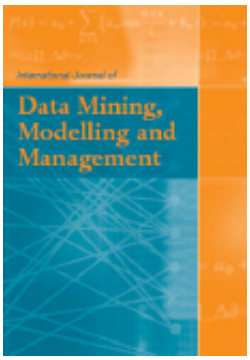


International Journal of Data Mining, Modelling and Management

ISSN online: 1759-1171 - ISSN print: 1759-1163
https://www.inderscience.com/ijdmmm

K-means and DBSCAN for look-alike sound-alike medicines issue

Souad Moufok, Anas Mouattah, Khalid Hachemi

DOI: 10.1504/IJDMMM.2024.10057242

Article History:

Received:	21 February 2022
Last revised:	07 February 2023
Accepted:	26 March 2023
Published online:	22 January 2024

K-means and DBSCAN for look-alike sound-alike medicines issue

Souad Moufok, Anas Mouattah* and Khalid Hachemi

Institute of Maintenance and Industrial Safety,
University of Oran 2 Mohamed Ben Ahmed,
B.P 1015 El M'naouer 31000 Oran, Algeria
Email: moufok.souad@univ-oran2.dz
Email: mouattah.anas@univ-oran2.dz
Email: hachemi.khalid@univ-oran2.dz
*Corresponding author

Abstract: The goal of this study is to analyse the application of data mining techniques in clustering drug names based on their spelling similarity in order to reduce the occurrence of dispensing errors caused by look-alike sound-alike medicine confusion, as they considered one of the most common causes of dispensing errors. Two unsupervised data mining methods, k-means and DBSCAN, were used in conjunction with two similarity measures, BiSim and Levenshtein. The results of the study showed that the approach is effective in identifying potential confusable medicines, with BiSim-based k-means clustering being favoured with a silhouette score of 0.5.

Keywords: look-alike sound-alike; LASA; data mining; medication errors; dispensing errors; k-means; DBSCAN.

Reference to this paper should be made as follows: Moufok, S., Mouattah, A. and Hachemi, K. (2024) 'K-means and DBSCAN for look-alike sound-alike medicines issue', *Int. J. Data Mining, Modelling and Management*, Vol. 16, No. 1, pp.49–65.

Biographical notes: Souad Moufok is currently an Assistant Professor with the Institute of Maintenance and Industrial Safety, University of Oran 2 Mohammed Ben Ahmed, Algeria. Her research interests include patient safety, data mining and informatics.

Anas Mouattah is currently pursuing his PhD in Instrumentation at the Institute of Maintenance and Industrial Safety, University of Oran 2 Mohammed Ben Ahmed, Algeria. His research interests include patient safety, artificial intelligence, and embedded systems.

Khalid Hachemi is currently a Professor with the Institute of Maintenance and Industrial Safety, University of Oran 2 Mohammed Ben Ahmed, Algeria. He has authored or co-authored number of scientific articles. His research interests include patient safety, automation and industrial engineering.

1 Introduction

Medication dispensing error refers to a preventable difference between the patient's actual medication and the prescribed medication. This can take various forms, such as dispensing the wrong medicine, an incorrect dose form or concentration, or dispensing at the wrong time. This has become a major concern due to the potential fatal consequences, particularly for critically ill patients. The causes of these errors have been attributed to various factors, including being short-staffed or pressed for time, excessive workload, fatigue, disruptions during delivery, and issues with look-alike sound-alike (LASA) medications.

Studies have found that LASA issues are the second most common cause of dispensing errors, after overloads in command. These errors occur when similar medications have similar names, different medicines with the same name but different brands, abbreviated names or nicknames, or unclear verbal medication orders.

Data mining, which involves analysing large databases to generate useful information, it has been deployed in the healthcare sector to improve patient care, reduce healthcare costs, and enhance decision-making processes. It involves the use of statistical and machine learning techniques to analyse large and complex datasets generated by electronic health records, claims data, and other sources. Its usage has transcended to predict patients at risk of chronic conditions or readmission. Some physicians have managed to exploit it in clinical decision support systems to assist in diagnosis, treatment planning and for personalised medicine to tailor treatment plans based on a patient's specific genetic and medical history. It has also been used to address dispensing errors and improve patient safety. Some studies have proposed a classification model that considers three factors – medication name, characteristics, and environment – to address the LASA issue.

In this study, the authors aim to improve patient safety by addressing the LASA issue using data mining methods, specifically k-means and density-based spatial clustering of application with noise (DBSCAN) clustering combined with BiSim and Levenshtein similarity measures. The goal is to identify confusable medication names that could cause ambiguity for the dispenser.

The study is structured as follows: in Section 3, the clustering methods and similarity measures are explained. In Section 4, the results of cross-testing the clustering methods using both similarity measures are presented. Finally, a conclusion summarises the key points of the study.

2 Related work

In order to overcome the LASA issue, several approaches are suggested, either addressing all the forms of dispensing errors like the healthcare enabled barcode, RFID and automatic dispensing cabinets, or dedicated particularly to the current issue (Gates et al., 2021; Mouattah and Hachemi, 2021; Mulac et al., 2021).

Levenshtein distance and Bigram similarity algorithms are considered as kernels for studies aimed at identifying and highlighting medicines names 'pairs that might result in confusion, this can even be extended to new medicines names against those already in existence (Alsaeedi, 2020; Ding et al., 2021; Joachims, 1999; Lambert, 1997; Levenshtein et al., 1966; Millán-Hernández et al., 2019).

