IEEE P2520.1 Working Group #12
Meeting Minutes
28 March 2022
WG Chair: James Covington
WG Secretary: H. Troy Nagle (Interim)
Meeting link:
https://ncsu.zoom.us/j/95028992587?pwd=SzNKT0pXNW9UL1loZnZMT25jU1dPdz09

1. **Call to Order**
   Chair called meeting to order at 10:04 AM EST. He announced that the meeting was being recorded for the purpose of preparing minutes.

2. **Roll Call and Disclosure of Affiliation**
   Affiliation FAQs: http://standards.ieee.org/faqs/affiliation.html
   The Chair asked the participants to sign-in at this link:
   https://docs.google.com/spreadsheets/d/1x3Le7jd_5h3bgiNcYMZIfjIbzE2XdE0U8Daon00O8Ks/edit#gid=0.
   The Chair asked the Secretary to check for a quorum. No new members were participating. The List of Participants is shown in **Attachment A**. A quorum was achieved (16 of the 18 voting members were present).

3. **Approval of Agenda**
   The Chair asked for approval of the agenda. Troy Nagle made the motion; Susan Schiffman seconded. Without objection to unanimous consent, the motion was adopted.

4. **Approval of Previous Meeting Minutes**
   The Chair asked for approval of the February 28 Meeting Minutes as circulated. Susana Palma made the motion; Paul Kagan seconded. Without objection to unanimous consent, the motion was adopted.

5. **IEEE-SA Patent & Copyright Policies**
   a. Call for Patents
      https://development.standards.ieee.org/myproject/Public/mytools/mob/slideset.pdf
      Per standard IEEE SA WG meeting practice, the Chair reviewed the required policy regarding potentially essential patents. No one raised concerns for consideration.
   b. Copyright Policy https://standards.ieee.org/ipr/copyright-materials.html
      Per standard IEEE-SA WG meeting practice, the Chair reviewed the required policy regarding copyrights. There were no questions or concerns.

6. **Technical Presentation:**
   The major focus for this meeting was a presentation from Fengchun Tian covering recent data analysis experiments using the Silhouette scoring method. His slides are included in **Attachment B**. The presentation reviewed the processing steps for the method and presented results from three example data sets. Fengchen's conclusion stated the following:

- Calculating the Silhouette Coefficient (SC) after K-means can produce false clustering.
- Replace the SC with the mean Silhouette Value (MSV) without K-means. Instead use only the raw data or LDA.
- Advantage: We know the labels for each chemical thus we know the number of clusters. Thus, the MSV is a single measure reflecting the ability of the EUT in differentiating chemicals.

After the presentation, a Q&A session followed.

The Silhouette method, compared to other clustering techniques, has the advantage of being able to generate a single parameter value. Mathematically it is easy to implement. The disadvantage is that one can get artifacts, and this can generate more clusters than should occur.

Clustering: Diverse chemicals easily separate using PCA. PCA tools are readily available and makes no assumptions about the data. However, other reporting methods that look at Euclidean or Manhattan distance could be more useful as an unsupervised classifier. If using Euclidean distance, a confidence interval can be employed, and linear methods can determine overlapping data points. Very simple approaches use nearest neighbor, farthest neighbor, or Ward's method. Do we need to specify a clustering method, or let the manufacturer use any method they want (hierarchical, Sammon mapping, PCA, etc.)? The most important issue is how we score the clustering. The primary advantage of the Silhouette method is that we can specify a single value as a passing criterion. We want a method that is simple to calculate for all three of our performance levels. A major concern is that, in many practical applications, one can get non-Gaussian clusters. Can the Silhouette method separate non-Gaussian clusters, or will we get artifacts? In Ward's method, we can put confidence levels on each result and reduce the score to a single value.

We can make classification easier by our choice of chemical options in Appendix A. For example, Fox uses PCA, and it readily separates ethanol, acetone, and isopropanol. PCA is included in the signal processing options for most commercially available enose systems.

Next the Chair focused the discussion on the working draft of the standard. In the current version, specific levels for MSV are included as place holders. The final thresholds will be values much higher than 0.5. Shall we allow the operator to give the EUT the number of clusters to generate? Regarding the number of clusters required for each testing level, that is a decision we will make after running some examples through operational-EUT devices that are available to WG members.

PCA vs. LDA: It was suggested that we first try PCA and only resort to LDA if necessary. Should we allow the operator to avoid using PCA and use the raw data prior to computing the Silhouette coefficients?

Appendix A: The table of chemicals will have a range of choices. Simple, safe options will be at the top to facilitate passage of P2520.1. For example, we could include ethanol, isobutylene, and propanol. More application specific options will be included for P2520.x.1 standards.

Consensus: At this time, we agreed to use the MSV scoring method for Level 1.

Next steps: Troy and Susan are currently editing the draft standard. The Chair will send out a revised copy soon and highlight in green text some specific sections that he would like WG members to review at our next meeting.

7. **New Business/Activities for the Next Meeting**
   There was no New Business.

