Unsupervised clustering of materials properties using hierarchical techniques

Arafa S. Sobh*

Mechanical Engineering Department,
Helwan University,
Al Sikka Al Hadid Al Gharbeya,
Qism Helwan, Cairo Governorate, Egypt
Email: arafasobh@yahoo.com
*Corresponding author

Sameh A. Salem

Electronics, Communications and Computers Department,
Helwan University,
Al Sikka Al Hadid Al Gharbeya,
Qism Helwan, Cairo Governorate, Egypt
Email: sameh.asalem@yahoo.com

Rania Darwish and Mohammed Hussein

Mechanical Engineering Department,
Helwan University (On leave to BUE),
Al Sikka Al Hadid Al Gharbeya,
Qism Helwan, Cairo Governorate, Egypt
Email: raniaa04@yahoo.com
Email: mohammed.hussein@bue.edu.eg

Omar Karam

Faculty of Informatics and Computer Science,
Computer Science Department,
British University in Egypt (BUE),
Egypt
Email: omer.karam@bue.edu.eg

Abstract: Data mining (DM) algorithms arose as a promising and flourishing discipline at manufacturing and industrial engineering. This paper proposes an efficient decision support approach for manufacturing engineering. The proposed approach tackles clustering challenges for engineering materials properties. It adopts the hierarchal clustering for mining engineering materials properties. Extensive experiments and comparisons are conducted on three different real-world datasets for engineering materials properties. In addition, a study of different similarity measures is carried out to choose the best fit similarity measure to engineering material datasets. A comparison of the results with other competitors clearly shows the robustness of the proposed approach. Therefore, it is highly recommended to use the proposed approach as a scalable engineering material properties tool.

Keywords: decision support systems; manufacturing systems; linkage hierarchal algorithm; materials properties.

Reference to this paper should be made as follows: Sobh, A.S., Salem, S.A., Darwish, R., Hussein, M. and Karam, O. (2015) 'Unsupervised clustering of materials properties using hierarchical techniques', *Int. J. Collaborative Enterprise*, Vol. 5, Nos. 1/2, pp.74–88.

Biographical notes: Arafa S. Sobh graduated with a BSc and MSc in Mechanical Engineering Department both from Helwan University, Egypt, in 2001 and 2009, respectively. His research interests include material science applications, clustering algorithms, and data mining.

Sameh A. Salem is an Associate Professor since 2014 in Helwan University, Cairo, Egypt. In 2008, He received the degree of PhD in Engineering from Department of Electrical Engineering and Electronics, The University of Liverpool, UK. His research interests include clustering algorithms, machine learning, data mining, parallel computing, and cloud computing. Also, he is the Coordinator and Academic Advisor at the Department of Communication and Information Technology, Uninettuno University (Italy) in corporation with Faculty of Engineering, Helwan University (Egypt). Furthermore, he is reviewing several proposals and research projects at the National Telecommunication Regulatory Authority (NTRA) – Egypt.

Rania Darwish is an Assistant Professor at Helwan University. She received her PhD in Communication Engineering from Helwan University, Cairo, Egypt.

Mohammed Hussein is currently a Professor at the Mechanical Engineering Department, Helwan University, Egypt. He obtained his PhD from Helwan University and California State Polytechnic University (channel system) in 1991. He is now on-leave to the BUE, Cairo Egypt. In 2000, He was appointed as an Associate Professor. He has many research papers in scientific journals and conferences proceedings. His research work includes visual inspection, quality engineering, supply chains and material control, and technology transfer.

Omar Karam is a Professor of Information Systems at the Faculty of Informatics and Computer Science of the British University in Egypt (BUE). He is on secondment from Ain Shams University, Egypt. He obtained his BSc (1980) in Engineering (Communications and Electro-physics) from Alexandria University and his MSc (1987) and PhD (1993) in Computer Engineering from North Carolina State University. Prior to the BUE, he was the Director of the Egyptian Universities Network (EUN). He has extensive professional experience in ICT management and is the author or co-author of more than 60 scientific publications in international journals and conferences. His current research interests are data mining and warehousing, interconnection networks, optoelectronic computing and ICT in education.

1 Introduction

Knowledge is the most valuable asset for a manufacturing enterprise, as it enables a business to differentiate itself from competitors and to compete efficiently and effectively to the best of its ability. Knowledge exists in all business functions, including purchasing, marketing, designing, production, maintenance and distribution; however, discovering knowledge could be notoriously difficult to identify, capture and manage (Harding et al., 2006). Drawn to adopting data mining concepts emerges as a promising discipline that could facilitate the discovery of the required knowledge stored at the manufacturing databases and warehouses.

Studying and analysing the properties of materials (e.g., mechanical, chemical, physical, thermal, electrical, optical, etc.) have a substantial impact at the manufacturing community (Al-Mubaid et al., 2009). Moreover, it could also accelerate the research process and guide the development of new materials with selected engineering properties (Sander et al., 2003).

However, most production engineers who perform manufacturing processes have little knowledge about engineering materials properties that have a direct effect on final product assessments (Jeswiet et al., 2015). In case of modifying or make a new product, design engineers who are acquainted and understand the relationships between manufacturing processes that will be performed on the materials and the changes that will be undergone to material properties (Garcia-Munoz, 2014; Ronowicz et al., 2015).

Moreover, certain material properties have a direct effect on others such as chemical composition properties. Chemical composition means alloying elements that are merged together in the molten state and then solidified. Carbon and iron contents, for example, have a direct effect on mechanical strength property. Iron and chromium content influence chemical corrosion property. In addition, materials phase transformations are important to designer and manufacturer because these transformations influenced the mechanical and physical properties (Callister Jr. and Rethwisch, 2010).

