CHEMICAL ENGINEERING TRANSACTIONS
VOL. 43, 2015
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Chief Editors: Sauro Pierucci, Jiří J. Klemeš

ISBN 978-88-95608-34-1; **ISSN** 2283-9216
DOI: 10.3303/CET1543205

Ensemble Clustering for Fault Diagnosis in Industrial Plants

Sameer Al-Dahidi[a], Francesco Di Maio[a]*, Piero Baraldi[a], and Enrico Zio[a,b]

[a]Energy Department, Politecnico di Milano, Milan, Italy
[b]Chair on Systems Science and the Energetic challenge, Fondation EDF, Centrale Supélec, Paris, France
francesco.dimaio@polimi.it

In this paper, we propose an unsupervised ensemble clustering approach for fault diagnosis in industrial plants. The basic idea is to combine multiple base clusterings of operational transients of industrial equipment, when the number of clusters in the final ensemble clustering (*P**) is unknown. In practice, a Cluster-based Similarity Partitioning Algorithm (*CSPA*) is employed to quantify the co-association matrix that describes the similarity among the different base clusterings and, then, a Spectral Clustering technique embedding an unsupervised *K*-Means algorithm is used to find the optimum number of clusters of *P** based on Silhouette validity index calculation. The identified clusters allow distinguishing different operational behaviors of the equipment. The proposed approach is verified with respect to an artificial case study representative of the signal trend behavior of an industrial equipment during shut-down operations. The obtained results have been compared with those achieved by a state-of-art approach, known as Cluster-based Similarity Partitioning and Serial Graph Partitioning and Fill-reducing Matrix Ordering Algorithms (*CSPA-METIS*): the results show that the novel approach is able to identify the final ensemble clustering with a lower misclassification rate than the *CSPA-METIS* approach.

1. Introduction

In industries such as chemical, oil and gas, and nuclear, equipment is subjected to several causes of performance degradation (e.g., presence of manufacturing defects, wear and tear) that lead systems to work in anomalous conditions (Baraldi et al. 2013a). Capturing the different operational conditions of this equipment, detecting the onset of abnormal conditions and classifying them in different types can aid the decision maker to decide a proper maintenance intervention and, hence, increase equipment availability and system safety, while reducing overall corrective maintenance costs (Piccinini and Demichela, 2008; Al-Dahidi et al. 2014; Demichela and Camuncoli, 2014).
Fault diagnosis aims at partitioning the collected data representative of different operational conditions of the equipment into dissimilar groups (whose number may be “a priori” unknown) such that data belonging to the same group are more similar than those belonging to the other groups. Once the groups are identified, one can distinguish, among these, anomalous behaviors of the equipment (Baraldi et al. 2013a).
In this paper, we consider the practical case in which the number of groups is a priori unknown and formulate the problem as an unsupervised classification problem aimed at partitioning the data into homogeneous clusters so that those data belonging to the same cluster are very similar to each other and dissimilar to those of the other clusters (Salvador, 2002).
Several clustering algorithms have been proposed and practically used to solve unsupervised classification problems, for example *K*-Means (Liao and Bolt, 2002), Self-Organizing Maps (*SOM*) (Al-Dahidi, 2014), Fuzzy *C*-Means (*FCM*) (Di Maio et al. 2011; Di Maio et al. 2012; Baraldi et al. 2013a), Classification Tree (Baraldi et al. 2012) and Spectral clustering (Von Luxburg, 2007; Baraldi et al. 2013a; Baraldi et al. 2013b; Baraldi et al. 2014). However, there is no unique clustering algorithm capable of correctly identifying the underlying structure of any kind of dataset. Even the application of different clustering algorithms to the same set of data, or of the same algorithm with different parameter settings, leads to different clustering results (Fred and Jain, 2005; Fern and Lin, 2008; Vega-Pons and Ruiz-Shulcloper, 2011). Without any prior knowledge on the

