0.58
Completeness	K-Means	High	0.86	0.38	0.55	0.65
		Medium	0.84	0.33	0.5	0.62
		Low	0.83	0.3	0.46	0.6
	Farthest First	High	0.73	0.27	0.45	0.52
		Medium	0.64	0.22	0.42	0.48
		Low	0.66	0.23	0.42	0.5
	DBScan	High	0.79	0.26	0.42	0.56
		Medium	0.77	0.22	0.36	0.53
		Low	0.74	0.17	0.3	0.47
	XMeans	High	0.86	0.29	0.45	0.57
		Medium	0.85	0.26	0.42	0.54
		Low	0.84	0.24	0.39	0.52
Consistency	K-Means	High	0.86	0.37	0.54	0.64
		Medium	0.84	0.31	0.47	0.59
		Low	0.82	0.26	0.42	0.55
	Farthest First	High	0.65	0.24	0.44	0.49
		Medium	0.52	0.19	0.4	0.41
		Low	0.53	0.18	0.38	0.39
	DBScan	High	0.79	0.26	0.43	0.56
		Medium	0.77	0.21	0.35	0.52
		Low	0.73	0.17	0.3	0.47
	XMeans	High	0.86	0.3	0.47	0.59

		Medium	0.84	0.25	0.41	0.53
		Low	0.83	0.22	0.37	0.5
Timeliness	K-Means	High	0.86	0.39	0.56	0.65
		Medium	0.87	0.4	0.58	0.66
		Low	0.87	0.4	0.57	0.65
	Farthest First	High	0.74	0.3	0.5	0.58
		Medium	0.72	0.29	0.48	0.54
		Low	0.71	0.27	0.47	0.54
	DBScan	High	0.71	0.21	0.38	0.46
		Medium	0.69	0.18	0.32	0.42
		Low	0.66	0.15	0.29	0.38
	XMeans	High	0.86	0.3	0.47	0.56
		Medium	0.87	0.29	0.46	0.56
		Low	0.86	0.32	0.49	0.58

Table 7: Page Blocks: averages

			Rand	Jaccard	FM	F1
Accuracy	K-Means	High	0.43	0.32	0.55	0.57
		Medium	0.43	0.32	0.54	0.57
		Low	0.42	0.31	0.53	0.55
	Farthest First	High	0.82	0.81	0.9	0.86
		Medium	0.82	0.82	0.91	0.87
		Low	0.82	0.82	0.9	0.87
	DBScan	High	0.82	0.81	0.9	0.86
		Medium	0.81	0.8	0.89	0.86
		Low	0.81	0.8	0.89	0.85
	XMeans	High	0.35	0.22	0.45	0.47
		Medium	0.35	0.21	0.44	0.46
		Low	0.34	0.21	0.44	0.44
Completeness	K-Means	High	0.43	0.32	0.55	0.57
		Medium	0.44	0.33	0.55	0.59
		Low	0.45	0.34	0.57	0.62
	Farthest First	High	0.8	0.8	0.89	0.86
		Medium	0.77	0.76	0.86	0.84
		Low	0.82	0.82	0.9	0.86
	DBScan	High	0.83	0.83	0.91	0.87
		Medium	0.83	0.83	0.91	0.86
		Low	0.83	0.83	0.91	0.86
	XMeans	High	0.36	0.22	0.45	0.48
		Medium	0.36	0.22	0.46	0.49
		Low	0.37	0.24	0.47	0.49
Consistency	K-Means	High	0.44	0.33	0.56	0.6
		Medium	0.42	0.31	0.54	0.58
		Low	0.41	0.3	0.52	0.56
	Farthest First	High	0.82	0.81	0.9	0.87
		Medium	0.81	0.81	0.9	0.86
		Low	0.76	0.75	0.85	0.83
	DBScan	High	0.82	0.81	0.9	0.86
		Medium	0.8	0.79	0.89	0.85
		Low	0.79	0.78	0.88	0.85
	XMeans	High	0.35	0.21	0.44	0.46

		Medium	0.34	0.2	0.43	0.45
		Low	0.33	0.2	0.42	0.43
Timeliness	K-Means	High	0.43	0.32	0.55	0.6
		Medium	0.41	0.3	0.52	0.58
		Low	0.39	0.28	0.5	0.56
	Farthest First	High	0.82	0.82	0.9	0.87
		Medium	0.77	0.76	0.86	0.83
		Low	0.8	0.8	0.89	0.85
	DBScan	High	0.83	0.83	0.91	0.86
		Medium	0.86	0.85	0.92	0.87
		Low	0.88	0.87	0.93	0.87
	XMeans	High	0.36	0.22	0.45	0.47
		Medium	0.39	0.28	0.5	0.55
		Low	0.37	0.25	0.48	0.51