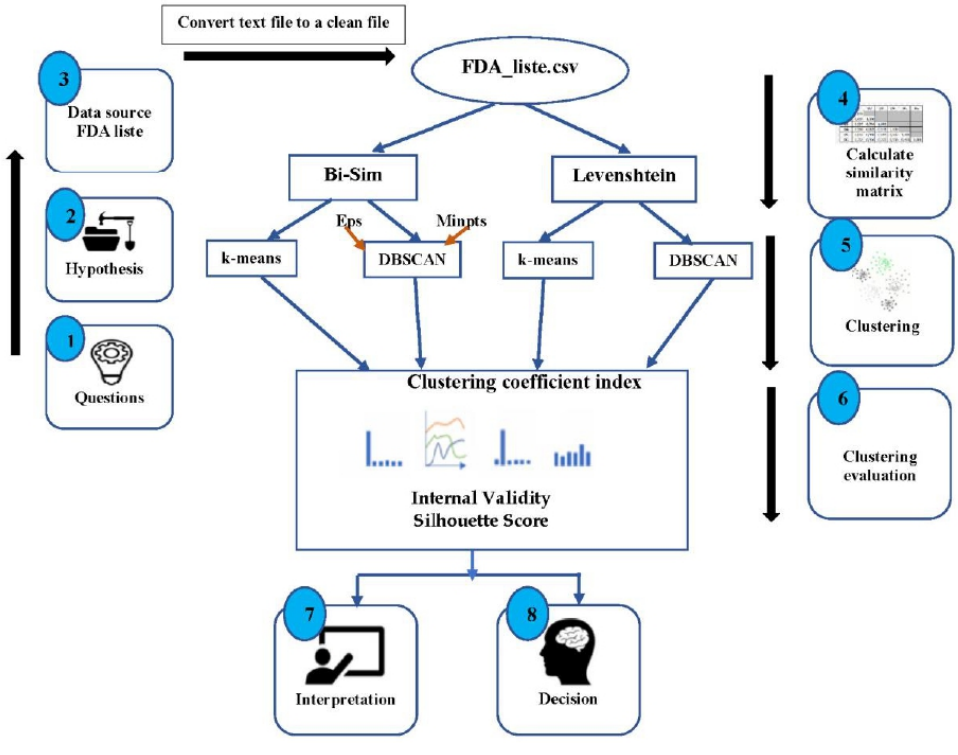
To address the visual side (look-alike) of the name writing, the 'tall man lettering' approach has been proposed, where the medicines names are suggested to be written in lowercase; only the part of the text which is critical is highlighted using the uppercase (e.g. cycloSERINE and cycloSPORINE). This solution brought an improvement in reducing the visual based confusion of medicines names that may lead to the dispensing error, and it has big approbation among healthcare actors (Filik et al., 2010).

On the other hand, approaches are suggesting identifying similar medicines names based on the phonetic characteristics of each letter. But this necessitates the phonetic transcription of the names, but it permits highlighting confusable medicines names that cannot be detected using the visual methods, in addition to blurred verbal orders that may cause misperception (Hadwan et al., 2021; Wang et al., 2015).

In the absence of experimental studies on the LASA effect, which are considered decisive, the approaches addressing both the visual and acoustic sides of the issue using combination of orthographic and phonetic methods remain the highest standard in highlighting critical medicines names (Lambert et al., 1999; Millán-Hernández et al., 2016).

Another proposed solution is 'trade dress', where the idea is to link every medicine with one or more of its semblance characteristics; including colour, shape …, etc. But, as much as this seems distinguishing, it cannot be proved to be efficient especially with the amount of new medicines entering the market continuously (Ciociano and Bagnasco, 2014; Sim and Robertson, 2008).

Figure 1 Research methodology block-diagram (see online version for colours)



Alternative approaches addressed the issue using hybrid data mining approach, where the classification model is based on logistic regression, retrieving a decision tree from the cases of dispensing errors and a medical database, this permits the detection of the confusable drugs with precisions attending 83%. Other strategies suggest the detection of phonetic and orthographic similarity (look-alike and sound-alike) between pairs of drug names using genetic algorithms, and a set of similarity measures, (like Editext, NED, LCSR …, etc.) (Chen et al., 2011; McLachlan et al., 2020; Tseng et al., 2007; Wang et al., 2015; Watterson et al., 2021).

Taking in consideration spacing between confusable medicines, in this paper, we suggest applying a combination k-means and DBSCAN on both BiSim and Lewenstein resultant spacing matrices, in addition, we study the feasibility of exploiting the resultant number of clusters using DBSCAN as a starting point of splitting in the k-means algorithm and vice versa. The following block diagram, Figure 1 displays the flow of this study.

3 Material and method

Before applying the unsupervised clustering algorithms, a similarity distance intra pairs is put in place. In this section, we explain the obtainment of the similarity matrices using Levenshtein and BiSim, the application of k-means and DBSCAN on the spacing matrices along with our approach in combining them.

In studying the LASA medicines issue, it is inevitable to pass by the Food and Drug Administration list of confusable medicines; which we referred to it by the FDA list in this paper, as it constitutes a reference for all the papers treating the same topic. In our work, we applied the clustering algorithms on the FDA list, and we discussed the results accordingly.

3.1 Similarity matrix

We have the similarity between two drugs $S_{ab} = S_{ba}$ and each drug name is checked against the other names and not itself. Hence, the attended is a square matrix of $n \times n$ where n is the number of drugs.