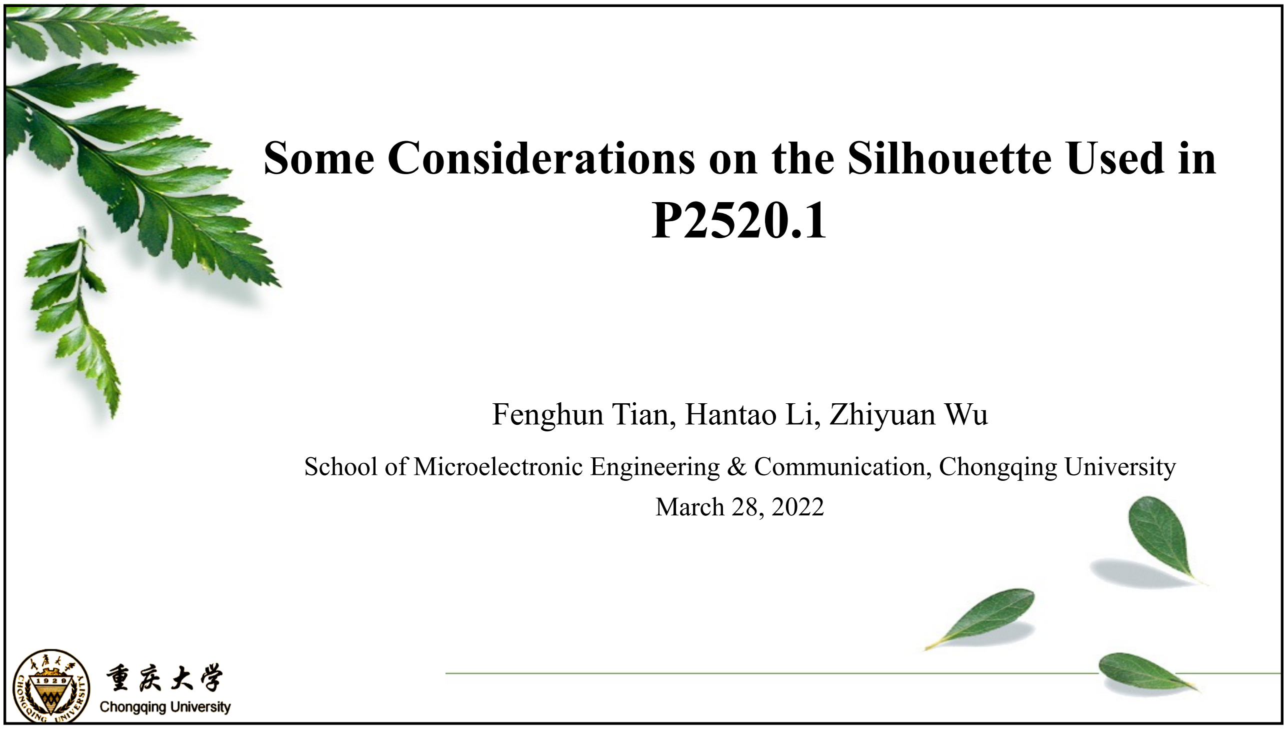
8. **Future Meetings**
   The Chair announced the next meeting of the WG will take place on April 25.

9. **Adjourn**
   The meeting time-period having expired, without objection to unanimous consent, the Chair adjourned the meeting at 11:05 AM.

Attachment A: Participants (18)

NAME	AFFILIATION
Carlos Diaz	Ambiente et Odora
Christopher Jensen	Self
Cyril Herrier	Aryballe
Duke Oeba	Self, Oregon State University
Ehsan Danesh	Alphasense Ltd
Etienne Bultel	Aryballe
Ettore Massera	ENEA
Fengchun Tian	Chongqing University
Hua-Yao Li	Huazhong University of Science and Technology
James Covington	Professor, School of Engineering, University of Warwick
Katayoun Emadzadeh	Self
Krishna Persaud	University of Manchester
Paul Kagan	AWLDM Systems
Radislav Potyrailo	GE Research
Sandrine Isz	Alpha-MOS
Susan Schiffman	NC State University
Susana Palma	NOVA University of Lisbon
Troy Nagle	NC State University

Attachment B: Presentation Slides



Some Considerations on the Silhouette Used in P2520.1

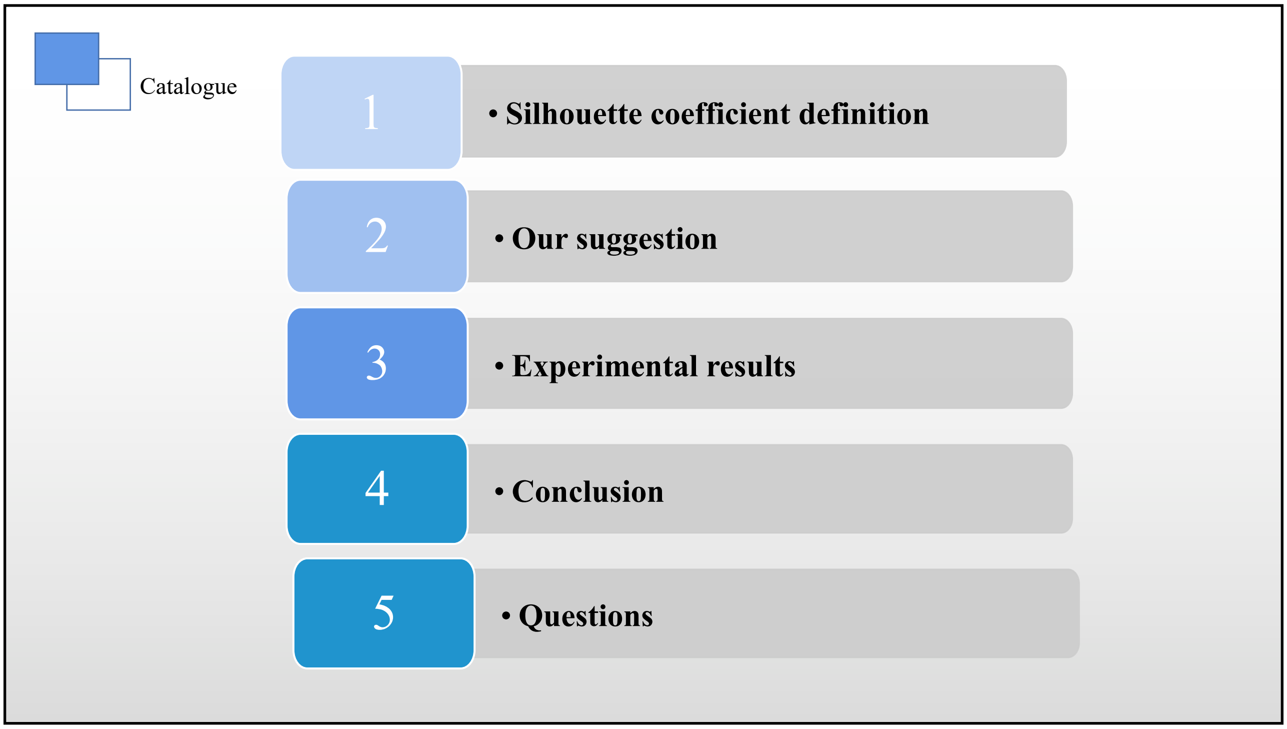
Fenghun Tian, Hantao Li, Zhiyuan Wu

School of Microelectronic Engineering & Communication, Chongqing University

March 28, 2022

重庆大学
Chongqing University

1

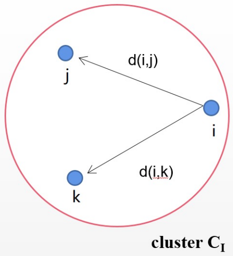


Catalogue

1. • **Silhouette coefficient definition**
2. • **Our suggestion**
3. • **Experimental results**
4. • **Conclusion**
5. • **Questions**