Table 8: Pen Digits: averages

			Rand	Jaccard	FM	F1
Accuracy	K-Means	High	0.91	0.42	0.6	0.71
		Medium	0.91	0.39	0.56	0.69
		Low	0.9	0.37	0.55	0.68
	Farthest First	High	0.56	0.14	0.31	0.33
		Medium	0.51	0.12	0.3	0.3
		Low	0.46	0.12	0.29	0.28
	DBScan	High	0.26	0.12	0.33	0.24
		Medium	0.23	0.1	0.29	0.18
		Low	0.28	0.1	0.28	0.18
	XMeans	High	0.89	0.38	0.56	0.67
		Medium	0.89	0.37	0.54	0.66
		Low	0.89	0.34	0.52	0.64
Completeness	K-Means	High	0.91	0.42	0.59	0.71
		Medium	0.9	0.38	0.55	0.68
		Low	0.89	0.32	0.49	0.64
	Farthest First	High	0.82	0.22	0.38	0.48
		medium	0.81	0.21	0.37	0.47
		Low	0.79	0.2	0.35	0.46
	DBScan	High	0.78	0.25	0.44	0.55
		Medium	0.79	0.21	0.38	0.52
		Low	0.78	0.18	0.33	0.49
	XMeans	High	0.88	0.36	0.54	0.64
		Medium	0.87	0.33	0.51	0.62
		Low	0.86	0.3	0.47	0.6
Consistency	K-Means	High	0.91	0.4	0.57	0.69
		Medium	0.89	0.33	0.5	0.65
		Low	0.88	0.27	0.43	0.6
	Farthest First	High	0.82	0.22	0.37	0.47
		Medium	0.8	0.18	0.33	0.43
		Low	0.79	0.16	0.29	0.39
	DBScan	High	0.78	0.24	0.43	0.55
		Medium	0.77	0.2	0.36	0.51
		Low	0.76	0.17	0.31	0.5
	XMeans	High	0.87	0.34	0.51	0.63

		Medium	0.86	0.28	0.45	0.58
		Low	0.84	0.23	0.39	0.52
Timeliness	K-Means	High	0.91	0.41	0.58	0.7
		Medium	0.91	0.4	0.57	0.69
		Low	0.91	0.39	0.56	0.69
	Farthest First	High	0.51	0.14	0.32	0.33
		Medium	0.48	0.12	0.3	0.3
		Low	0.47	0.12	0.3	0.29
	DBScan	High	0.23	0.11	0.3	0.2
		Medium	0.34	0.1	0.27	0.18
		Low	0.43	0.1	0.26	0.18
	XMeans	High	0.89	0.39	0.56	0.67
		Medium	0.87	0.35	0.53	0.63
		Low	0.87	0.33	0.51	0.6

Table 9: Wall Following: averages

			Rand	Jaccard	FM	F1
Accuracy	K-Means	High	0.59	0.21	0.35	0.43
		Medium	0.59	0.21	0.35	0.43
		Low	0.59	0.21	0.35	0.43
	Farthest First	High	0.44	0.3	0.5	0.48
		Medium	0.41	0.31	0.52	0.48
		Low	0.39	0.32	0.54	0.49
	DBScan	High	0.36	0.32	0.55	0.48
		Medium	0.36	0.33	0.56	0.49
		Low	0.38	0.32	0.54	0.48
	XMeans	High	0.63	0.14	0.27	0.34
		Medium	0.62	0.16	0.29	0.36
		Low	0.61	0.17	0.3	0.37
Completeness	K-Means	High	0.6	0.21	0.35	0.43
		Medium	0.6	0.21	0.35	0.43
		Low	0.6	0.21	0.34	0.42
	Farthest First	High	0.52	0.26	0.42	0.45
		Medium	0.51	0.26	0.43	0.46
		Low	0.47	0.28	0.46	0.47
	DBScan	High	0.44	0.27	0.45	0.45
		Medium	0.44	0.28	0.46	0.45
		Low	0.43	0.29	0.47	0.46
	XMeans	High	0.6	0.21	0.35	0.42
		Medium	0.6	0.21	0.34	0.42
		Low	0.6	0.21	0.34	0.42
Consistency	K-Means	High	0.6	0.21	0.34	0.42
		Medium	0.59	0.2	0.34	0.41
		Low	0.59	0.19	0.33	0.39
	Farthest First	High	0.52	0.26	0.43	0.46
		Medium	0.52	0.26	0.42	0.45
		Low	0.52	0.26	0.42	0.45
	DBScan	High	0.46	0.26	0.43	0.44
		Medium	0.48	0.25	0.41	0.43
		Low	0.5	0.24	0.4	0.42
	XMeans	High	0.6	0.21	0.34	0.42