Thus, the choice of the suitable materials with the appropriate properties for a certain manufacturing application requirements necessitate a decision support system rather than relying on expertise engineers (Gorunescu, 2011; Nie et al., 2009). Decision support system helps engineers at analysing and exploring vast amount of variables concerned with each operation that performed on materials (Gupta et al., 2015). Therefore, decision support tools could elevate system reliability, maintenance quality, meanwhile cutting of processing time and cost (Horng et al., 2011; Cebrail and Esra, 2010).

Therefore, this work is oriented towards developing a scalable decision support system that is tolerated for discovering and analysing materials properties to benefit the manufacturing community and industrial engineers. The proposed approach employs data mining techniques to support the decision-making processes. The main concern of this work is tackling the unsupervised clustering for engineering materials properties. In this study, a linkage hierarchal clustering techniques applied to explore three different-sized material properties databases. The remaining of the paper is organised as follows: Section 2 presents the related work, while Section 3 presents the hierarchal clustering algorithms. Section 4 shows engineering materials databases. The proposed unsupervised generic model is presented in Section 5. Experimental results and conclusion will be discussed in Sections 6 and 7, respectively.

2 Related work

Substantial attention has been drawn towards studying engineering materials properties. The first attempts served the field of prediction modelling engineering materials properties were mathematical models that gain acceptance as conducting predictive tools (Agarwal et al., 2014). Mathematical predictive models could be classified into three basic models: deterministic models (Monsia, 2012; Kirlova et al., 2009), stochastic models (Jurgens et al., 2012; Dong et al., 2013) and probabilistic models (Michael and Mital, 1998; Köhler et al., 2007). However, all the previously mentioned models are attempted to solve one- or two-dimensional problem. Moreover, most of these models interested to construct their predictive mathematical models to deal with one class, family of materials or a few members of materials. Additionally, these models lakes the required accuracy and robustness because of the difficulty and complexity of materials' variables (Agarwal et al., 2014).

Clustering, as suggested tool, unsupervised DM technique is the art of how to find groups or clusters in datasets, based on their similarity (Salem, 2007; Kanungo et al., 2002). Therefore, clustering techniques are helpful in decision support system and finding variables relationship (Krishnakumari and Vivekanandan, 2009). The clustering process could be summarised into two main steps. The first one is evaluating the similarity measurement that can be performed by one of the following measurements such as Euclidean, Manhattan or Pearson. The next step is building an algorithm to construct clusters based on the chosen criterion which can be either partition or hierarchal approach (Gorunescu, 2011; Salem, 2007; Han and Camber, 2006; Maurizio, 2011).

According to the authors' knowledge, most of the previous efforts that are related to high-dimensionality problems are basically depending on partition techniques, such as *K*-means or its developed techniques. The authors in Shi et al. (2011) used the classification power of *k*-means to cluster the similar dense points in 3D manufacturing scan. However, the authors in Yiakopoulos et al. (2011) investigated the *k*-means as rolling bearing fault detection technique for industrial environment. Also, the authors in Haghtalab et al. (2015) considered *k*-means as robust unsupervised technique for monitoring manufacturing processes and applications that need control charts. In addition, the authors in Chien et al. (2007) utilised *k*-means as decision support tool in semiconductor manufacturing processes and they showed how engineers experience is not enough to find root causes of defects. The authors in Gupta et al. (2015) exploits data mining techniques in material properties knowledge discovery to obtain the optimal process parameters in manufacturing pellet products.

Moreover, there are some trail efforts to develop clustering performance, such as Ben Khediri et al. (2012) who adopted fault detection techniques consists of Kernel *k*-means and support vector domain description to monitor and detect Etch Metal process alarms. A hybrid self-organising map (SOM) then *K*-means followed by C4.5 decision tree classifier were used by Tsai, (2012) to identify soldering defect patterns. Doreswamy and Hemanth (2012) carried out a study on *k*-means with different similarity measure functions and proposed a new similarity measure called design specification distance (DSD) function for the selection of engineering materials.

Thus, it could be seen that most previously mentioned efforts concentrate on manufacturing engineering applications and its' fields. However, studying materials properties that served most of manufacturing engineering applications are not investigated as worthy. Additionally, most of researchers use low-dimensionality datasets. Meanwhile, the fore mentioned efforts neglect inclusions or outliers patterns, but neglecting outliers' patterns could minimise the clustering accuracy (Salem, 2007; Doreswamy and Hemanth, 2012).

Finally, most of the fore-mentioned efforts employed *K*-means algorithms which suffer some problems. Most importantly, the *K*-means algorithm needs a prior determination for the initial number of clusters, however, this a prior determination is not always accurate or well defined by users. Also, *k*-means algorithm needs to run several times to reach stability of results. Further, the Euclidean distance is the most commonly similarity measurement used with *K*-mean algorithm and it bypasses some datasets. Where outliers obtained by the Euclidean distance may be considered as important data for decision-making Geetha and Arock (2009) and Abu Abbas (2008).

Hierarchal clustering techniques overcome the problem involved with *k*-means and also can be considered as robust and effective techniques for materials decision support systems (Chauhan and Vaish, 2013). Chauhan and Vaish (2013) presented a hierarchal clustering as an effective technique that capable to solve the problem of hard coating material selection. They also constructed a hierarchal decision-making approach to select magnetic materials (Chauhan and Vaish, 2012).