2

1.1 Definition of Silhouette



Assume the data have been clustered via any technique, such as k-means, into κ clusters. For data point $i \in C_I$,

$$\boxed{a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)} \qquad (1)$$

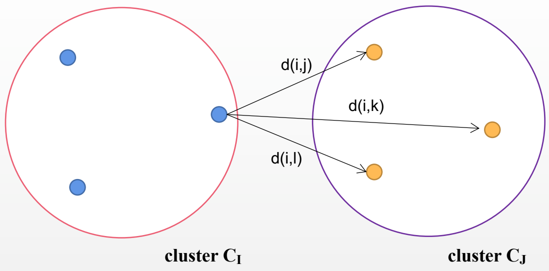
a measure of how well i is assigned to its cluster

$a(i)$ be the mean distance between i and all other points in the same cluster,

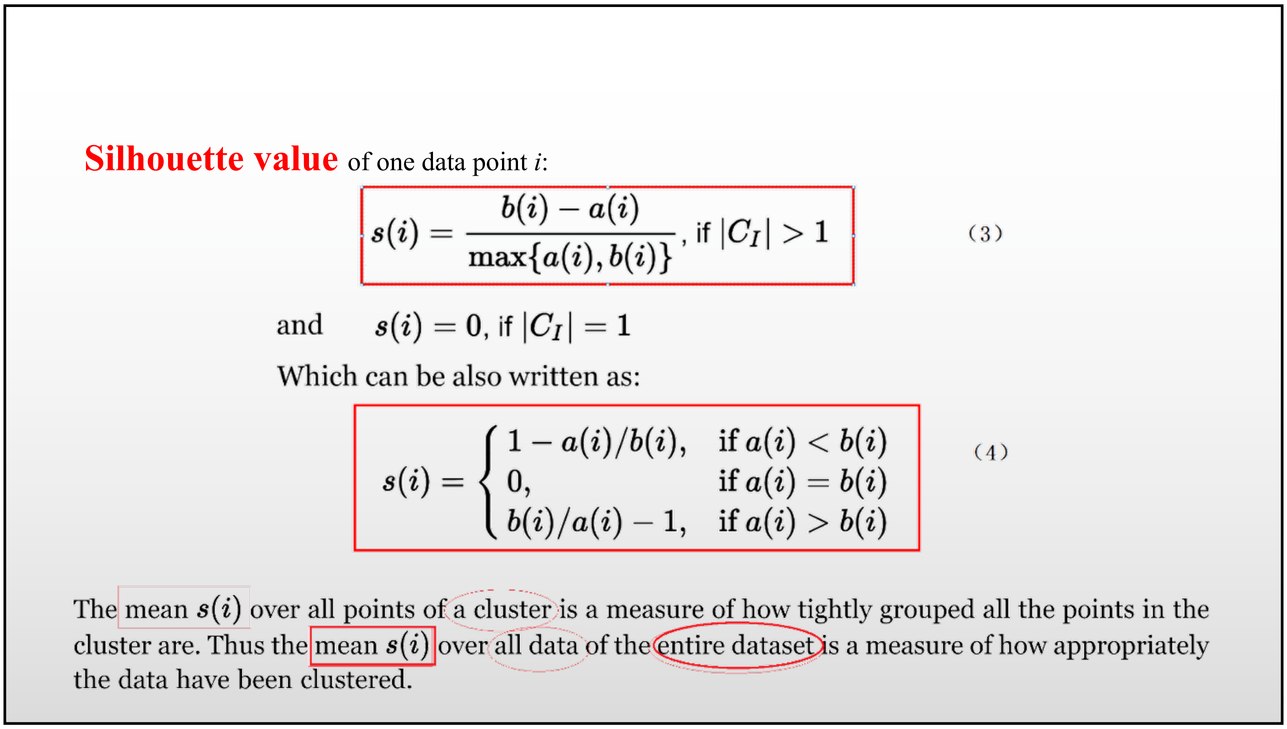
$|C_I|$ is the number of points belonging to cluster i,

$d(i,j)$ is the distance between data points i and j in the cluster C_I

$$\boxed{b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)} \qquad (2)$$



Where $b(i)$ is the smallest mean distance of point i to all points in any other cluster, of which i is not a member. The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.



Silhouette value of one data point *i*:

$$\boxed{s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \text{ if } |C_I| > 1} \qquad (3)$$