		Medium	0.6	0.2	0.33	0.4
		Low	0.59	0.19	0.33	0.39
Timeliness	K-Means	High	0.59	0.21	0.35	0.43
		Medium	0.6	0.2	0.34	0.42
		Low	0.6	0.2	0.34	0.41
	Farthest First	High	0.37	0.32	0.55	0.48
		Medium	0.39	0.31	0.53	0.48
		Low	0.43	0.29	0.48	0.46
	DBScan	High	0.38	0.31	0.53	0.48
		Medium	0.44	0.28	0.46	0.45
		Low	0.47	0.26	0.43	0.42
	XMeans	High	0.59	0.21	0.35	0.43
		Medium	0.59	0.21	0.35	0.42
		Low	0.56	0.26	0.41	0.48

As it shown in the table with accuracy influence 10, in most cases accepted accuracy levels produce small negative impact on clustering quality. In case of KMeans an FarthestFirst algorithms and medium accuracy level the changes are even insignificant. Therefore we supposed that percentage of incorrect data for medium and low accuracy levels can be increased.

Table 10: Accuracy: statistical results

			Rand	Jaccard	FM	F1
K-Means	Medium	Cardio	▼	■	■	□
		Image	▼	▼	▼	□
		Page Blocks	□	□	□	▼
		Pen Digits	▼	▼	■	▼
		Wall Following	▼	□	□	□
K-Means	Low	Cardio	▼	■	■	▼
		Image	▼	■	■	▼
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	■	■	▼
		Wall Following	▼	▼	▼	▼
FarthestFirst	Medium	Cardio	□	■	■	□
		Image	□	▼	▼	▼
		Page Blocks	□	□	□	□
		Pen Digits	□	▼	▼	□
		Wall Following	▼	□	▲	□
FarthestFirst	Low	Cardio	■	■	□	■
		Image	■	▼	▼	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	■	▼	▼	■
		Wall Following	■	▲	♦	□
DBScan	Medium	Cardio	□	■	■	■
		Image	▼	▼	▼	▼
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▲	▼	▼	■
		Wall Following	▲	▼	▼	▼
DBScan	Low	Cardio	□	■	■	■
		Image	▲	▼	▼	▼
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▲	▼	■	■

		Wall Following	▲	▼	▼	▼
XMeans	Medium	Cardio	▼	▼	▼	□
		Image	▼	▼	▼	▼
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	▼	▼	▼
		Wall Following	▼	▲	▲	▲
XMeans	Low	Cardio	▼	▼	▼	▼
		Image	▼	▼	■	▼
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	■	■	▼
		Wall Following	▼	▲	▲	▲

According to table 11 we can observe that low completeness levels have high negative impact on Cardiotocography and Image Segmentation datasets, low negative impact on Pen Digits and Page Blocks data, and different variants of impact on Wall Following dataset, depending on the clustering algorithm. In any case accepted completeness levels affect on clustering quality more than accuracy ones.

Table 11: Completeness: statistical results

			Rand	Jaccard	FM	F1
K-Means	Medium	Cardio	■	■	■	■
		Image	▼	■	■	■
		Page Blocks	▲	▲	▲	▲
		Pen Digits	▼	■	■	▼
		Wall Following	▼	□	□	□
K-Means	Low	Cardio	■	■	■	■
		Image	▼	■	■	■
		Page Blocks	▲	▲	▲	▲
		Pen Digits	▼	■	■	■
		Wall Following	▼	▼	▼	▼
FarthestFirst	Medium	Cardio	■	■	■	■
		Image	■	■	▼	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	□	▼	▼	□
		Wall Following	□	□	□	▲
FarthestFirst	Low	Cardio	■	■	■	■
		Image	■	■	▼	▼
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	▼	▼	▼
		Wall Following	■	▲	▲	▲
DBScan	Medium	Cardio	■	■	■	■
		Image	▼	■	■	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▲	■	■	▼
		Wall Following	▼	▲	▲	▲
DBScan	Low	Cardio	■	■	■	■
		Image	▼	■	■	■
		Page Blocks	▼	▼	▼	▼