3.1.1 Levenshtein distance

Levenshtein distance was first suggested by Levenshtein et al. (1966), it is devoted to compare the similarity between two strings with regard to characters' order. Using two strings as input parameters, Levenshtein's algorithm calculates the minimum amount of deletion, insertion and substitution required to transform a string into another string.

The pseudo code of the Levenshtein distance algorithm is shown in Algorithm 1.

Algorithm 1 Levenshtein distance algorithm

- **Input:**
  $D = \{T_1, T_2, \ldots, T_L\}$ is the data set where each T_i represents a drug name.
  $T_i = \{U_1, U_2, \ldots, U_R\}$ is a character of each drug name.
- **Output:** $M = D \cdot D$ is a distance matrix.

```
1   String x (size n)
2   String y (size m)
3   Create an empty matrix 'M' of size (n + 1) × (m + 1), where the row and column headers
    correspond to characters of string x and y, respectively.
4   if string x.length = 0 then return string y.length
5   if string y.length = 0 then return string x.length
6   for each i = 1 to n
7   M(i + 1, 1) = i
8   end of i
9   for j = 1 to m do
10  M(1, j + 1) = 0
11  end of j
12  for each character i of x to n do
13  for each character j of y to n do
14  if x(i) = y(j) then cost = 0
15  if x(i) <> y(j) then cost = 1
16  M(i + 1, j + 1) = minimum (
    M(i – 1, j) + 1                 //deletion
    M(I, j – 1) + 1                 //insertion
    M(i – 1, j – 1) + cost)         //substitution
17  end of j
18  end of i
19  return M(n, m)
```

3.1.2 BiSIM similarity

The orthographic similarity measure BiSim proposed by Dagan et al. (1994) belongs to the family of n-gram where $n = 2$, measures and calculation of the number of common bigrams between two words is by adding single symbol at the beginning of each word which increases the important of the correspondence of the initial letter with a cross link absence contain that guarantees the sequentially of letter matches. This measure calculates the similarity between two strings by dividing the result of the similarity Nsim by the length of the longest string.

Given the drug names x and y as sequences of size n and m, respectively. Each character of the two strings x and y is represented by i and j, respectively.

Nsim is defined as:

$$\text{Nsim}(i, j) = \max(M(i-1, j), M(i, j-1), M(i-1, j-1) + S\left(x_i x_{i-1}, y_j y_{j-1}\right) \tag{1}$$

where:

$$S\left(x_1, x_2, ..., x_n, y_1, y_2, ..., y_m\right) = \begin{cases} \frac{1}{2} & \text{if } x_i = y_j \\ 0 & \text{if } x_i \neq y_j \end{cases} \tag{2}$$

The pseudo code of BiSim algorithm is shown in Algorithm 2.

Algorithm 2 BiSim similarity algorithm

- **Input:**
  $D = \{T_1, T_2, \ldots, T_L\}$ is the data set where each T_i represents a drug name.
  $T_i = \{U_1, U_2, \ldots, U_R\}$ is a character of each drug name.
- **Output:** $M = D \cdot D$ is a similarity matrix.

```
String x (size n)
String y (size m)
Create an empty matrix 'M' of size (n + 1) × (m + 1), where the row and column headers correspond to characters of string x and y, respectively.
for i = 1 to n do
M(i, 1) = 0        //Symbol added at the beginning of the string x
end of i
for j = 1 to m do
M(1, j) = 0        //Symbol added at the beginning of the string y
end of j
for each character i of x to n do
for each character j of y to n do
M(i, j) = Nsim(i, j)
end of j
end of i
return M(n, m)/max(n, m)
```

3.2 Clustering algorithms

The matrix of similarity and distance of all intra-pair of drug names are used as data input for clustering algorithms. In order to classify drug names into similar groups, two main techniques are exploited, k-means and DBSCAN, with two different distance measures.

The main idea of the present work is to calculate DBSCAN clustering on the data, by retrieving the number of clusters found that represents the value of K provided, then use this value as a starting point of splitting in the k-means algorithm.

3.2.1 DBSCAN clustering

The execution of DBSCAN algorithm has two stages:

- Stage 1: determining the optimal epsilon (Eps) value.
- Stage 2: application of DBSCAN algorithm using the optimal Eps value.

3.2.1.1 Stage 1: determining the optimal value of Eps

The execution of the DBSCAN algorithm requires the Epsilon (Eps) parameter, for this, we use the DMDBSCAN method (Han et al., 2012) to determine the Eps value

automatically based on the k-nearest neighbours (K-NN) method, by calculating the distance of each point with the n nearest neighbouring points.

3.2.1.2 Stage 2: DBSCAN algorithm

DBSCAN is a dataset clustering algorithm proposed by Ester et al. (1996) where the objective of the process is to identify a dense region, which can be measured by the number of objects close to a given point in order to assign each point of a dataset to a cluster, or to consider it as a noise.

Two main parameters are required for DBSCAN: epsilon ('Eps') and minimum points ('MinPts'), where Eps defines the radius of the neighbourhood region, called the neighbourhood of x, and MinPts represents the minimum number of neighbours in Eps.

Each point x in the dataset can be considered as:

- *Core point:* if the point P with a neighbour count greater than or equal to MinPts.
- *Border point:* if the number of its neighbours is less than MinPts, but it belongs to the Eps neighbours of some core point.
- *Noise point:* is any point that is not a core point or a border point.

The pseudo code of the DBSCAN algorithm is shown in Algorithm 3.