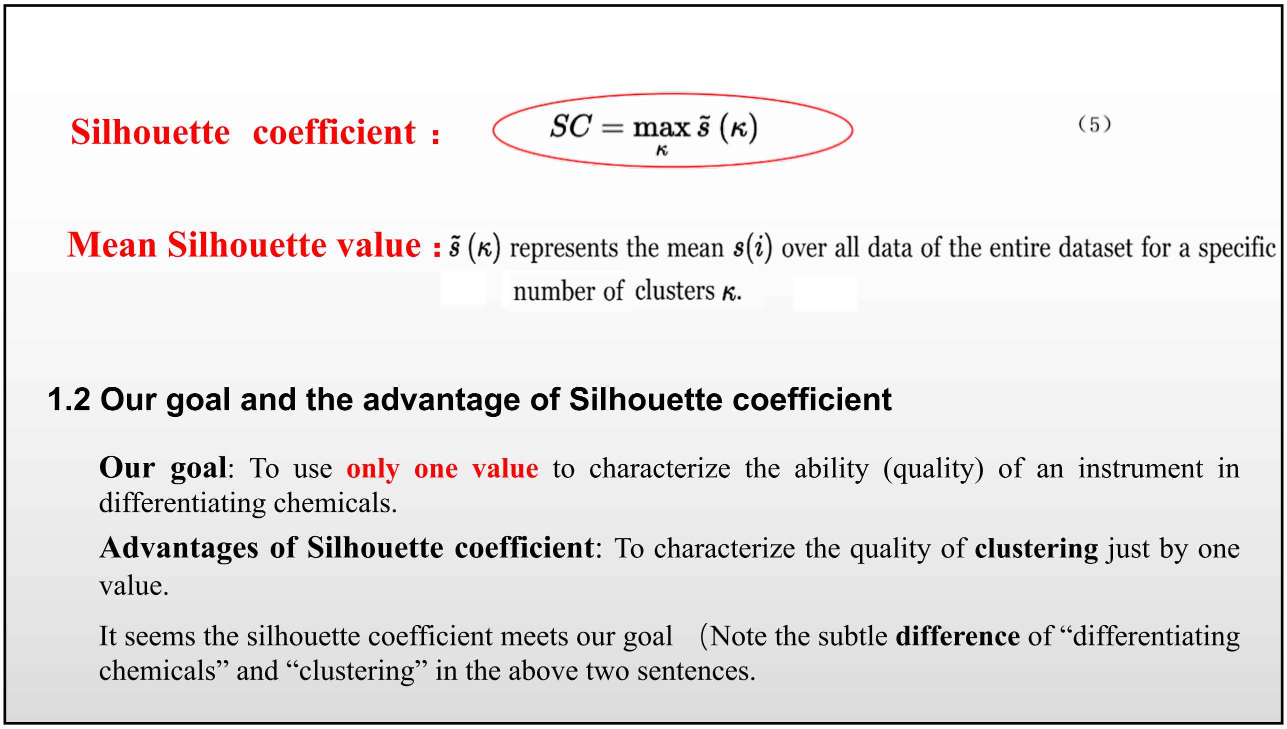
and $s(i) = 0$, if $|C_I| = 1$

Which can be also written as:

$$\boxed{s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}} \qquad (4)$$

The mean $s(i)$ over all points of a cluster is a measure of how tightly grouped all the points in the cluster are. Thus the mean $s(i)$ over all data of the entire dataset is a measure of how appropriately the data have been clustered.

5



Silhouette coefficient : $$\boxed{SC = \max_{\kappa} \tilde{s}(\kappa)} \qquad (5)$$

Mean Silhouette value : $\tilde{s}(\kappa)$ represents the mean $s(i)$ over all data of the entire dataset for a specific number of clusters κ.

1.2 Our goal and the advantage of Silhouette coefficient

Our goal: To use **only one value** to characterize the ability (quality) of an instrument in differentiating chemicals.

Advantages of Silhouette coefficient: To characterize the quality of **clustering** just by one value.

It seems the silhouette coefficient meets our goal （Note the subtle **difference** of "differentiating chemicals" and "clustering" in the above two sentences.

6

1.3 The problem lies in Silhouette coefficient calculation after K-means clustering

Example 1

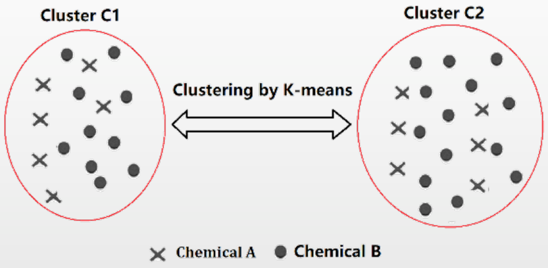


Fig. 1 The two clusters obtained by K-means clustering

Falsely clustered case

↓

It seems two clusters are fairly isolated

↓

Its *SC* will be quite high

↓

But a high SC does not mean a good ability of differentiating chemicals

1.3 The problem lies in Silhouette coefficient calculation after K-means clustering

Example 2

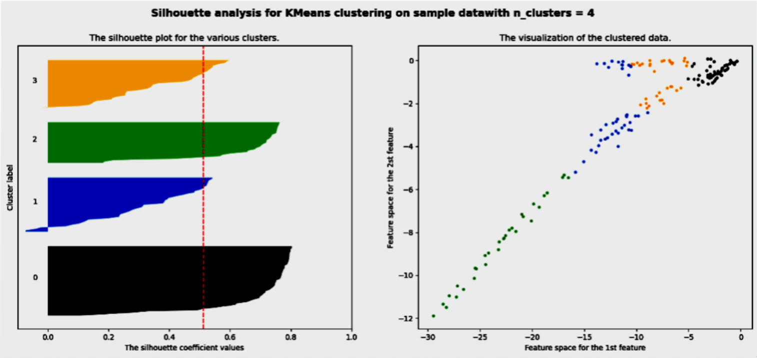


Fig. 2 Silhouette after K-means(left) and visualization of the clustered data points by PCA (right)

$$SC = \max_{\kappa} \tilde{s}(\kappa) \qquad (5)$$

Falsely optimal number of clustering k=3 （True k=4）

↓

The Silhouette coefficient *SC* after K-means is **0.607** with **k=3**, while $\tilde{s}(\kappa)$ =**0.513** for **k=4**.

↓

Silhouette coefficient calculating after the clustering of K-means cannot correctly characterize the real differentiating ability of the instrument due to the existence of misclassification

Data from UCI database: http://archive.ics.uci.edu/ml/datasets/Twin+gas+sensor+arrays; (Cf. https://doi.org/10.1016/j.snb.2016.05.089).

2 Our suggestion

We can still use the Silhouette method after making two subtle modifications:

(1) Since we know the exact category and number of our chemicals (i.e., the *k* in Eq. (5) is fixed), we do not use Eq. (5) to search for the optimal *k* again. Instead, we just calculate the mean silhouette value $\tilde{s}\left(\kappa\right)$ with **the fixed *k*.**

$$\xcancel{SC = \max_{\kappa} \tilde{s}\left(\kappa\right)} \qquad (5)$$

(2) Never use K-means as clustering method. Instead, use raw data and the data after LDA to calculate $\tilde{s}\left(\kappa\right)$, respectively.