		Pen Digits	▼	■	■	■
		Wall Following	▼	▲	▲	▲
XMeans	Medium	Cardio	■	■	■	■
		Image	▼	▼	▼	▼
		Page Blocks	▲	▲	▲	▲
		Pen Digits	▼	▼	▼	▼
		Wall Following	▼	▼	▼	▼
XMeans	Low	Cardio	■	■	■	■
		Image	▼	■	■	▼
		Page Blocks	▲	▲	▲	▲
		Pen Digits	▼	■	■	■
		Wall Following	▼	□	▼	□

Consistency errors are dependent on dataset structure too 12. They produce low negative effect on Page Blocks data (because it is biased) and on Wall Following data (because it has bad cluster structure) and high negative effect on other datasets. Consistency has the most negative impact on cluster validity. All used algorithms work badly with such kind of errors.

Table 12: Consistency: statistical results

			Rand	Jaccard	FM	F1
K-Means	Medium	Cardio	▼	■	■	■
		Image	▼	■	■	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	■	■	■
		Wall Following	▼	▼	▼	▼
K-Means	Low	Cardio	▼	■	■	■
		Image	▼	■	■	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	■	■	■
		Wall Following	▼	▼	▼	▼
FarthestFirst	Medium	Cardio	■	■	■	■
		Image	■	■	■	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	■	■	■
		Wall Following	□	□	□	□
FarthestFirst	Low	Cardio	■	■	■	■
		Image	■	■	■	■
		Page Blocks	■	■	■	■
		Pen Digits	▼	■	■	■
		Wall Following	□	▼	□	▼
DBScan	Medium	Cardio	■	■	■	■
		Image	■	■	■	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▲	■	■	■
		Wall Following	▲	▼	▼	▼
DBScan	Low	Cardio	■	■	■	■
		Image	▼	■	■	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	■	■	■

		Wall Following	♦	▼	▼	▼
XMeans	Medium	Cardio	■	■	■	■
		Image	▼	■	■	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	■	■	■
		Wall Following	▼	▼	▼	▼
XMeans	Low	Cardio	■	■	■	■
		Image	▼	■	■	■
		Page Blocks	▼	▼	▼	▼
		Pen Digits	▼	■	■	■
		Wall Following	▼	▼	▼	□

According to table 13 partitioning algorithms have unsigniffficant quality changes (in case of KMeans and FarthestFirst algorithms) or even some improvement (in case of XMeans algorithms) on different timeliness levels. DBScan algorithm has low negative quality changes on medium timeliness level and high negative changes on low timeliness level. But we should note that for some datasets (for example Pen Digits) difference between DBScan clustering on ideal data and DBScan clustering on high quality data is already very high.

Table 13: Timeliness: statistical results

			Rand	Jaccard	FM	F1
K-Means	Medium	Cardio	□	□	□	□
		Image	▲	▲	▲	□
		Page Blocks	▼	▼	▼	▼
		Pen Digits	□	▼	▼	▼
		Wall Following	▲	▼	▼	▼
K-Means	Low	Cardio	□	□	□	□
		Image	▲	□	□	□
		Page Blocks	■	■	■	■
		Pen Digits	□	▼	▼	▼
		Wall Following	▲	▼	▼	▼
FarthestFirst	Medium	Cardio	▲	▲	▲	▲
		Image	□	▼	▼	▼
		Page Blocks	□	□	□	□
		Pen Digits	□	▼	▼	▼
		Wall Following	▲	▼	▼	▼
FarthestFirst	Low	Cardio	□	□	□	□
		Image	▼	▼	▼	■
		Page Blocks	□	□	□	□
		Pen Digits	□	▼	▼	■
		Wall Following	♦	▼	□	▼
DBScan	Medium	Cardio	▼	■	■	■
		Image	▼	▼	■	■
		Page Blocks	▲	▲	▲	▲
		Pen Digits	▲	▼	▼	▼
		Wall Following	▲	▼	■	▼
DBScan	Low	Cardio	■	■	■	■

		Image	■	■	■	■
		Page Blocks	▲	▲	▲	▲
		Pen Digits	♦	▼	■	▼
		Wall Following	♦	■	■	■
XMeans	Medium	Cardio	▲	♦	♦	♦
		Image	□	□	□	□
		Page Blocks	▲	♦	♦	♦
		Pen Digits	▼	■	▼	▼
		Wall Following	▲	▼	▼	▼
XMeans	Low	Cardio	▲	♦	♦	♦
		Image	□	▲	▲	▲
		Page Blocks	▲	▲	▲	▲
		Pen Digits	▼	■	■	■
		Wall Following	▼	♦	♦	♦