Algorithm 3 DBSCAN algorithm

-	**Input:**
	D = dataset.
	Eps = radius.
	MinPts = minimum number of neighbourhoods.
-	**Output:** K = clusters.
1	Calculate the similarity matrix using Levenshtein distance or BiSim similarity measure.
2	$K = 0$.
3	**for** each unvisited point P of dataset D **do**
4	Mark P as visited
5	**if** (size of $N <$ MinPts) **then**
6	Mark P as Noise
7	**end if**
8	**if** (size of N >= MinPts) **then**
9	Mark P as a core point
10	Add point P to cluster K
11	**end if**
12	**for** each point P in N, repeat steps 2 to 11
13	**end for**
14	Return K.

3.2.2 K-means clustering

K-means clustering is one of the most widely used data mining methods, its objective is summed up in dividing a dataset into k clusters, where k is considered as an input variable determined by the user and may be different from the optimal number of clusters. To solve this problem, in this work, we used the result of the number of clusters obtained by applying the DBSCAN algorithm as the number of inputs of the averaging algorithm k-means.

The pseudo code of the k-means algorithm is shown in Algorithm 4.

Algorithm 4 K-means algorithm

-	**Input:**
	Number of clusters K.
	Similarity matrix or Distance matrix.
-	**Output:** C = clustering result $\{C1, C2, \ldots, Ck\}$.
1	Choose randomly K points (one line of the data matrix), these points are the centres of the clusters (called controid)
2	Repeat.
3	Assign each point (element of the data matrix) to the group whose centre is nearest to it.
4	Recalculate the centre of each cluster and modify the centroid.
5	Until convergence of the algorithm to a stable partition.

In the unsupervised ML, the silhouette score is the reference regarding the precision of the clustering, it has a value in [–1; 1], where as much as this value gets close to 1, it means that the medicines in the same cluster are close to each other, and less match to the other clusters at the same time in the matter of semblance. The opposite is true when the silhouette value gets close to –1.

4 Results and discussion

In this section, we reveal and discuss the results of applying the above methods on the list of medicines that might sound ambiguous for the healthcare actors and threatening.

4.1 Clustering using DBSCAN

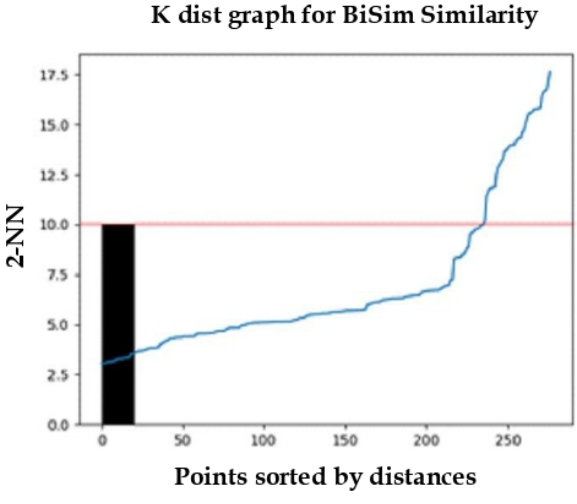
4.1.1 Optimal Eps and Minpts

The obtained is curved graph, where the optimal value for Eps is at the graph maximum. The illustrated results in Figure 2 show the optimal values of Eps using the nearest neighbours approach; which allows defining the closest neighbours of each expression as well as the distances, where $k = 2$ represents the value of MinPts. In Table 1, the variation of silhouette score of DBSCAN clustering on corresponds of the variation of Minpts.

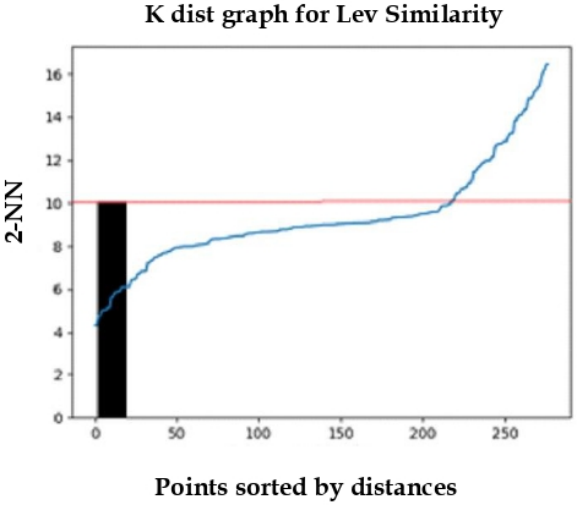
Table 1 The variation of sihouette score of DBSCAN clustering on corresponds with the variation of Minpts

			BiSim			
Minpts	2		3	4	5	
Silhouette	0.45		–0.058	–0.065	–0.004	
			Levenschtein			
Minpts	2	3	4	5	6	7
Silhouette	0.394	0.259	0.255	0.371	0.351	0.356

Figure 2 K-dist graph for drug-names for both (a) BiSim, (b) Levenshtein (see online version for colours)



(a)



(b)