Kantar and Keskin (2013) used linkage hierarchal to investigate the relationship between electrical consumption and economic growth. In addition, Willett (1982) recommended one of six hierarchal clustering techniques in predicting molecular property of compounds materials. And for important reengineering application in such way redesign of work flow and processes, Cui and Chae (2011) recommended a linkage hierarchal clustering as decision support system for identifying components.

As discussed above hierarchal techniques can be considered as a powerful clustering techniques and have a wide applications in manufacturing fields. Its strength gained by using it with different similarities such as average, weight, single, complete, etc. However, it can be concluded that the most applicable clustering criteria are average, complete and single linkages. But also, single linkage suffers spread out of clusters that called chaining and therefore clusters will be not compact enough, however, complete linkage has crowding shortage that means compact but not far enough apart (Tibshirani, 2013). Therefore, it can be recommended the average linkage owing to its ability to avoid these problems by taking the average pairwise similarity, so that the clusters will be compact and far apart (Salem, 2007; Tibshirani, 2013).

Not only, is the aim of the present work to recommend a hierarchal clustering technique but also to investigate and recommend one of different distance measurements. The distance measurements that will be evaluated and compared are Euclidean, Squared Euclidean, Jaccard and Cosine (Murtagh and Contreras, 2011).

Therefore, this paper presents a data mining-based decision support system that targets materials properties clustering. Apart from previously mentioned efforts, this work employs linkage hierarchal clustering technique. The proposed approach tolerated for low and high-dimensionality materials datasets. Three different material properties datasets have been utilised and collected from on line material site with a total of 75 materials (http://www.matweb.com).

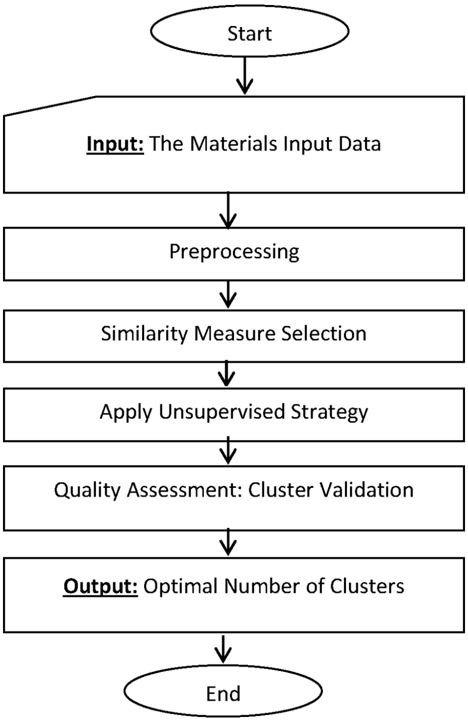
3 The proposed unsupervised generic model

This subsection presents the proposed generic model adopted to add a powerful tool for decision-making systems. The procedure of the generic model is presented in Figure 1.

It can be denoted that the proposed model consists of six elements or steps of inputs illustrated as following: the first step concerns the materials dataset to be clustered; it should be entered in tabulated matrix form. Normalisation is a pre-processing step performed on each dataset to make all the attributes of materials have the same weights as illustrated in Section 4. Step number 3 in the proposed model decides which clustering algorithm will be applied that could be either by partitional or hierarchal techniques also the criteria of dissimilarity between clusters that may be selected for hierarchal techniques: average, complete or single. Distance measure should be determined in step number 4; Euclidean, Squared-Euclidean, Cosine, Jaccard, etc. The fifth Step is a supplementary step performed to evaluate and decide the optimal number of clusters by step number 6.

The following sections present some points of the proposed approach. Firstly, Section 3.1 describes the exploited linkage hierarchal algorithm. Secondly, Section 3.2 illustrates validation indices. Thirdly, Section 3.3 describes the employed material properties datasets.

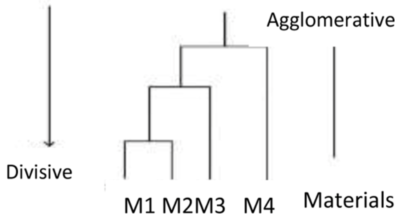
Figure 1 Proposed unsupervised generic model


Start
Input: The Materials Input Data
Preprocessing
Similarity Measure Selection
Apply Unsupervised Strategy
Quality Assessment: Cluster Validation
Output: Optimal Number of Clusters
End

3.1 Hierarchal clustering algorithms

Hierarchical algorithms, non-partition algorithms, are unsupervised clustering techniques in which clustering criterion either divisive method from up to down or merging process successively is perform from bottom to up that can be named agglomerative method (Chauhan and Vaish, 2012). The output form of these algorithms is a dendrogram tree, shown in Figure 2, which consider each material in the dataset as individual cluster, and successively merge the most similar or closest pair of clusters according to a similarity measure (Salem, 2007).