3. Experimental results

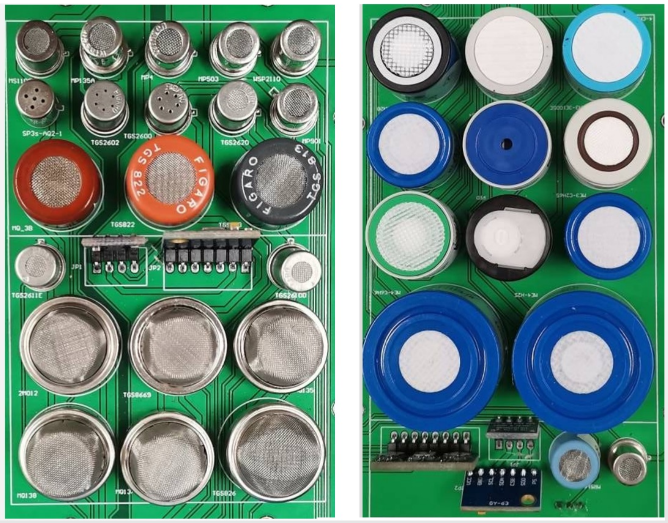
3.1 Experiment 1: Pure chemicals testing

Table1 1 The fourteen gases used in our experiments

	Chemicals	CAS	Category	LD_{50}(mg/kg)
1	n-hexane	110-54-3	Hydrocarbon	25000
2	Acetone	67-64-1	Ketone	5800
3	Acetaldehyde	75-07-0	Aldehyde	1930
4	Formaldehyde	50-00-0	Aldehyde	800
5	Toluene	108-88-3	Benzene	5000
6	Benzol	71-43-2	Benzene	3306
7	n-butylamine	109-73-9	Amine	366
8	Ethanol	64-17-5	Alcohol	7060
9	2-Propanol	67-63-0	Alcohol	5840
10	Tetrahydrofuran	109-99-9	Ether	1650
11	Acetic acid	64-19-7	Hydroxy acid	3300
12	Ethyl acetate	141-78-6	Ester	5620
13	Tetrachloroethene	127-18-4	halogenated hydrocarbon	3005
14	Ammonia	7664-41-7	Inorganic compound	350

Note: LD50 (median lethal dose) is the index which characterize the toxicity of a chemical. Smaller LD50 means stronger toxicity.

The data set consists of 210 samples collected from a self-made e-nose which comprises an array of 37 gas sensors plus temperature, humidity and air pressure sensors. (**Note: 210 samples = 14 gases ×3 concentrations ×5 repetitions**).

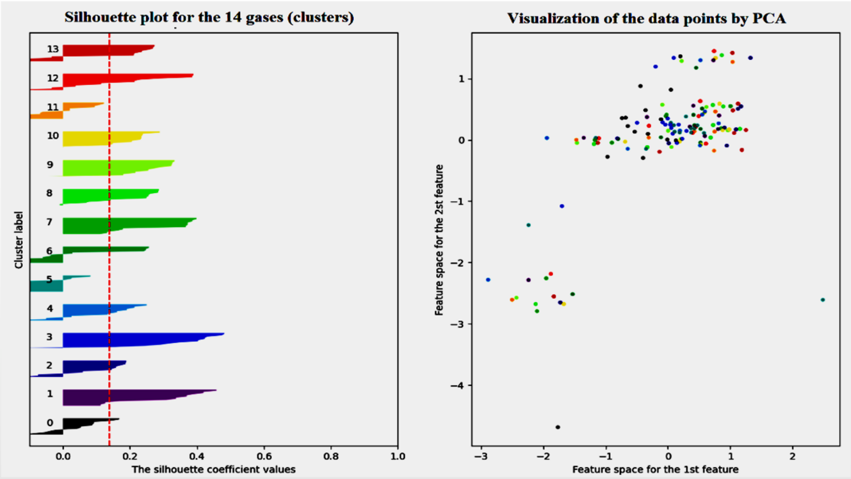


No.	Type Number	No.	Type Number	No.	Type Number
1	MG812	14	MS1100	27	TGS2611E
2	MR516	15	MP135A	28	TGS2610D
3	ME4-H2S	16	MP4	29	2M012
4	ME4-C6H6	17	MP503	30	TGS8669
5	ME3-C2H6S	18	WSP2110	31	MQ135
6	PID-AH	19	SP3-AQ2-1	32	MQ138
7	4S	20	TGS2602	33	MQ137
8	NH3-3E100SE	21	TGS2600	34	TGS826
9	4HS+	22	TGS2620	35	SMD1001
10	ME3-CH2O	23	MP901	36	SMD1007
11	4-CH3SH-10	24	MQ_3B	37	SMD1013B
12	4ETO-10	25	TGS822	38	MS5611-01BA03
13	4OXV	26	TGS813	39	SHT3X

The array of 37 gas sensors plus temperature/humidity and air pressure sensors used in our electronic nose

11

3.1.1 Calculate Silhouette value without clustering (i.e., use raw data only)



$\tilde{s}\left(\kappa\right)$=0.139

Fig. 3 Calculating silhouette value without clustering (left) and the visualization of data points by PCA (right)

12

3.1.2 Calculate Silhouette value after K-means clustering

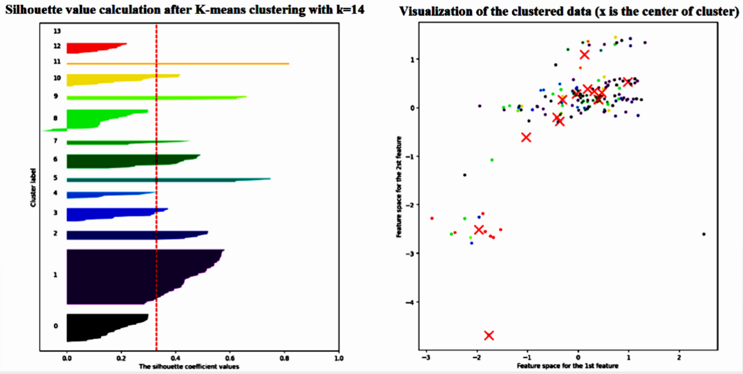


Fig. 4 Calculating silhouette value after K-means clustering (left) and the visualization of data points by PCA (right)

