Clustering Crude Oil Samples Using Swarm Intelligence

F. Ferreira[1], T. Ciodaro[1], J. M. de Seixas[1], G. Xavier[2], and A. Torres[3]

[1] Signal Processing Lab, COPPE/Poli - Federal University of Rio de Janeiro
fferreira,ciodaro,seixas@lps.ufrj.br
[2] PETROBRAS Research & Development Center
gilberto.xavier@petrobras.com.br
[3] FAT - State University of Rio de Janeiro
artorres.uerj@gmail.com

Abstract. The identification patterns in the crude oil intrinsic qualities provides useful information for the refinery operation and logistics. The *a priori* information concerning the characteristics expected by a given crude oil improves the logistic concerning which refineries should process this crude, together with pricing and marketing. This article presents the results of data mining models applied to a generic database of crude oil samples. Only information available in the crude oil intrinsic qualities is used. Clustering techniques based on bio-inspired algorithms are applied to the data samples in order to extract structured patterns from data. Three algorithms were used: PSO, FSS and ABC. Particles and fishes represent the possible clustering solutions. ABC represents solutions as food sources to be evaluated by bees. The silhouette index was used as the fitness function to be optimized. The results were later evaluated using other clustering quality index. The algorithms were able to find patterns beyond the standard oil classification, which considers only the oil density measurement.

Keywords: Oil and Gas, Swarm Clustering, Data Mining, Pattern Recognition

1 Introduction

The oil logistic and refining processes are dependent on the type and characteristics of the crude oil feedstock. In fact, the refinery operation is designed to process crude oils with certain characteristics, followed by the transport logistic to the refinery. Depending on the conditions, different crude oils are blended in order to match the characteristics supported by a given refinery.

Crude oil (a.k.a petroleum) is the world's most important fossil fuel today and has invaluable economical importance. The current amount of exploitable quantities of conventional crude oil is estimated at billions barrels. Although, the proven quantities are huge, oil sources are obviously not inexhaustible. According to the International Energy Agency this will lead to shortness in the exploitable

amounts of crude oils around 2030 and hence to a proportional increase of the price of energy.

This scenario and the growing environmental concern has urged the search for alternative and sustainable energy sources and for a more efficient and 'greener' use of crude oils and related products. As a consequence, several development efforts have been undertaken in order to improve the cracking processes of high density crude oil fractions in order to make viable the processing of such heavy feedstocks as tar sands or less valuable by-products as coke and bitumen. In parallel to these, studies have been developed with the main goal of reducing fossil fuel combustion pollutants like sulfur, vanadium, copper and nickel.

Lastly, a myriad of new characterization techniques have been also developed in order to unveil the details of the crude oil composition as such information is inexorable to the best use of the crude oil. The petroleum molecular composition is as variable as the geologic circumstances during its formation. A large variety of organic chemical compounds, which easily exceed 10^4 different molecular structures, occur in a crude oil and the main fraction is formed by hydrocarbons like alkanes, cycloalkanes and (polycyclic) aromatic compounds.

These facts lead to the conclusion that crude oils can have several different compositions and hence various chemical and physical properties. As a matter of fact, the knowledge of these properties is very important along the whole crude oil supply chain, from the exploitation to the transportation and refining. For that reason, the determination of physical-chemical profiles or so called crude oil assay is common practice in the oil industry.

As such, a crude oil assay is the chemical evaluation of a crude oil feedstock and it comprises relatively simple tests to determine, for instance, its density, viscosity and total sulfur content, as well rather complex ones such as the characterization of the boiling range fractions. Most of these tests are carried out by standard methods developed by the American Society for Testing and Materials International (ASTM International) and the Energy Institute, formely known as the Institute of Petroleum (IP). Nevetherless, most of these methods are rather slow, elaborate and expensive, requiring large volumes of the crude oil to be sampled and sent to the laboratory.

An alternative approach to laboratory tests (a.k.a. wet analysis) is to combine currently available computational intelligence and machine learning methods with the large amounts of data contained in hundreds of even thousands of conventional crude oil assays in order to generate AI-based crude oil assays. Therefore, since there is a great variety of crude oils, a clever approach to develop accurate models is first try to cluster the crude oils by similarity in a few number of groups and then develop bespoke models for each group. Classification of crude oils in groups is an old problem faced by the oil industry professionals and along the years several different crude oil classification methodologies were proposed.

This article presents the results of data mining models applied to a generic database of crude oil properties. Only the most basic intrinsic quality information is used at the moment. Clustering techniques based on bio-inspired algorithms are applied to the data samples in order to extract structured patterns from data.

This article is organized as follows: the Section 2 describes the algorithms that will be used to create the clustering models, while the Section 3 describes the database used to extract the crude oil characteristics. Thereafter, the methodology is presented in the Section 4. Then, the results are shown in the Section 5. Finally, conclusions and possible future studies are derived in Section 6.

2 Swarm Clustering

Clustering data is widely applied in the scientific literature for machine learning and data mining purposes in a large number of different applications. Data clustering models can be interpreted as a compact data representation or as a generative model [16].