Figure 2 Agglomerative and divisive hierarchal techniques



Dissimilarity between two clusters in hierarchal algorithms uses different linkage criteria in comparisons; average, complete and single as shown in Figure 3. Similarity between the two clusters that will be merged, based on the average distance between all possible pairs of objects in the clusters, can be calculated by equation (1) as follows:

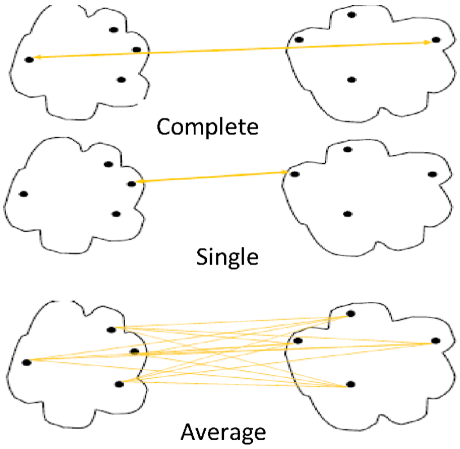
$$\|H-G\| = \frac{1}{|H|.|G|.}\sum\|x-y\| \tag{1}$$

$$\|H-G\| = \max\|x-y\| \tag{2}$$

$$\|H-G\| = \min\|H-G\|, \tag{3}$$

where x and y objects belonging to H and G clusters, respectively, and represented the two clusters. $|H|$ and $|G|$ represent the number of objects in the two clusters. However, the dissimilarity between two clusters for complete and single linkage are based on the maximal and minimal distance between the objects belonging to the corresponding clusters, respectively (Salem, 2007; Kantar and Keskin, 2013), as presented in equations (2) and (3).

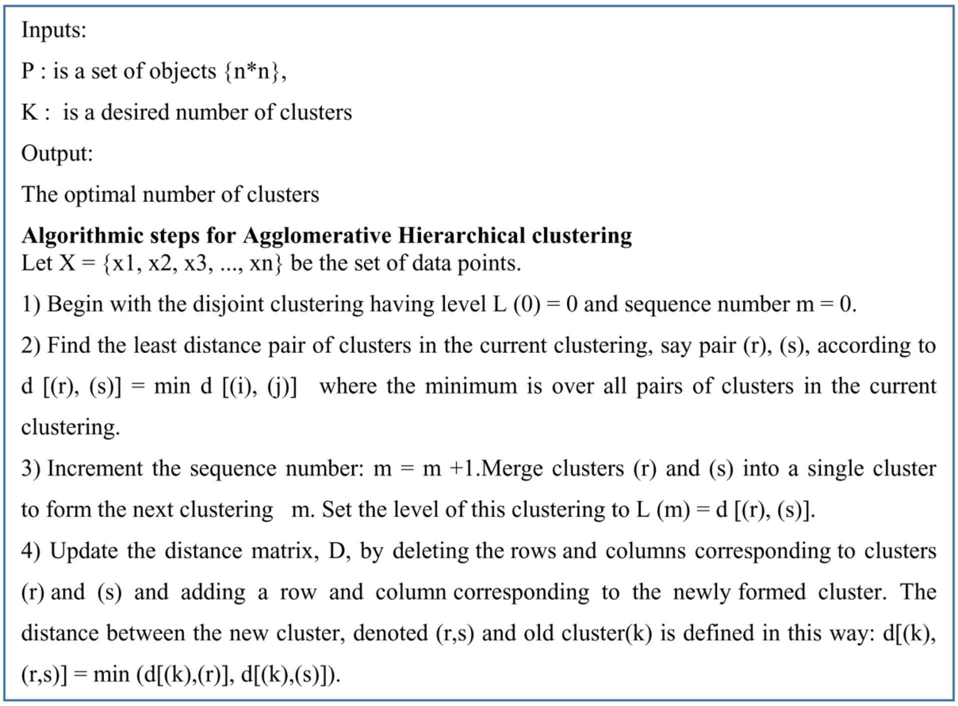
Figure 3 Examples of hierarchical clustering: (a) single linkage; (b) complete linkage and (c) average linkage (see online version for colours)



Source: Salem (2007)

Figure 4 describes the algorithmic steps of the employed hierarchical clustering technique.

Figure 4 Agglomerative hierarchical clustering adopted from Murtagh and Contreras (2011) (see online version for colours)



Inputs:

P : is a set of objects {n*n},

K : is a desired number of clusters

Output:

The optimal number of clusters

Algorithmic steps for Agglomerative Hierarchical clustering
Let X = {x1, x2, x3, ..., xn} be the set of data points.

1) Begin with the disjoint clustering having level L (0) = 0 and sequence number m = 0.

2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d [(r), (s)] = min d [(i), (j)] where the minimum is over all pairs of clusters in the current clustering.

3) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L (m) = d [(r), (s)].

4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).

3.2 Validation indices

Validation indices could be considered as effective assessment and evaluation standards and criteria to provide degree of confidence for the clustering results obtained from the used algorithm. There are more than one validation measurements can be used such as Dunn's index, Calinski-Harabasz index, Davies-Bouldin index, C-index and Silhouette index (Salem, 2007; Bolshakova and Azuaje, 2003). The latter is suitable for estimating only the first choice or best partition (Bolshakova and Azuaje, 2003). Silhouette measure could be used as a confidence indicator on the membership of the *i*th sample in cluster *Xj*. The Silhouette index for the *i*th sample in cluster *Xj* is determined by equation (4).

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{4}$$

where *a*(*i*) is the average distance between the *i*th sample and all of the samples included in *Xj*; and *b*(*i*) is the minimum average distance between the *i*th sample and all of the samples clustered in X_k ($k = 1, \ldots, c$; $k \neq j$) 'max' is the maximum operator. It can be denoted that values of *S*(*i*) can be ranged from (–1) to (1) When a *S*(*i*) is close to 1, that means the *i*th sample has been 'well–clustered', i.e., it was assigned to an appropriate cluster. When $S(i)$ is close to zero, it suggests that the ith sample could also be assigned to the nearest neighbouring cluster. If $S(i)$ is close to (–1) it means that such a sample has been 'misclassified' (Bolshakova and Azuaje, 2003).