Note: there existed many falsely classified data points by K-means

Table 2 Mean Silhouette values $\tilde{s}\,(\kappa)$ after K-means clustering with different k

k	$\tilde{s}\,(\kappa)$	k	$\tilde{s}\,(\kappa)$
2	0.270	12	0.267
3	0.246	13	0.220
4	0.268	14	0.330
5	0.272	15	0.320
6	0.300	16	0.303
7	0.300	17	0.317
8	0.311	18	0.327
9	0.317	19	0.343
10	0.336	20	0.352
11	0.310		

$SC = \max_{\kappa} \tilde{s}\,(\kappa)$ =0.352 with k=20

$\tilde{s}\,(\kappa)$ = 0.330

for k=14 (true number of gas category/cluster)

3.1.3 Calculate Silhouette value after LDA

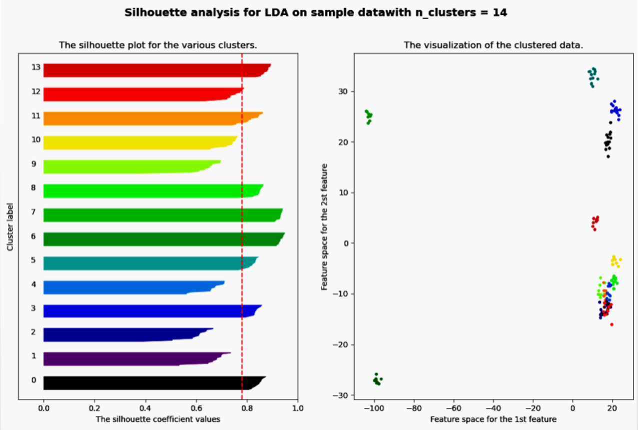


Fig. 5 Silhouette after LDA (left) and the visualized data points (right)

LDA is used for **dimensionality reduction**, not for classification (Because all information of the chemicals are known in our case, see Table 1)

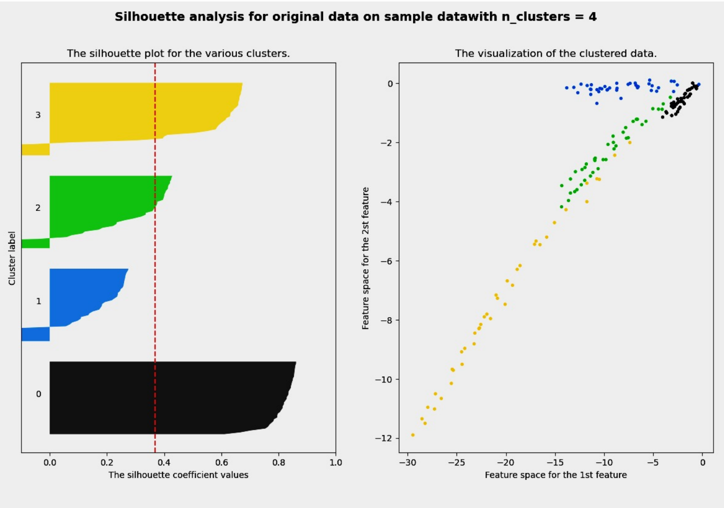
Mean Silhouette value $\tilde{s}\,(\kappa)$=0.787

for k=14 (True number of category/cluster)

Since all data are labeled, so there is no need to consider the classification accuracy even if there are overlaps among clusters.

3.2 Experiment 2 Public data in UCI database

3.2.1 Calculate silhouette value without K-means clustering（raw data)


Silhouette analysis for original data on sample datawith n_clusters = 4

The silhouette plot for the various clusters.

The visualization of the clustered data.

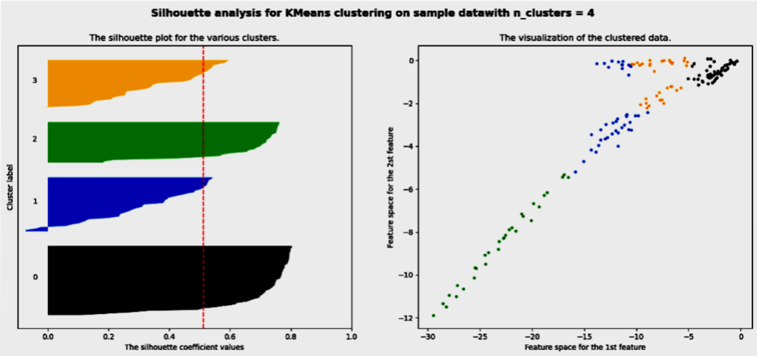
$\tilde{s}\left(\kappa\right)$=0.368

for k=4 (True number of clusters)

Fig. 6 Silhouette (left) and visualization of the clustered data by PCA (right)

Data from UCI database: http://archive.ics.uci.edu/ml/datasets/Twin+gas+sensor+arrays; (Cf. https://doi.org/10.1016/j.snb.2016.05.089).

3.2.2 Calculate silhouette value after K-means clustering


Silhouette analysis for KMeans clustering on sample datawith n_clusters = 4

The silhouette plot for the various clusters.

The visualization of the clustered data.

Fig. 7 Silhouette (left) and visualization of the clustered data by PCA (right)

Data are from UCI database: http://archive.ics.uci.edu/ml/datasets/Twin+gas+sensor+arrays; (Cf. https://doi.org/10.1016/j.snb.2016.05.089).

$$SC = \max_{\kappa} \tilde{s}\left(\kappa\right) \qquad (5)$$

Falsely optimal number of clustering k=3 （True k=4）

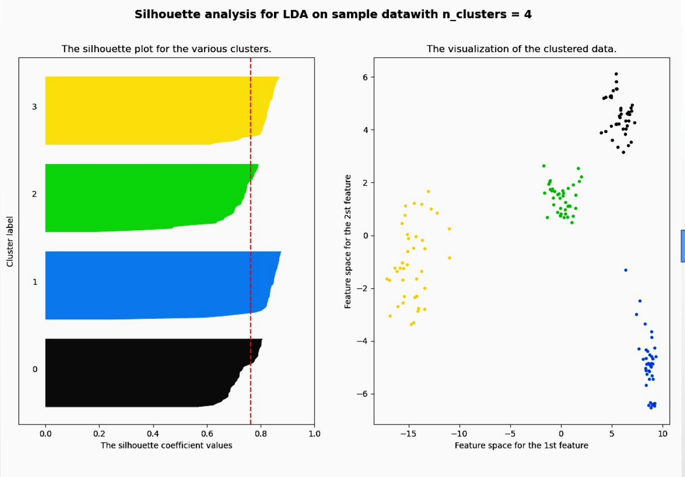
↓

The Silhouette coefficient SC after K-means is **0.607** with **k=3**, while $\tilde{s}\left(\kappa\right)$ =**0.513** for **k=4**.