Please cite this article as: Al-Dahidi S., Di Maio F., Baraldi P., Zio E., 2015, Ensemble clustering for fault diagnosis in industrial plants, Chemical Engineering Transactions, 43, 1225-1230 DOI: 10.3303/CET1543205

underlying structure of the dataset, one can only say that the different results obtained are equally plausible (Vega-Pons and Ruiz-Shulcloper, 2011).
To proceed with this uncertainty in the clustering model, ensemble approaches have been proposed (Strehl and Ghosh, 2002). A typical ensemble clustering scheme is shown in Figure 1. For a given dataset $\overline{\overline{X}}$, ensemble clustering usually entails the collection of the results from multiple base clusterings. The base clusterings composing the ensemble can be different because of the different clustering algorithm used or because of the different data features extracted from $\overline{\overline{X}}$ upon which (some) clustering algorithm is applied. The outcome of the individual base clusterings are, then, aggregated into the final ensemble clustering *P** by a given method of aggregation (Strehl and Ghosh, 2002; Topchy et al. 2005; Chen, 2007; Vega-Pons and Ruiz-Shulcloper, 2011).

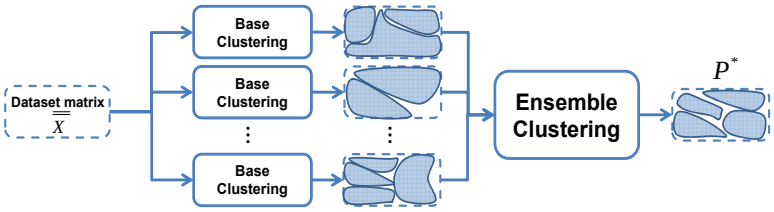

Figure 1: Scheme of ensemble clustering approach

Several methods have been used to obtain the final ensemble clustering. For example Graph and Hypergraph partitioning algorithms, such as the Cluster-based Similarity Partitioning (*CSPA*), construct a graph from the similarities among the base clusterings, and cluster it using a graphic-based clustering algorithm such as Serial Graph Partitioning and Fill-reducing Matrix Ordering Algorithm (*METIS*) (Karypis and Kumar, 1995; Strehl and Ghosh, 2002), for a predetermined number of clusters *M* in the final ensemble clustering (Topchy et al. 2005). In this paper, *CSPA* and *METIS* algorithms have been taken as reference for comparison because *CSPA-METIS* is the simplest and often best performing method for ensemble aggregation (Strehl and Ghosh, 2002).
However, for most industrial applications, the "a priori" knowledge of the number of clusters *M* to be found in the final clustering is rarely available (Chakaravathy and Ghosh, 1996; Strehl and Ghosh, 2002). For this reason, an objective of the present work is to develop a novel unsupervised ensemble clustering approach capable of identifying the final ensemble clustering without any previous knowledge of *M*.
To this aim, we propose to replace *METIS* algorithm with spectral clustering (Von Luxburg, 2007; Baraldi et al. 2013a) and Silhouette validity index (Rousseeuw, 1987) to automatically determine *M*. Practically, the base clustering results are summarized in a co-association matrix by pairwise similarity computation. Then, a spectral clustering technique (Von Luxburg, 2007; Baraldi et al. 2013a), embedding the unsupervised *K*-Means algorithm, is fed by the similarity matrix values for mining the clusters that are formed by the most similar data. The optimum number of clusters *C** is selected among several candidates $C_{candidate}$, based on the morphology of the obtained final ensemble clusters, evaluated by the Silhouette validity index that measures the similarity of the data belonging to the same cluster and the dissimilarity of these in the other clusters (a large silhouette value indicates that the obtained clusters of the final ensemble clustering are well separated and compacted (Rousseeuw, 1987)). The proposed approach is tested with respect to an artificial case study representative of the signal trend behavior of industrial equipment during shut-down transients. The results obtained have been compared with those achieved by *CSPA-METIS* approach.
The remaining of this paper is organized as follows. In Section 2, the novel unsupervised ensemble clustering approach is proposed. The artificial case study representative of the signal trend behavior of industrial equipment during shut-down transients is introduced in Section 3. Furthermore, the results obtained with the application of the novel approach to the artificial case and the comparison with *CSPA-METIS*, are discussed in Section 4. Finally, Section 5 concludes the paper with some considerations.