3.3 Engineering materials datasets

Throughout this work, three different datasets collected from two different sources have been exploited. Table 1 describes the exploited datasets' specifications. Dataset one of ferrous-ferrous materials; steels and cast irons materials and dataset two concerned ferrous–non-ferrous materials, however, dataset three named metallic-non-metallic materials. Datasets one and three are collected from mat-website (http://www.matweb.com) but, dataset two is obtained from reference (Al-Mubaid et al., 2009). These three materials properties datasets are of different families and sizes. Tracked features or attributes are composed of thermal, electrical, physical, mechanical and chemical compositions. Datasets are organised in files of tabular format in which instances and variables are tabulated in rows and columns. Normalisation and weighed voting techniques are proposed to improve the prediction of the number of clusters. The three datasets are normalised by one of the three normalisation methods; Min-max., decimal scaling and Z-score method (Han and Camber, 2006). Z-score or Zero-mean can be performed by using equation (4) in which an attribute P are normalised based on the mean and standard deviation. A value d of P is normalised to D' by equation (5)

$$D' = \frac{(d - \text{mean}(P))}{\text{Std}(P)}, \tag{5}$$

where D' is the normalised attribute, mean (P) refers to the mean of attribute P and Std (P) refers to the standard deviation of attribute P (Jain et al., 2005; Jain and Bhandare, 2011).

Table 1 Description of materials datasets

Datasets	*No. of materials*	*No. of features*	*No. of classes*
Ferrous-ferrous	24	24	2
Ferrous-nonferrous	18	84	2
Metallic-nonmetallic	33	24	2

3.3.1 (Ferrous-ferrous) dataset

Dataset one is sampled from ferrous families that containing iron element. Two basic different subfamilies are collected: Steel and Cast iron materials which can be differentiated by percentage of carbon. This dataset consists of 24 materials or objects. Each material is characterised by 24 features.

3.3.2 (Ferrous–non-ferrous) dataset

Dataset two is sampled from two different families: ferrous and non-ferrous materials. This dataset consists of 18 materials or objects. Each material is characterised by 84 features.

3.3.3 (Metallic–non-metallic) dataset

Dataset three is sampled from two different families: metallic and non-metallic. This dataset consists of 33 materials or objects. Each material is characterised by 24 features.

The following sections shows and demonstrate the main aim of the present work that is the investigation of the effectiveness and robustness of the proposed hierarchal algorithms along with different similarity measurements compared with *K*-means algorithm results.

4 Experimental results

In this section, a series of experiments were carried out to examine the performance of hierarchical and partitional algorithms on clustering the engineering materials. In these experiments, each clustering algorithm is tested on different real-world datasets using several similarity measures at different number of clusters. All the experiments have been implemented through Matlab®. It should be noted that each clustering algorithm can produce some clusters regardless of whether or not clusters existence. Therefore, it is essential to assess the quality of clustering results with assessment criteria which have no preferences to any algorithm. In this paper, silhouette index (Salem, 2007; Han and Camber, 2006; Shi et al., 2011; Kantar and Keskin, 2013) is used. Its value ranges between –1 and 1: a value near 1 indicates that the point Mi is affected to the right cluster whereas a value near –1 indicates that the point should be affected to another cluster. In this context, the best clustering results are corresponding to higher values of silhouette index.

4.1 Evaluation of Hierarchal techniques with different clustering criterion

A series of experiments with different number of clusters *K* ($K = 2, \ldots, 8$) were carried out using variations of hierarchical algorithms namely hierarchical with single linkage (H-single), hierarchical with average linkage (H-avg.), hierarchical with complete linkage (H-comp.) and K-means as partitional clustering algorithm.

Figures 5–7 show the validation index of four clustering algorithms on dataset 1 vs. the number of clusters that are considered above, where the number of clusters corresponding to the highest validation value is expected the true number of clusters. Furthermore, higher values of cluster validation indicate better clustering quality. As shown for the three figures, the value of $K = 2$ represents the best fit number of clusters which actually coincides with the true number of clusters in the underlying three datasets, namely Steel and cast irons, ferrous and non-ferrous, Metallic and non-metallic, respectively. It should be noted that the contrast of validation values at different number of clusters are low. This is owing to the low separation between clusters and the existence of overlap structures in these datasets.

Referring to Figure 5 of ferrous–ferrous results, the average and single linkages have the highest validation scores than complete linkage along different number of clusters. However results of ferrous–non-ferrous, Figure 6 shows that average and complete linkages have the highest validation scores than single linkage in this case. On the other hand, the highest scores are the average and complete linkages for metallic–non-metallic dataset as shown in Figure 7.

Figure 5 Validation indices of clustering results of Ferrous-ferrous dataset using different clustering algorithms (see online version for colours)