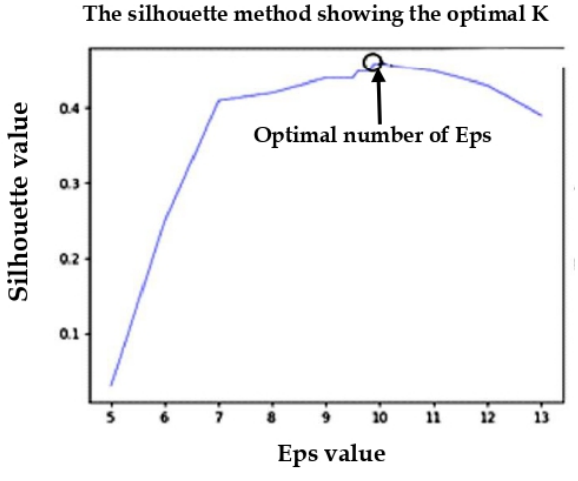
As known, the choice of Minpts depends on the dimension of the data where it has to be equal or greater to it. In this case, as we are clustering spacing measures which are one scale values, we have k = MinPts = 2 which correspond to the formula suggested by Sander et al. (1998) in equation (3):

$$\text{MinPts} = 2 * \text{dimension of the data} \tag{3}$$

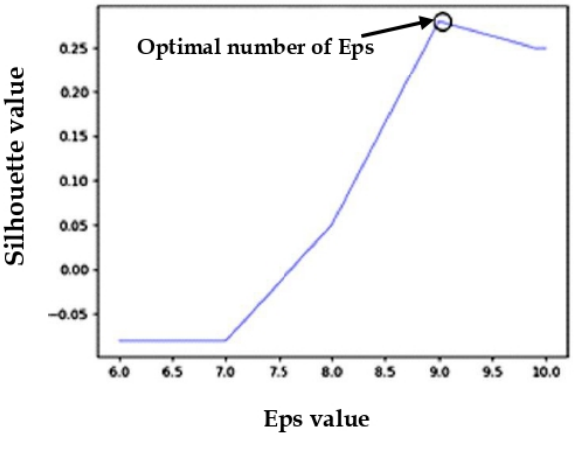
$$\text{MinPts} = 1 * 2 = 2 \tag{4}$$

Using the BiSim and Levenshtein similarity matrices as input data, the large variation in the graph is respectively equal to 10 [Figures 2(a) and 2(b)], whilst the red line and black bar are used to roughly identify how abruptly the graph has changed.

Figure 3 Optimal values of Eps per silhouette coefficient with both measures; (a) BiSim, (b) Levenshtein (see online version for colours)



(a)

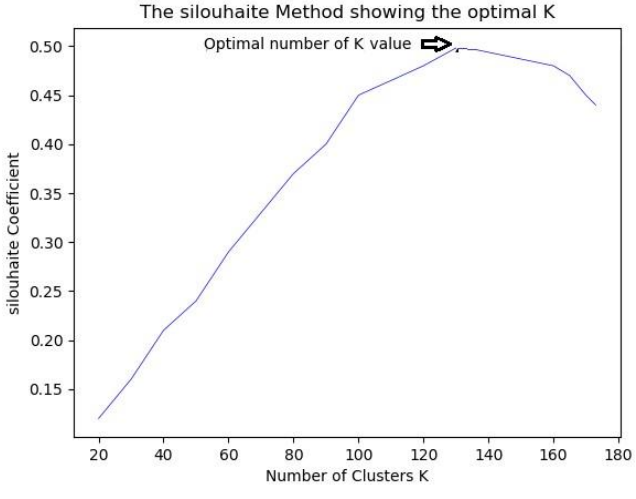


(b)

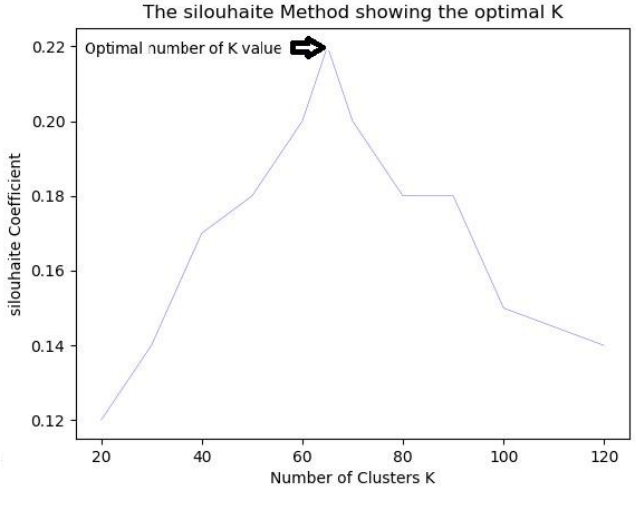
4.1.2 Clustering

Attained results using the DBSCAN clustering algorithm; based on BiSim measure are 105 drug groups with a silhouette coefficient equal to 0.44. On the other hand, 95 drug clusters were obtained with a silhouette coefficient equal to 0.25 based on Levenshtein similarity measure.

Figure 4 The variation of the value of the number of clusters K according to the silhouette coefficient score using the (a) BiSim, (b) Levenshtein similarity measure (see online version for colours)



(a)



(b)

4.2 Clustering using k-means

4.2.1 Optimal number of clusters

Knowing that clustering using the k-means algorithm requires an optimal number of clusters K, we used the silhouette coefficient, which allows evaluating the clustering performance. The latter is calculated for each of the two similarity measures BiSim and Levenshtein. The closer the coefficient value is to 1, the better the clustering result is.