2. The novel unsupervised ensemble clustering approach

In this Section, the novel unsupervised ensemble clustering approach is proposed to overcome the need of having "a priori" knowledge of the number of clusters *M* in the final ensemble clustering. The flowchart for the method is sketched in Figure 2. The algorithm goes along the following two phases: a procedure (i.e., *CSPA*) for establishing a similarity matrix $\overline{\overline{S}}$ and a novel ensemble procedure for revealing the "hidden" structure *P** of $\overline{\overline{S}}$ by adopting spectral clustering and Silhouette validity index.

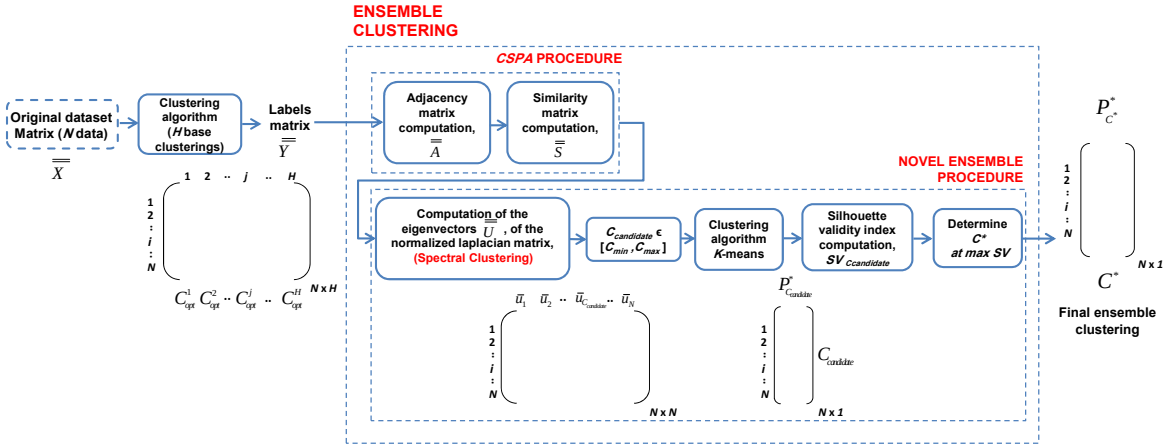

Figure 2: Flowchart of the proposed approach

We consider *N* data belonging to the dataset $\overline{\overline{X}}$ that are clustered into *H* base clusterings. For each *j*-th base clustering, *j*=1,…,*H,* each datum is labeled by an integer number ranging in $[1,C_{opt}^{j}]$, where C_{opt}^{j} is the number of clusters for each *j*-th base clustering. The problem of clustering the *N* data is, thus, transformed into an aggregation problem of the base clusterings outcomes $\overline{\overline{Y}}$ of size *N*x*H*.

The algorithm entails five main steps, where the first two are the *CSPA* procedure and the last three are the novel ensemble procedure:

Step 1: *Adjacency matrix computation*. In practice, for each *j*-th base clustering, if two data belong to the same cluster they are considered similar, i.e., similarity μ=1, and if not they are dissimilar, i.e., similarity μ=0. Thus, an adjacency binary similarity matrix, $\overline{\overline{A}}$, is built by aggregating the similarities of the *H* base clusterings (Strehl and Ghosh, 2002).

Step 2: *Similarity matrix computation*. From the adjacency binary similarity matrix, $\overline{\overline{A}}$, the overall similarity matrix $\overline{\overline{S}}$, is computed as the entry-wise average of the *H* base clusterings, i.e., $\overline{\overline{S}} = \frac{1}{H}\overline{\overline{A}}\,\overline{\overline{A}}^{T}$ (Strehl and Ghosh, 2002). In this way, each entry of the similarity matrix has a value in [0,1], which is proportional to how likely a pair of data is, when grouped together.