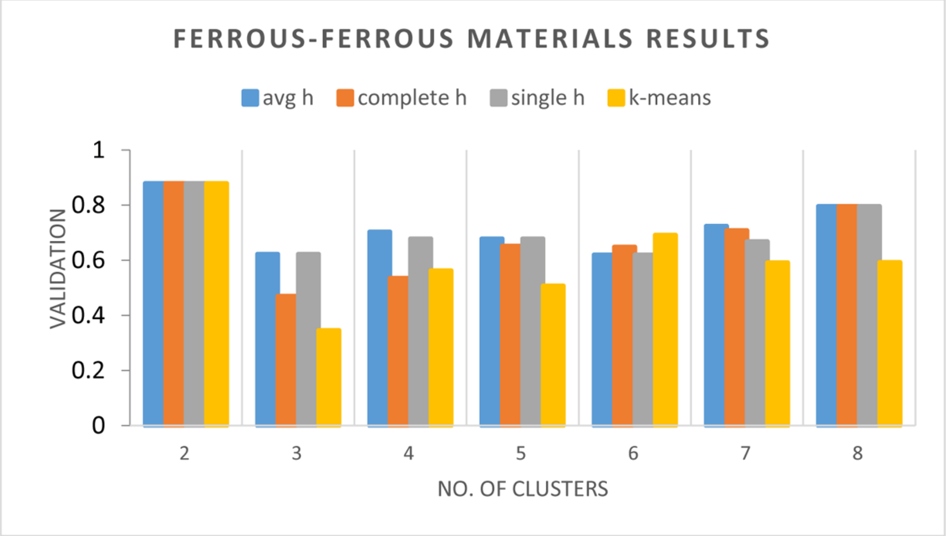
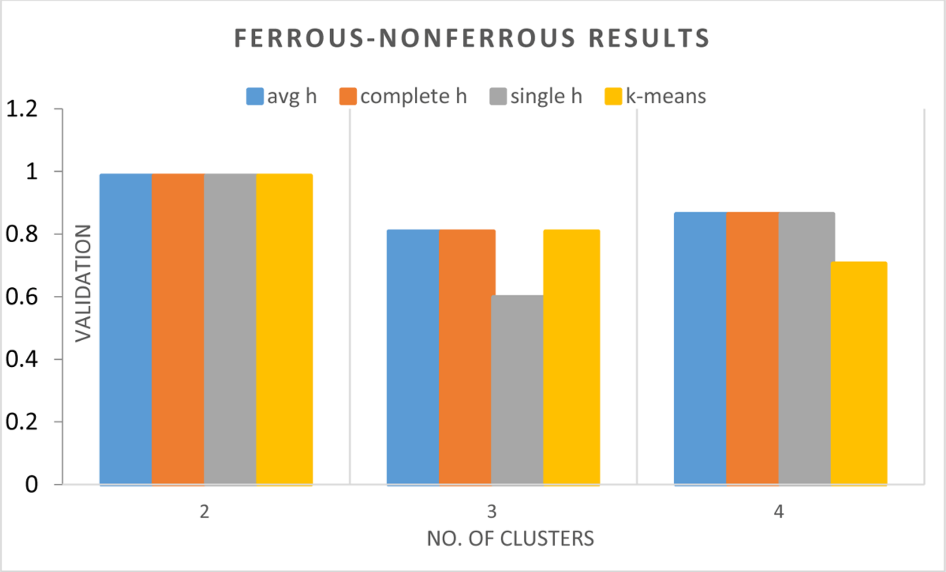
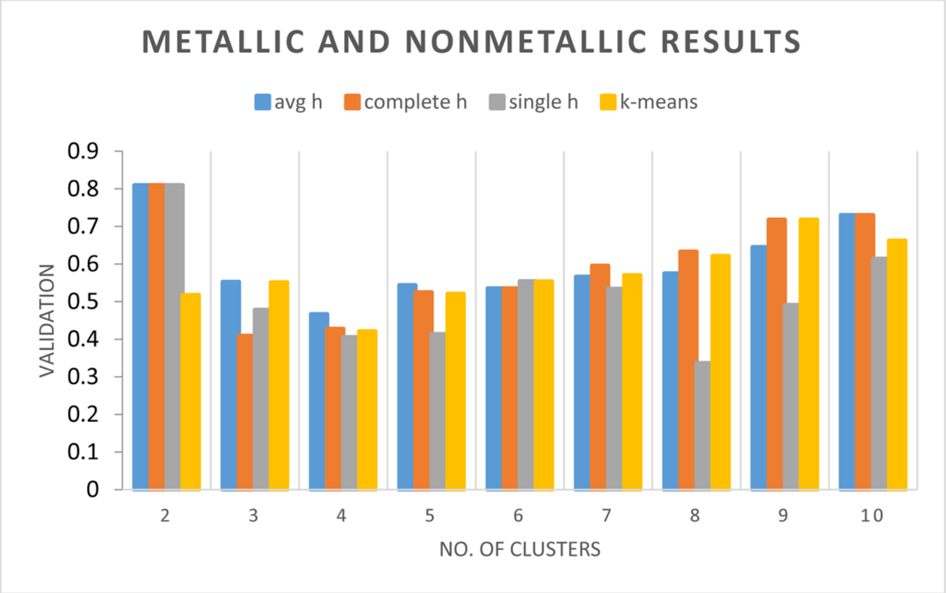


Figure 6 Validation indices of clustering results of Ferrous-nonferrous using different clustering algorithms (see online version for colours)



Based on the above results and discussion, it could be concluded that linkage hierarchal algorithms produce more consistent results for materials properties than *K*-means for the three datasets. Moreover, average linkage hierarchal clustering technique is consistently the best fit algorithm for the three different datasets.

Figure 7 Validation indices clustering results of metallic and nonmetallic using different clustering algorithms (see online version for colours)



4.2 *Evaluation of hierarchal clustering with different similarity measures*

In this section, another series of experiments were carried out to examine the performance of average hierarchical technique on clustering the engineering materials. In these experiments, the proposed technique is tested on the same three datasets using several distance measures at different number of clusters.

Tables 2–4 illustrate the results of exploiting different distance measurements namely: Euclidean, Squared Euclidean (S.Euclidean), Cosine and Jaccard similarities, respectively.