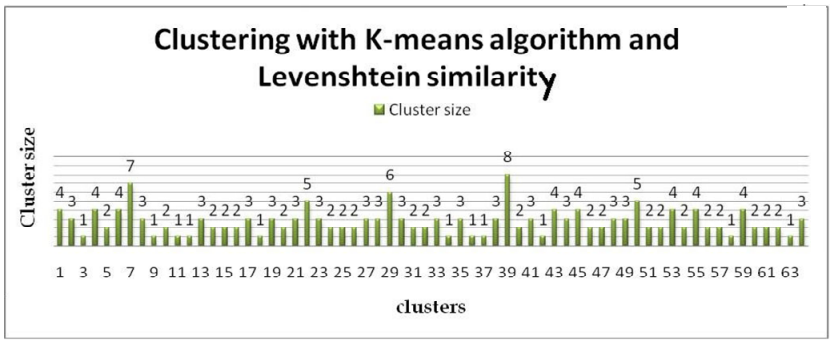
Figure 4 shows that the optimal number of clusters is equal to 131 with a silhouette score equal to 0.50 using the BiSim similarity measure, whereas the optimal number of clusters equals 65 with a silhouette score equal to 0.22 using the lev similarity measure.

4.2.2 Clustering

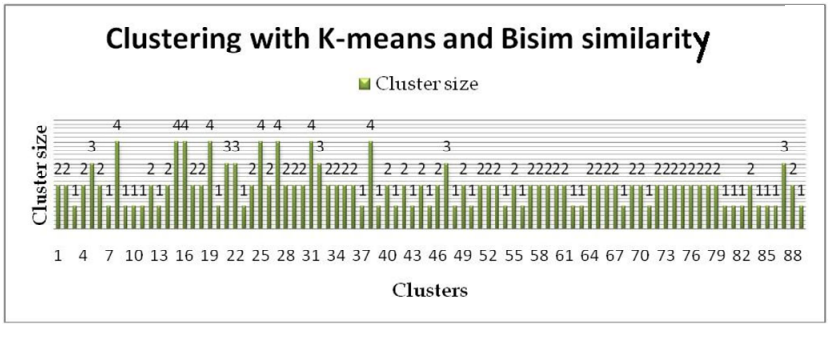
Using the k-means clustering algorithm, we obtained the clusters shown in Figure 5 in the form of a histogram. The number of bars represents the number of clusters found, where each of the bars represents the drug set judged to be similar based on orthographic similarity.

In terms of precision, the k-means clustering algorithm using the BiSim similarity measure has better performance than the DBSCAN algorithm with a silhouette score equal to 0.5 compared to the DBSCAN algorithm which has a silhouette score equal to 0.44. Figure 6 shows the obtained results of comparison in terms of precision.

Figure 5 Clustering results using k-means algorithm, (a) clustering using Levenshtein similarity, (b) clustering using BiSim similarity (see online version for colours)



(a)



(b)

Noted that for both of the clustering algorithms k-means and DBSCAN, BiSim has performed better than Levenstein for the same algorithm, however, for a deaf comparison, we see that Levenstein worked better with DBSCAN over k-means contrariwise BiSim with k-means. This can be traced to the nature of the similarity tests them self, as Levenshtein measures the number of edits required to transform one string into another, while BiSim considers the presence of bi-grams (consecutive characters) in measuring the similarity, which means, passing from the first word to the second is faster using Levenstein. This is translated in the low optimal number of clusters obtained using Levenstein compared to it using BiSim (95% 105 for DBSCAN and 65% 130 for k-means).

The reason behind the leverage of BiSim based k-means compared to DBSCAN algorithm is that the latter considers a large number of drug names as Noise. This is linked to dimension of the data, as the obtained spacing measures are single scale while DBSCAN is conceived for multi-dimensional clustering. Moreover, combining the two clustering algorithms by using the number of clusters obtained using DBSCAN as a starting point for splitting in the k-means does not have the best performance overall, still, it is in the best quarter in terms of silhouette score and it's performed better than DBSCAN alone (Figure 3). This means that we can expect the best performance of k-means in clustering by at most 25% of error if we already have the data of DBSCAN.

Table 2(a) represents a sample of the treated medicines clustering, according to the orthographic similarity of a set of drug names extracted from the FDA list. Table 2(b) shows the obtained grouping using the BiSim based k-means clustering approach.

Table 2(a) A sample of medicine names cited in the FDA list

1	sulfaSALAzine	14	sulfiSOXAZOLE	27	NovoLIN
2	DAPTOmycin	15	eriBULin	28	CeleXA
3	hydrALAZINE	16	NexAVAR	29	epiRUBicin
4	romiDEPsin	17	hydrOXYzine	30	levOCARNitine
5	HYDROmorphone	18	oxyCODONE	31	CeleBREX
6	levoFLOXacin	19	hydroCHLOROthiazide	32	romiPLOStim
7	NexIUM	20	idaruCIZUmab	33	HYDROcodone
8	penicillAMINE	21	SUFentanil	34	SORAfenib
9	fentaNYL	22	valACYclovir	35	DACTINomycin
10	predniSONE	23	acetoHEXAMIDE	36	methazolAMIDE
11	NIFEdipine	24	HumuLIN	37	acetaZOLAMIDE
12	oxyMORphone	25	chlordiazePOXIDE	38	methylPREDNISolone
13	valGANciclovir	26	penicillin	39	HumaLOG

By considering the FDA list as a reference for medicines clustering, what is seen is, except for the 2nd, 6th, 9th and 12th cluster, all of the pairs in the FDA list are included in the same clusters.