Step 3: *Spectral Clustering*. Spectral clustering aims at transforming the similarity matrix $\overline{\overline{S}}$ into a normalized laplacian matrix $\overline{\overline{L}}_{rs}$. Then, for the obtained $\overline{\overline{L}}_{rs}$, the eigenvectors $\overline{u}_1,\overline{u}_2,...,\overline{u}_{C_{candidate}},...,\overline{u}_N$ are calculated corresponding to the computed eigenvalues in ascending order $\lambda_1,\lambda_2,...,\lambda_{C_{candidate}},...,\lambda_N$, and are stored in a matrix $\overline{\overline{U}}$ with a size *N*x*N*, where $C_{candidate}$=[C_{min},C_{max}], and C_{min} and C_{max} are the possible minimum and maximum numbers of clusters in the final ensemble clustering, respectively.

Step 4: *Clustering algorithm*. For each candidate number of clusters $C_{candidate}$, the eigenvectors $\overline{u}_1,\overline{u}_2,...,\overline{u}_{C_{candidate}}$ associated to the $C_{candidate}$ smallest eigenvalues of its laplacian matrix $\overline{\overline{L}}_{rs}$ are calculated. In this way, the reduced matrix of $\overline{\overline{U}}$ with a size *N*x$C_{candidate}$ is fed to a clustering algorithm to find the final ensemble clustering $P_{C_{candidate}}^{*}$. In this work, we resort to the *K*-means algorithm as one of the most popular clustering methods (Su and Chou, 2001; Fern and Lin, 2008).

Step 5: *Final ensemble clustering selection*. For each $C_{candidate}$ the obtained ensemble clustering $P_{C_{candidate}}^{*}$ is evaluated by computing its Silhouette validity index $SV_{C_{candidate}}$ (Rousseeuw, 1987). The most appropriate ensemble clustering $P_{C^*}^{*}$ is the one for which the Silhouette reaches a maximum, i.e., clusters are well separated and compacted (Rousseeuw, 1987).

3. Artificial case study

An artificial case study has been designed to generate *N*=149 data representative of the signal trend behavior of an industrial equipment, e.g., a rotating machine, during shut-down operations. Each datum is described by *H*=3 features representative of the equipment condition, e.g., vibration signals, and of the environmental and operational conditions that can influence the equipment behavior, e.g., vacuum and temperature signals. These data are stored in a matrix $\overline{\overline{X}}$ of a size 149x3.

The objective is to reveal the "hidden" structure *P** of the dataset $\overline{\overline{X}}$ by identifying groups of data with similar functional behaviors, representative of different operational conditions of the equipment. Without loss of generality, it is assumed that the operational conditions of the industrial equipment are *M*=7: 1) three classes of normal condition (*NC1*, *NC2*, *NC3*), 2) three classes of abnormal condition (*AC1*, *AC2*, *AC3*), and 3) one class of outliers (i.e., unknown behaviours). The dataset $\overline{\overline{X}}$ is pictorially shown in Figure 3. The values of each *j*-th feature $\overline{\overline{X}}_j$, *j*=1,2,3, for different classes of data have been created by randomly sampling their realization from different univariate distribution functions (1 to 10 in Figure 3) whose combination characterizes the class.