Table 2 Results of dataset one for average linkage with different distance measures

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$
Euclidean	*0.8792*	0.6219	0.7027	0.6777	0.62	0.7238	0.7957
S.Euclidean	*0.8792*	0.6118	*0.1371*	*0.1216*	*–0.0289*	0.122	*0.3366*
Cosine	*0.8792*	0.6219	0.678	0.6518	0.6396	0.7238	0.7957
Jaccard	0.1361	–0.5588	–0.5015	–0.4418	–0.3817	–0.2772	–0.2011

Table 3 Results of dataset two for average linkage with different similarity measurements

	$K = 2$	$K = 3$	$K = 4$
Euclidean	*0.9861*	0.8078	0.8633
S. Euclidean	*0.9861*	*0.5977*	*0.8633*
Cosine	*0.9861*	0.8078	0.8633
Jaccard	–0.6923	–0.6088	–0.5048

Table 4 Results of dataset three for average linkage with different similarity measurements

	K = 2	*K = 3*	*K = 4*	*K = 5*	*K = 6*	*K = 7*	*K = 8*	*K = 9*	*K = 10*
Euclidean	*0.8094*	0.5525	0.4666	0.5435	0.5352	0.5663	0.575	0.645	0.73
S.Euclidean	*0.5608*	*0.4519*	*0.2663*	*0.1541*	*0.209*	*0.2881*	*0.3317*	*0.2898*	*0.2604*
Cosine	*0.8094*	0.5525	0.4666	0.5435	0.5352	0.5663	0.6329	0.645	0.73
Jaccard	–0.3907	–0.3432	–0.363	–0.3042	–0.2467	–0.22	–0.1731	–0.2176	–0.196

By studying and analysing Tables 2–4, it can be seen that average linkage has the highest validation values for both Cosine and Euclidean measurements through all number of clusters. Interestingly, Cosine and Euclidean for all linkage hierarchal have the greatest value at number of cluster 2, but S-Euclidean is the more sensitive to clusters' number changing. So that, it could be concluded that, S-Euclidean similarity-based average linkage hierarchal approach is the most applicable for these datasets. After analysing the results of clustering algorithms it could be concluded that, average linkage hierarchal algorithms are recommended for small and large datasets with low- and high-dimensionality over K-means for engineering material properties. In addition, S-Euclidean is the most applicable similarity measurement for our problem.

5 Conclusion and future work

In this paper, an efficient and robust decision support approach to manufacturing engineering has been introduced. As demonstrated, the proposed approach could be beneficial in releasing several challenges to materials properties. In this context, the predication of optimal cluster partitions along with the best fit similarity measure to engineering materials have been achieved. Experimental results on real-world datasets have showed the reliability and robustness of the proposed approach in the predication of engineering materials properties.

In future, the proposed approach can be integrated with different validation indices in cooperation with alternative similarity measures to examine in order to the structure bias in the data. Moreover, further investigations are needed to examine the effect of feature reduction on the clustering performance.

References

Abu Abbas, O. (2008) 'Comparisons between data clustering algorithms', *The International Arab Journal of Information Technology*, Vol. 5, pp.320–325.

Agarwal, K., Shivpuri, R. and Bonthapally, V. (2014) 'Process-structure-microstructure relationship in hot strip rolling of steels using statistical data mining', *Procedia Engineering*, Vol. 81, pp.90–95.

Al-Mubaid, H., Abouel Nasr, E.S. and Hussein, M. (2009) 'A methodology for mining material properties with unsupervised learning', *Int. J. Rapid Manufacturing*, Vol. 1, pp.237–252.

Ben Khediri, I., Weihs, C. and Limam, M. (2012) 'Kernel k-means clustering based local support vector domain description fault detection of multimodal processes', *Expert Systems with Applications*, Vol. 39, pp.2166–2171.

Bolshakova, N. and Azuaje, F. (2003) 'Cluster validation techniques for genome expression data', *Signal Processing*, Vol. 83.

Callister Jr., W.D. (2001) *Fundamentals of Materials Science and Engineering*, John Wiley and Sons, Inc., 5th ed., Ch. 1.

Callister Jr., W.D. and Rethwisch, D.G. (2010) *Materials Science and Engineering an Introduction*, 8th ed., Chapter 1, Wiley.

Cebrail and Esra (2010) 'Implementing a data mining solution for enhancing carpet manufacturing productivity', *Knowledge-Based Systems*, Vol. 23.

Chauhan, A. and Vaish, R. (2012) 'Magnetic material selection using multiple attribute decision making approach', *Materials & Design*, Vol. 36, pp.1–5.

Chauhan, A. and Vaish, R. (2013) 'Hard coating material selection using multi-criteria decision making', *Materials and Design*, Vol. 44.

Chien, C., Wang, W. and Cheng, J. (2007) 'Data mining for yield enhancement in semiconductor manufacturing and an empirical study', *Expert Systems with Applications*, Vol. 33, pp.192–198.

Cui, J.F. and Chae, H.S. (2011) 'Applying agglomerative hierarchical clustering algorithms to component identification for legacy systems', *Information and Software Technology*, Vol. 53, pp.601–614.

Dong, L., Gamal, S.H. and Atluri, S.N. (2013) 'Stochastic macro material properties, through direct stochastic modeling of heterogeneous microstructures with randomness of constituent properties and topologies, by using Trefftz computational grains (TCG)', *CMC*, Vol. 37, pp.1–21.

Doreswamy and Hemanth. K.S. (2012) 'A novel design specification distance (DSD) based K-mean clustering performance evaluation on engineering materials' data base', *International Journal of Computer Applications*, Vol. 55, pp.26–33.