The presence of the 'intruders' in the 3rd, 4th, 7th and the 11th (prednisone, fentanyl, chlordiazePOXIDE and idaruCIZUmab) is linked to the absence of their partner (of the FDA list) in the sample, which led to their integration in the closest clusters in term of semblance, or even forming new clusters; which is the case for the 2nd the 9th and 12th clusters. While for the 6th cluster the situation is different, there is a divergence between

the pair resultant of the followed approach and those of FDA list, as HYDROcodone is paired with oxyCODONE and oxyMORphone is paired with HYDROmorphone in the FDA list despite the presence of the pair (oxyCODONE, oxyMORphone) in both of them, this can be tracked to the nature of this clustering, since any drug name has belong to only one cluster at a time, while for hydroCODONE and hydroMORPHONE it is understandable, as despite their absence from the FDA list, they pretty much look alike-sound alike grammatically, phonetically and syllabically, we can even say that they are missed from the FDA list.

Table 2(b) Resultant clusters of executing BiSim based k-means clustering approach (the combination) (see online version for colours)

Cluster ID	*Similar drugs names*	*Cluster ID*	*Similar drugs names*
1	acetaZOLAMIDE	9	levOCARNitine
	acetoHEXAMIDE		levoFLOXacin
2	methazolAMIDE	10	CeleBREX
	methylPREDNISolone		CeleXA
	SORAfenib		
	SUFentanil		
3	penicillAMINE	11	epiRUBicin
	penicillin		eriBULin
	predniSONE		fentaNYL
4	chlordiazePOXIDE	12	NexAVAR
	HumaLOG		NexIUM
	HumuLIN		NIFEdipine
			NovoLIN
5	valACYclovir	13	DACTINomycin
	valGANciclovir		DAPTOmycin
6	HYDROcodone	14	romiDEPsin
	HYDROmorphone		romiPLOStim
7	hydrALAZINE	15	sulfaSALAzine
	hydroCHLOROthiazide		sulfiSOXAZOLE
	hydrOXYzine		
	idaruCIZUmab		
8	oxyCODONE	16	guaiFENesin
	oxyMORphone		guanFACINE

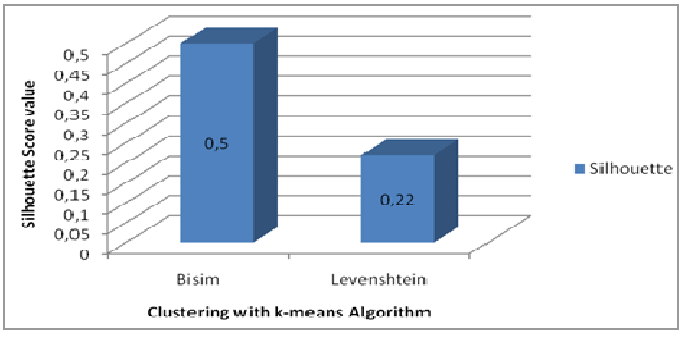
Notes: Yellow – Combination that exists in the FDA list.
Blue – Intruder.
Purple/red – Combination that does not exist in the FDA list.

Looking at the close intra-members semblance of the same clusters, even in the absence of their partners of the FDA list, we can say that clustering using data mining methods is feasible and can even be used in the detection of new confusable drug names.

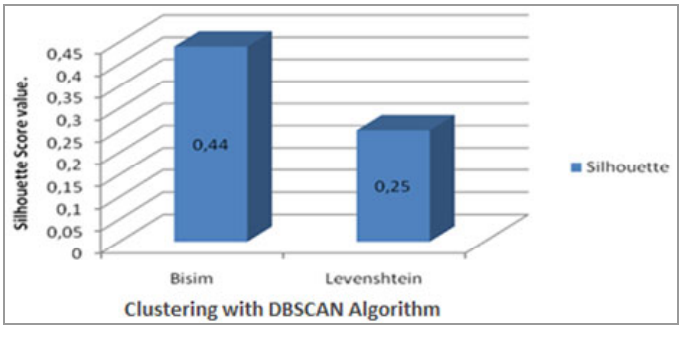
This can be used to reduce medication dispensing errors by taking extra precautions with medicines of the same cluster in drugs classification and positioning, and automatic

dispensing cabinet filling, nevertheless, other technologies can also be included such as bare code or RFID, and this by submitting critical/confusable medicines to special management strategies, as in the end, it does not matter if it is combination, extra work or extra time as long as it is for better patient hospitality sake.

Figure 6 (a) Comparison of k-means with BiSim and Levenshtein similarities in terms of clustering accuracy and (b) Comparison of DBSCAN with BiSim and Levenshtein similarities in terms of clustering accuracy (see online version for colours)



(a)



(b)

5 Conclusions

In this paper, and with the intention of improving the patient safety by addressing LASA medicines issue, we studied the application of unsupervised clustering approaches, k-means and DBSCAN on two similarity measures, BiSim and Levenshtein conjunctionally. The best out of the combinations is obtained by BiSim based k-means clustering, with a silhouette score equal to 0.5.

This can be exploited in reducing medication errors whether in dispensing, classification, filling an automatic dispensing cabinet …, etc. and this by taking cautious measures for medicines of the same clusters. It may be extended to the prediction of confusable drug names.

The clustering using the above approaches can be blamed for its singularity; which means each medicine can only belong to one cluster at a time. On the other hand, it is

capable of spotting confusable medicines that have not been detected before. Hence, associating this approach along with other data mining approaches is the best option for the sake of improving patient safety.

References

Alsaeedi, A. (2020) 'A survey of term weighting schemes for text classification', *International Journal of Data Mining, Modelling and Management*, Vol. 12, No. 2, pp.237–254.