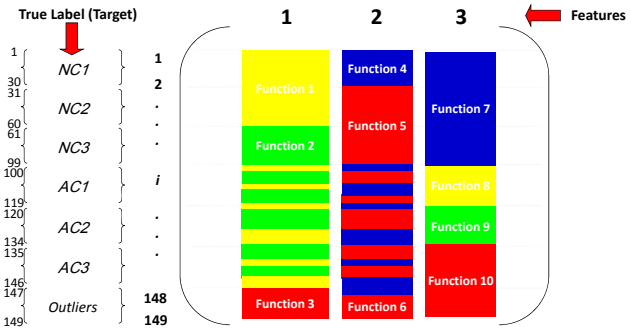

Figure 3: The seven operational conditions of the artificial case study

As shown in Figure 3, clustering each *j*-th feature independently may reveal only some groups of the "hidden" operational conditions of the industrial equipment indicated in Figure 3, whereas only a final ensemble clustering would enlighten all the *M*=7 clusters. For example, clustering the feature *j*=1 cannot reveal any abnormal operational condition. The objective is, thus, to aggregate these base clusterings into a final ensemble clustering *P** capable of identifying the "true" grouping of the shut-down transients of the industrial equipment.

To mine the clusters shown in Figure 3, the *j*-th base clustering outcomes are obtained by a *K*-means unsupervised learning algorithm (Vlachos et al. 2003). For identifying the correct number of clusters C_{opt}^{j} for each base clustering, Davies-Bouldin (*DB*) validity criterion has been used (Davies and Bouldin, 1979): the minimum *DB* value is reached for the number of clusters which gives optimal separation and compactness (Davies and Bouldin, 1979). Table 1 reports the optimum number of clusters C_{opt}^{j} obtained for each base clustering. For validation of the *DB* validity criterion to decide C_{opt}^{j}, we use the information on the real classes to which the data belong, to calculate the misclassification rate (Table 1) (it is worth noticing that in real industrial applications the real class is unknown).

Table 1: Optimum numbers of clusters and misclassification rates of clustering for the three features

Features	C_{opt}^{j}	Misclassification rate
j=1	2	7.4%
j=2	2	4.6%
j=3	4	6.8%

The obtained base clustering labels for each feature have been, then, stored in a matrix $\overline{\overline{Y}}$ of a size 149x3. The application of the clustering ensemble approach aims at finding the final ensemble clustering of the data. In the following Section, our novel approach is applied and compared with *CSPA-METIS* approach.

4. Application of the novel approach to the artificial case study

In this Section, the application of the novel ensemble clustering approach is described according to the steps presented in Section 3 and then, the results obtained are compared to those achieved by the *CSPA-METIS* approach.

Given the similarity matrix $\overline{\overline{S}}$, we calculate $\overline{\overline{L}}_{rs}$ and its eigenvectors $\overline{u}_1, \overline{u}_2, ..., \overline{u}_{C_{candidate}}, ..., \overline{u}_{149}$ and the corresponding eigenvalues $\lambda_1, \lambda_2, ..., \lambda_{C_{candidate}}, ..., \lambda_{149}$. The obtained eigenvectors are stored in the matrix $\overline{\overline{U}}$ with size 149x149. The number *M* of clusters of final ensemble clustering is selected according to the values of Silhouette for different numbers of clusters $C_{candidate}$ that span the interval [2,16], where the lower bound (2) is the minimum number of base clusters (see Table 1), whereas the upper bound (16) is the number of the largest combination of the three base clusters (i.e., 2x2x4): the optimum number of clusters *C** in the final ensemble clustering is the value at which the Silhouette value is maximized, i.e., *C** =6 (star in Figure 4 (left)). Figure 4 (right) shows the results of the aggregation obtained by using the novel and the reference approaches. The Figure shows the *N*=149 data in chronological order from top to bottom with the associated true clustering labels represented in different shades of color. It is worth mentioning that there is no correspondence between the similar colors of the true and the aggregation results. The application of *CSPA-METIS* approach leads us to distinguish clearly only three clusters, i.e., *NC1*, *NC2* and *NC3*, whereas the remaining data have not been correctly clustered. Comparing the obtained clustering results with the true clustering, one can calculate the misclassification rate to be equal to 34.9% (52 out of 149 data incorrectly classified), which is not a satisfactory result. On the other hand, the application of the novel approach leads us to recognize clearly six out of seven operational conditions with a reduced misclassification rate 3.4% (5 out of 149 data incorrectly classified) compared to the *CSPA-METIS* approach.