↓

Silhouette coefficient calculating after the clustering of K-means cannot correctly characterize the real differentiating ability of the instrument due to the existence of misclassification

3.2.3 Calculate silhouette value after LDA



$\tilde{s}\,(\kappa)$=0.763

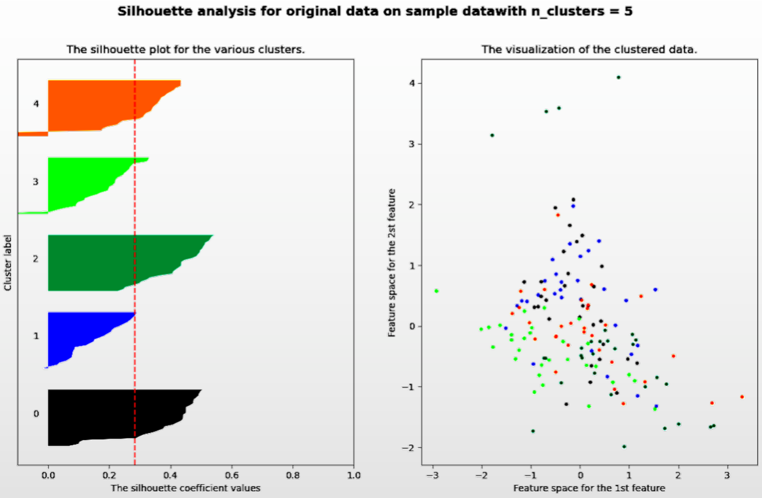
for *k*=4 (True number of clusters)

Fig. 8 Silhouette (left) and visualization of the clustered data points (right)

Data are from UCI database: http://archive.ics.uci.edu/ml/datasets/Twin+gas+sensor+arrays; (Cf. https://doi.org/10.1016/j.snb.2016.05.089).

3.3 Experiment 3: Chinese medicine sorting

3.3.1 Calculate silhouette value without K-means clustering



The data set consist of 150 samples collected from the same e-nose above. There are 5 kinds of Chinese medicine. (150 samples=5 kinds ×30 repetitions)

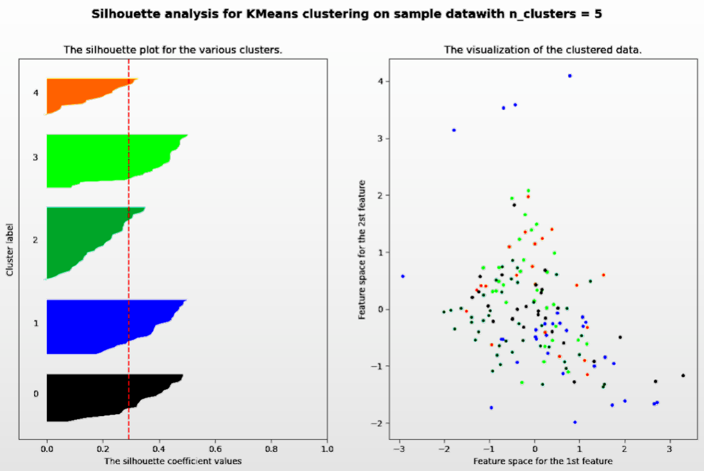
$\tilde{s}\,(\kappa)$=0.284

for *k*=5 (True number of clusters)

Note:
The number of points in each cluster is equal Because each kind has 30 samples.

Fig. 9 Calculating silhouette value without clustering (left) and the visualization of data points by PCA (right)

3.3.2 Calculate silhouette value after K-means clustering



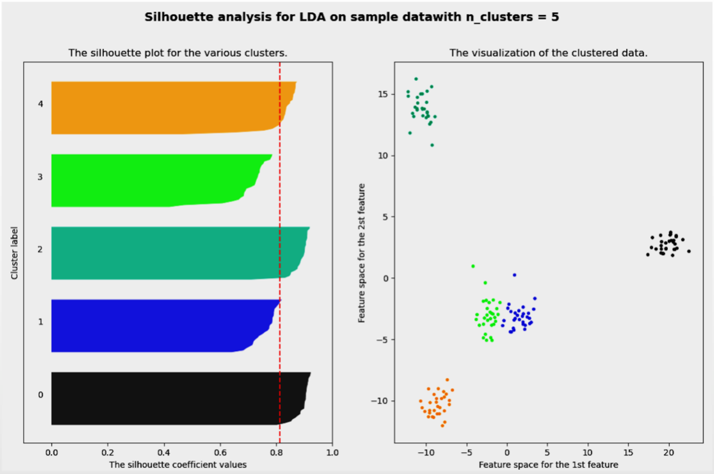
$\tilde{s}\,(\kappa)$=0.293

for *k*=5 (True number of clusters)

Note: There are many falsely classified data points by K-means, so it can be found that the number of points in each cluster is not equal.

Fig. 10 Silhouette values calculated after K-means clustering (left) and the visualized data points (right)

3.3.3 Calculate silhouette value after LDA



$\tilde{s}\,(\kappa)$=0.813

for *k*=5 (True number of clusters)

No falsely clustering problem

Fig. 11 Silhouette values calculated after LDA

4 Conclusion

To calculate Silhouette Coefficient after K-means has the problems :

(1) falsely clustering,

(2) need two variables (accuracy of classification, **Silhouette coefficient, *SC*)** to characterize an instrument.

We can use only one variable, the **mean Silhouette Value** $\tilde{s}(\kappa)$ to replace the ***SC*** in our case. To calculate the $\tilde{s}(\kappa)$ without K-means with either of the following methods:

Use the raw data only: the $\tilde{s}(\kappa)$ is small in many cases

LDA: a commonly used supervised method with proper $\tilde{s}(\kappa)$

Advantage：Since we know the exact information of all the clusters, it makes full fusion of the known information of labels (category of Chemicals) and the number of clusters (each chemical is a cluster) . So we can use only one value, i.e., the $\tilde{s}(\kappa)$ to reflect the ability of an instrument in differentiating chemicals.