Chen, L-C., Chen, C-H., Chen, H-M. and Tseng, V.S. (2011) 'Hybrid data mining approaches for prevention of drug dispensing errors', *Journal of Intelligent Information Systems*, Vol. 36, No. 3, pp.305–327.

Ciociano, N. and Bagnasco, L. (2014) 'Look alike/sound alike drugs: a literature review on causes and solutions', *International Journal of Clinical Pharmacy*, Vol. 36, No. 2, pp.233–242.

Dagan, I., Pereira, F. and Lee, L. (1994) *Similarity-Based Estimation of Word Cooccurrence Probabilities*, ArXiv Preprint Cmp-Lg/9405001.

Ding, H., Copeland, K., Greenhaw, R., Mills, B., Hileman, C. and Kieu, V. (2021) *Drug Name Correction of Medication Records from Aeromedical Certification Exams*, No. DOT/FAA/AM-21/23.

Ester, M., Kriegel, H-P., Sander, J., Xu, X. et al. (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise', in *KDD*, Vol. 96, pp.226–231.

Filik, R., Price, J., Darker, I., Gerrett, D., Purdy, K. and Gale, A. (2010) 'The influence of tall man lettering on drug name confusion', *Drug Safety*, Vol. 33, No. 8, pp.677–687.

Gates, P.J., Hardie, R-A., Raban, M.Z., Li, L. and Westbrook, J.I. (2021) 'How effective are electronic medication systems in reducing medication error rates and associated harm among hospital inpatients? A systematic review and meta-analysis', *Journal of the American Medical Informatics Association*, Vol. 28, No. 1, pp.167–176.

Hadwan, M., Al-Hagery, M.A., Al-Sanabani, M. and Al-Hagree, S. (2021) 'Soft bigram distance for names matching', *Peer J. Computer Science*, Vol. 7, p.e465, https://doi.org/10.7717/peerj-cs.465.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining Concepts and Techniques*, 3rd ed., University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

Joachims, T. (1999) 'Making large-scale support vector machine learning practical, advances in kernel methods', *Support Vector Learning*, pp.169–184 [online] https://dl.acm.org/doi/10.5555/299094.299104.

Lambert, B.L. (1997) 'Predicting look-alike and sound-alike medication errors', *American Journal of Health-System Pharmacy*, Vol. 54, No. 10, pp.1161–1171.

Lambert, B.L., Lin, S-J., Chang, K-Y. and Gandhi, S.K. (1999) 'Similarity as a risk factor in drug-name confusion errors: the look-alike (orthographic) and sound-alike (phonetic) model', *Medical Care*, Vol. 37, No. 12, pp.1214–1225.

Levenshtein, V.I. et al. (1966) 'Binary codes capable of correcting deletions, insertions, and reversals', in *Soviet Physics Doklady*, Vol. 10, No. 8, pp.707–710.

McLachlan, S., Dube, K., Hitman, G.A., Fenton, N.E. and Kyrimi, E. (2020) 'Bayesian networks in healthcare: distribution by medical condition', *Artificial Intelligence in Medicine*, Vol. 107, p.101912, https://doi.org/10.1016/j.artmed.2020.101912.

Millán-Hernández, C.E. et al. (2016) *Detección de Nombres de Medicamentos Confusos Por Su Parecido Ortográfico o Fonético Mediante Un Algoritmo Genético*, Autonomous University of the State of Mexico, http://hdl.handle.net/20.500.11799/65510.

Millán-Hernández, C.E., Garcia-Hernández, R.A., Ledeneva, Y. and Hernández-Castañeda, A. (2019) 'Soft bigram similarity to identify confusable drug names', in *Mexican Conference on Pattern Recognition*, pp.433–442.

Mouattah, A. and Hachemi, K. (2021) 'Estimation of medication dispensing errors (MDEs) as tracked by passive RFID-based solution', *International Journal of Healthcare Information Systems and Informatics (IJHISI*), Vol. 16, No. 3, pp.89–104.

Mulac, A., Mathiesen, L., Taxis, K. and Granås, A.G. (2021) 'Barcode medication administration technology use in hospital practice: a mixed-methods observational study of policy deviations', *BMJ Quality & Safety*, Vol. 30, No. 12, pp.1021–1030.

Sander, J., Ester, M., Kriegel, H-P. and Xu, X. (1998) 'Density-based clustering in spatial databases: the algorithm Gdbscan and its applications', *Data Mining and Knowledge Discovery*, Vol. 2, pp.169–194, https://doi.org/10.1023/A:1009745219419.

Sim, K.R., and Robertson, H.E. (2008) *The Canadian Regime for Protecting Against Pharmaceutical Trademark Confusion and Mistakes*, Trademark Rep., Vol. 98, p.1253.

Tseng, V.S., Chen, C-H., Chen, H-M., Chang, H-J. and Yu, C-T. (2007) 'Analysis and prevention of dispension errors by using data mining techniques', in *2007 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pp.65–70.

Wang, G., Jung, K., Winnenburg, R. and Shah, N.H. (2015) 'A method for systematic discovery of adverse drug events from clinical notes', *Journal of the American Medical Informatics Association*, Vol. 22, No. 6, pp.1196–1204.

Watterson, T.L., Stone, J.A., Brown, R., Xiong, K.Z., Schiefelbein, A., Ramly, E., Kleinschmidt, P., Semanik, M., Craddock, L., Pitts, S. et al. (2021) 'CancelRx: a health IT tool to reduce medication discrepancies in the outpatient setting', *Journal of the American Medical Informatics Association*, Vol. 28, No. 7, pp.1526–1533.