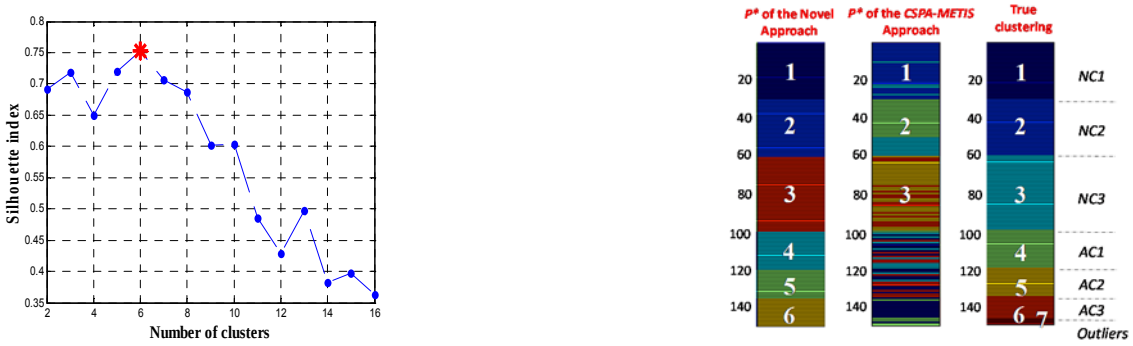

Figure 4: Silhouette values vs. cluster numbers (left) and the obtained final ensemble clusterings obtained by the novel and the reference approaches vs. the true clustering (right)

As last remark, it is worth mentioning that the outliers (three transients – class 7) using the proposed approach have not been grouped together in a separate cluster: this depends on the capability of the base clustering algorithm in recognizing the outliers (Topchy et al. 2005). For example, the optimum number of clusters for the feature *j*=1 is $C_{opt}^{1} = 2$ (see Table 1), whereas it should be equal to 3 (see Figure 3).

5. Conclusions

In this work, a novel unsupervised ensemble clustering approach is proposed to construct a final ensemble clustering *P** from *H* individual base clustering outcomes. The method is based on Spectral Clustering, embedding an unsupervised *K*-Means algorithm that is fed by a pairwise similarity computation, so that a co-association matrix summarizes the similarity among the data and the clusters are formed by the most similar data. The optimum number of clusters is selected among several candidates based on Silhouette validity index that quantifies the morphology of the obtained clusters and gives reason of the similarity of data belonging to the same cluster and, at the same time, of dissimilarity with those in the other clusters: a large silhouette value indicates that the obtained clusters of the final ensemble clustering are well separated and compacted. The proposed approach has been successfully tested with respect to an artificial case study properly designed to reproduce the signal trend behavior of industrial equipment during shut-down transients. The results obtained have been compared to those achieved by the *CSPA-METIS* approach of literature. The continuation of this work will consider the application of the method to real datasets collected during past operation of industrial equipment such as, for example, a nuclear power plant turbine.

Acknowledgements

The participation of Sameer Al-Dahidi and Piero Baraldi to this research is supported by the European Union Project INNovation through Human Factors in risk analysis and management (INNHF, www.innhf.eu) funded by the 7th framework program FP7-PEOPLE-2011- Initial Training Network: Marie-Curie Action. The participation of Enrico Zio to this research is partially supported by the China NSFC under grant number 71231001.

References

Al-Dahidi S., 2014, The Use of Self Organizing Maps for Diagnosing Faults in Motor Bearings, Safety and Reliability: Methodology and Applications - Proceedings of the European Safety and Reliability Conference, ESREL 2014, Wroclaw, Poland, September 2014, 895-902.

Al-Dahidi S., Baraldi P., Di Maio F., Zio E., 2014, A novel fault detection system taking into account uncertainties in the reconstructed signals, Annals of Nuclear Energy, 73, 131–144.