Garcia-Munoz, S. (2014) 'Two novel methods to analyze the combined effect of multiple raw-materials and processing conditions on the product's final attributes: JRPLS and TPLS', *Chemometrics and Intelligent Laboratory Systems*, Vol. 133.

Geetha, T. and Arock, M. (2009) 'Effective hybrid PSO and K-means clustering algorithm for gene expression data', *Int. J. Rapid Manufacturing*, Vol. 1, pp.173–188.

Gorunescu, F. (2011) *Data Mining: Concepts, Models and Techniques*, 3rd ed., Chapter 5.

Gupta, A., Cecen, A., Goyal, S., Singh, A.K. and Kalidindi, S.R. (2015) 'Structure–property linkages using a data science approach: application to a non-metallic inclusion/steel composite system', *Acta Materialia*, Vol. 91, pp.239–254.

Haghtalab, S., Xanthopoulos, P. and Madani, K. (2015) 'Arobust unsupervised consensus control chart pattern recognition framework', *Expert Systems with Applications*, Vol. 42.

Han, J. and Camber, M. (2006) *Data Mining: Concepts and Techniques*, 2nd ed.

Harding, J.A., Shahbaz, M., Srinivas and Kusiak, A. (2006) 'Data mining in manufacturing: a review', *Journal of Manufacturing Science and Engineering*, Vol. 128, pp.969–976.

Horng, S-C., Yang, F-Y. and Lin, S-S. (2011) 'Hierarchical fuzzy clustering decision tree for classifying recipes of ion implanter', *Expert Systems with Applications*, Vol. 38.

Jain, A., Nandakumar, K. and Ross, A. (2005) 'Score normalization in multimodal biometric systems', *Pattern Recognition*, Vol. 38.

Jain, Y.K. and Bhandare, S.K. (2011) 'Min max normalization based data perturbation method for privacy protection', *International Journal of Computer &communication Technology*, Vol. 2.

Jeswiet, J., Archibald, J., Thorley, U. and De Souza, E. (2015) 'Energy use in premanufacture (Mining)', *Procedia CIRP*, Vol. 29.

Jurgens, D., Krosche, M. and Niekamp, R. (2012) 'A process for stochastic material analysis based on empirical data', *Technische Mechanic*, Vol. 32, pp.303–306.

Kantar, E. and Keskin, M. (2013) 'The relationships between electricity consumption and GDP in Asian countries, using hierarchical structure methods', *Physica A: Statistical Mechanics and Its Applications*, Vol. 392, pp.5678–5684.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y. (2002) 'An efficient K-Means clustering algorithm: analysis and implementation', *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 24, pp.881–892.

Kirlova, M.G., Hadjikov, L.M. and Stoyan (2009) 'Mathematical modeling of the visco-elastic properties of human abdominal fascia', *11th National Congress on Theoretical and Applied Mechanics*.

Köhler, J., DalsgaardSørensen, J. and Michael (2007) 'Probabilistic model code for design of timber structures', *Structural Safety*, Vol. 29, pp.255–267.

Krishnakumari, P. and Vivekanandan, K. (2009) 'Semi supervised ensemble clustering algorithm for high dimensional genomic data', *Int. J. Rapid Manufacturing*, Vol. 1, pp.222–236.

Maurizio, M. (2011) *Data Mining Concepts and Techniques*, 3rd ed.

Michael, V.K. and Mital (1998) *Probabilistic modeling of high-Temperature Material Properties of a 5-Harness 0/90 Sylramic Fiber/CVI-SiC/MI-SiC woven composite*, NASA/TM-20849.

Monsia, M.D. (2012) 'A mathematical model for predicting the nonlinear deformation response of viscoelastic materials', *International Journal of Applied Mathematics and Mechanics*, Vol. 8.

Murtagh, F. and Contreras, P. (2011) 'Methods of hierarchical clustering', *Computer Science*, Vol. 1.

Nie, G., Zhang, L., Liu, Y., Zheng, X. and Shi, Y. (2009) 'Decision analysis of data mining project based on Bayesian risk', *Expert Systems with Applications*, Vol. 36, pp.4589–4594.

Ronowicz, J., Thommes, M., Kleinebudde, P. and Krysiński, J. (2015) 'A data mining approach to optimize pellets manufacturing process based on a decision tree algorithm', *European Journal of Pharmaceutical Sciences*, Vol. 73, pp.44–48.

Salem, S.A. (2007) *Data Clustering and Partial Supervision with Some Parallel Developments*, University of Liverpool for the Degree of Doctor of Philosophy.

Sander, J., Qin, X., Lu, Z., Niu, N. and Kovarsky, A. (2003) 'Automatic extraction of clusters from hierarchical clustering representations', *The 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp.75–87.

Shi, B., Liang, J. and Liu, Q. (2011) 'Adaptive simplification of point cloud using K-means clustering', *Computer-Aided Design*, Vol. 43, pp.910–922.

Tibshirani, R. (2013) *Clustering 2: Hierarchal Clustering*, Vol. 36.

Tsai, T. (2012) 'Development of a soldering quality classifier system using a hybrid data mining approach', *Expert Systems with Applications*, Vol. 39, pp.5727–5738.

Willett, P. (1982) 'A comparison of some hierarchal agglomerative clustering algorithms for structure – property correlation', *Analytica Chimica Acta*, Vol. 136, pp.29–37.

Yiakopoulos, C.T., Gryllias, K.C. and Antoniadis, I.A. (2011) 'Rolling element bearing fault detection in industrial environments based on a K-means clustering approach', *Expert Systems with Applications*, Vol. 38.

Website

The material web, http://www.matweb.com