We also suggest that, don't use PCA before Silhouette calculation, because although in most cases, the results are fine, but in a few cases, we get bad results (See Table A3 in next page). It might be due to the lost of information during dimensionality reduction. So we'd better only use raw data (or after LDA) to calculate the mean Silhouette values.

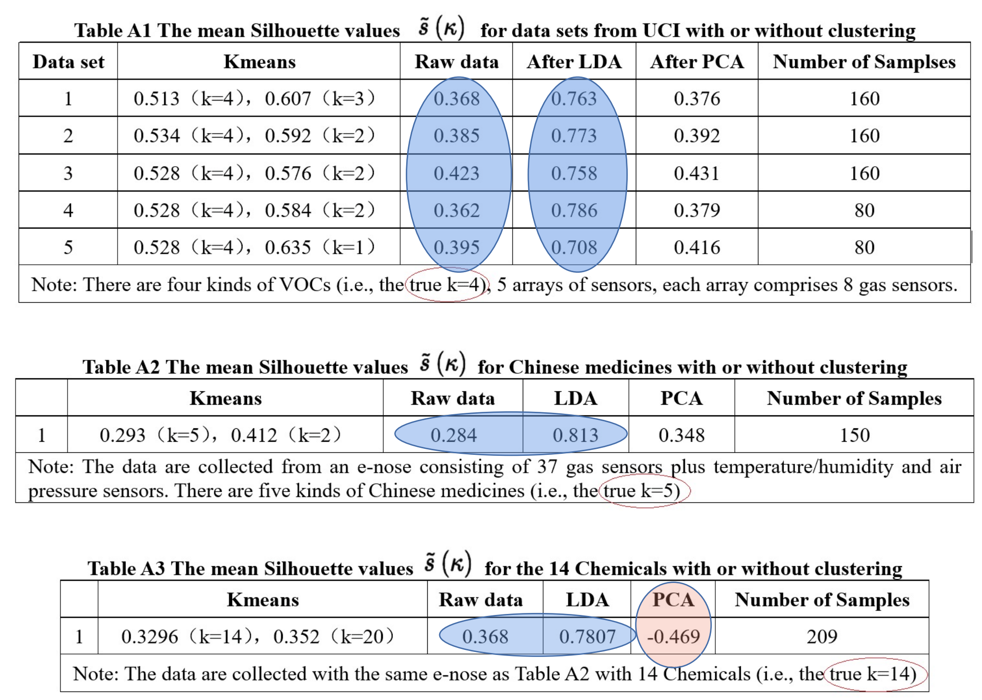


Table A1 The mean Silhouette values $\tilde{s}(\kappa)$ for data sets from UCI with or without clustering

Data set	Kmeans	Raw data	After LDA	After PCA	Number of Samplses
1	0.513（k=4），0.607（k=3）	0.368	0.763	0.376	160
2	0.534（k=4），0.592（k=2）	0.385	0.773	0.392	160
3	0.528（k=4），0.576（k=2）	0.423	0.758	0.431	160
4	0.528（k=4），0.584（k=2）	0.362	0.786	0.379	80
5	0.528（k=4），0.635（k=1）	0.395	0.708	0.416	80

Note: There are four kinds of VOCs (i.e., the true k=4), 5 arrays of sensors, each array comprises 8 gas sensors.

Table A2 The mean Silhouette values $\tilde{s}(\kappa)$ for Chinese medicines with or without clustering

	Kmeans	Raw data	LDA	PCA	Number of Samples
1	0.293（k=5），0.412（k=2）	0.284	0.813	0.348	150

Note: The data are collected from an e-nose consisting of 37 gas sensors plus temperature/humidity and air pressure sensors. There are five kinds of Chinese medicines (i.e., the true k=5)

Table A3 The mean Silhouette values $\tilde{s}(\kappa)$ for the 14 Chemicals with or without clustering

	Kmeans	Raw data	LDA	PCA	Number of Samples
1	0.3296（k=14），0.352（k=20）	0.368	0.7807	-0.469	209

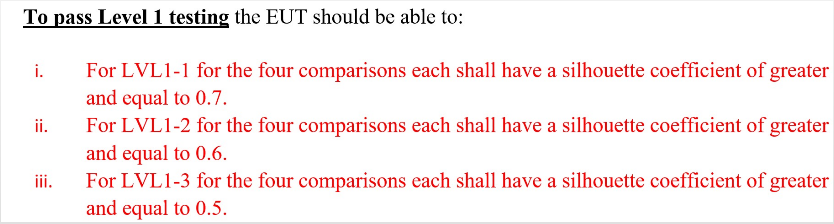
Note: The data are collected with the same e-nose as Table A2 with 14 Chemicals (i.e., the true k=14)

Question : Can we use LDA?

(1) Because the mean Silhouette values $\tilde{s}\,(\kappa)$ calculated with raw data are very low not only in many of our self-made e-nose cases, but also in some public database data (See the page above).

(2) LDA is a mature, linear, simple and commonly used algorithm either for dimensionality reduction or for classification. It is relatively objective.

(3) For those machine with low $\tilde{s}\,(\kappa)$ before LDA but high $\tilde{s}\,(\kappa)$ after LDA, the machine is still useful in many cases. If we just calculate $\tilde{s}\,(\kappa)$ without LDA , then a large number of e-noses might be failed to pass our standard. Besides, if we don't use LDA, maybe we have to modify the threshold (I don't know how much the $\tilde{s}\,(\kappa)$ will be for most e-noses, e.g., for Alpha MOS or PEN 3), and how we know the threshold of 0.7, 0.6 and 0.5 is ok (See below).



<u>To pass Level 1 testing</u> the EUT should be able to:

i. For LVL1-1 for the four comparisons each shall have a silhouette coefficient of greater and equal to 0.7.

ii. For LVL1-2 for the four comparisons each shall have a silhouette coefficient of greater and equal to 0.6.

iii. For LVL1-3 for the four comparisons each shall have a silhouette coefficient of greater and equal to 0.5.

(Page 11 of P2520.1 Rev 14)

23

Thanks!

24