Baraldi P., Di Maio F., Zio E., Sauco S., Droguett E., Magno C., 2012, Sensitivity Analysis of the Scale Deposition on Equipment of Oil Wells Plants, Chemical Engineering Transactions, 26, 327-332.

Baraldi P., Di Maio F., Zio E., 2013a, Unsupervised Clustering for Fault Diagnosis in Nuclear Power Plant Components, International Journal of Computational Intelligence Systems, 6 (4), 764-777.

Baraldi P., Di Maio F., Rigamonti M., Zio E., Seraoui R., 2013b, Transients Analysis of a Nuclear Power Plant Component for Fault Diagnosis, Chemical Engineering Transactions, 33, 895-900.

Baraldi P., Di Maio F., Rigamonti M., Zio E., Seraoui R., 2014, Unsupervised clustering of vibration signals for identifying anomalous conditions in a nuclear turbine, accepted, Journal of Intelligent and Fuzzy Systems.

Chakaravathy S. V., Ghosh J., 1996, Scale based clustering using a radial basis function network, IEEE Transactions on Neural Networks, 2(5), 1250–61.

Chen K., 2007, Trends in neural computation, Springer.

Davies D.L., Bouldin D.W., 1979, A cluster separation measure, IEEE Trans. Pattern Analysis and Machine Intelligence, 1, 224-227.

Demichela, M., Camuncoli, G., 2014, Risk based decision making. Discussion on two methodological milestones, Journal of Loss Prevention in the Process Industries, 28 (1), 01-108.

Di Maio F., Secchi P., Vantini S., Zio E., 2011, Fuzzy C-Means Clustering of Signal Functional Principal Components for Post-Processing Dynamic Scenarios of a Nuclear Power Plant Digital Instrumentation and Control System, IEEE – Transactions on Reliability, 60(2), June 2011, 415-425.

Di Maio F., Hu J., Tse P., Pecht M., Tsui K., Zio E., 2012, Ensemble-approaches for clustering health status of oil sand pumps, Expert Systems with Applications 39, 5 (2012) 4847-4859.

Dimitriadou E., Weingessel A., Homik K., 2001, Voting-merging: an ensemble method for clustering, In Proc. 2001 Int. Conf. Artificial Neural Networks (ICANN'01), 217-224.

Fern X. Z., Lin W., 2008, Cluster ensemble selection, Statistical Analysis and Data Mining, 1(3), 128-141.

Fred A. L., Jain A. K., 2005, Combining multiple clusterings using evidence accumulation, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(6), 835-850.

Karypis G., Kumar V., 1995, METIS - Unstructured Graph Partitioning and Sparse Matrix Ordering System, Version 2.0 (Technical report).

Liao T., Bolt B., 2002, Understanding and projecting the battle state, 23rd Army Science Conf., Orlando, FL.

Piccinini, N., Demichela, M., 2008, Risk based decision-making in plant design, Canadian Journal of Chemical Engineering, 86 (3), pp. 316-322.

Rousseeuw P., 1987, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 53–65.

Salvador A., 2002, Faults diagnosis in industrial processes with a hybrid diagnostic system, In MICAI 2002: Advances in Artificial Intelligence, 536-545. Springer Berlin Heidelberg.

Strehl A., Ghosh J., 2002, Cluster ensembles-a knowledge reuse framework for combining partitionings, In AAAI/IAAI, 93-99.

Su M. C., Chou C. H., 2001, A modified version of the K-means algorithm with a distance based on cluster symmetry, IEEE Transactions on pattern analysis and machine intelligence, 23(6), 674-680. ISO 690.

Topchy A., Jain A. K., Punch W., 2005, Clustering ensembles: Models of consensus and weak partitions, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(12), 1866-1881.

Vega-Pons S., Ruiz-Shulcloper J., 2011, A survey of clustering ensemble algorithms, International Journal of Pattern Recognition and Artificial Intelligence, 25(03), 337-372.

Von Luxburg U., 2007, A Tutorial on Spectral Clustering, Statistics and Computing, 17(4), 395-416.