Swarm intelligence algorithms have succeed in several pattern recognition tasks, frequently outperforming classical approaches when analyzing large and complex data sets [1, 9, 17]. These algorithms are based on the collective behavior inside a community of individuals interacting locally with each other and with their environment. Swarms use decentralized forms of control and self-organized in order to achieve its goals. The efficiency of natural swarms on complex problems solving inspired the development of computer systems that mimicking their behavior. Several approaches using bio-inspired algorithms were used for clustering problems [2]. Specially, three meta-heuristics have been highlighted in recently published works: Particle Swarm Optimization [15], Artificial Bee Colony [10] and Fish-School Search [6].

2.1 Particle Swarm Optimization (PSO)

PSO is a stochastic optimization algorithm based on the behavior observed in various social groups in nature, such as flocks of birds. The swarm represents the interactions among a number of individuals with no knowledge about the final flock's goal. Instead of it, the individual only knows its current state, it best past state and the neighbor that performs the best [15]. This way, each particle aims to achieve the success of its neighbors and, then, the whole population tends to accumulate on optimum regions of the search-space. Particle are vectors of dimension D located inside the defined search-space which represents the solution of the specified problem.

The most common implementation of this algorithm defines the particle behavior through two equations. The first one adjusts the particle speed and the second one, moves the particle towards it next position.

$$v_{id}^{(t+1)} \leftarrow \alpha v_{id}^{(t)} + U(0,\beta) \times (p_{id} - x_{id}^{(t)}) + U(0,\beta) \times (p_{gd} - x_{id}^{(t)}) \quad (1)$$

$$x_{id}^{(t+1)} \leftarrow x_{id}^{(t)} + v_{id}^{(t+1)} \quad (2)$$

where the particle index is designated by i, dimension by d, the particle's position by x_i, the current particle's speed by v_i, p_i is the best position found by i, α and β denote the inertia weight and the particle acceleration respectively, p_{gd} stands

for the best position found by any particle on the process so far and, finally, $U(0, \beta)$ is a random number generated in every movement.

The evaluation of the solutions vector represented by the particle i is computed by the function $f(x)$. The results are compared and the best position x_i is stored. Each particle surrounds the region centered in the local best position achieved by p_i and p_g. As the algorithm evolves, and the particle positions are updated, trajectories are diverted to new regions of the search-space, converging to an optimal global solution.

The Equation 1 can be rewritten as $v_i^{(t)} = x_i^{(t)} - x_i^{(t-1)}$, that is, the particle speed $v_i(t)$ is computed by the difference between the next position and the current one. Hence, it is possible to merge the two equation into only one:

$$x_{id}^{(t+1)} \leftarrow x_{id}^{(t)} + \alpha \times (x_{id}^{(t)} - x_{id}^{(t-1)}) + \sum U(0, \frac{\Phi}{2}) \times (p_{gd} - x_{id}^{(t)}) \qquad (3)$$

This way, the position update can be interpreted as the sum of three factors: the current position, the particle persistence and the social influence. Therefore, each particle start at it last position, persists towards the last direction and adjusts its trajectory according to a relation among the social influence and its current position.

2.2 Artificial Bee Colony (ABC)

Bees spend a fair part of their lives searching for food sources. Bee colonies have a decentralized food gathering system. This system can be adjusted in order to increase the amount of nectar collected [18].

Bees may appraise the food sources quality by measuring the quantity of nectar found. They also compute the distance between the source and the hive by understanding the amount of energy consumed in their displacement. Both estimations are shared with their companions by performing a waggle dance. Depending on the intensity of the dance forager bee recruit more mates for visiting the region. Bees that decide foraging without any guidance from other bees are called scouts. Therefore, the main idea behind the algorithms based on the bee foraging behavior is that forager bees have a potential solution to an optimization problem in their memory. This potential solution corresponds to the location of a food source and has an aggregated quality measure (objective function solution).

The implementation of the Artificial Bee Colony (ABC) uses three kinds of bees: Employed bees, Onlooker bees and Scout bees. Onlooker bee observes the employed bees to perform the waggle dance and then decide which food source will be visited. The scout bee is responsible by the random search throughout the search-space. For the described tasks, in this algorithm, half of the hive is made by employed bees and half by onlooker ones. For each food source, there is only one employed bee. The employed bees that are responsible by uninteresting sources become scout bees [11].

ABC presents three cycles:

1. Send the employee to the food source and measure the amount of nectar on it;
2. Onlookers choose which source food they will seek after the information provided by the employed bees, calculated by the expression:

$$p_i = \frac{\text{fit}_i}{\sum_{n=1}^{SN} \text{fit}_n} \tag{4}$$

   where SN is the number of food sources equal to the number of employed bees, and fit_i is the fitness of the solution.
3. Recruit scout bee for seeking source foods on unknown regions using the following expression:

$$v_{ij} = z_{ij} + \phi_{ij}(z_{ij} - z_{kj}) \tag{5}$$

   where $k \in 1, 2, ..., SN$ and $j \in 1, 2, ..., D$ are randomly chosen indexes.

The position of the food source represents a possible solution of the optimization problem and the amount of nectar represents the fitness of the solution.

2.3 Fish-School Search (FSS)

Each individual in a fish school has limited memory on the searching process. As the PSO algorithm, each element of a school represents a optima solution. The only information that the fish holds is its own weight and it is the only thing needed to indicate the best solution. In contrast to other algorithms, there is no need of knowing the best position, speed, direction, etc.

This algorithm is computed through three operators. Its execution occurs inside the "aquarium" which denotes the search-space. The available food is the fitness function to be optimized [6]. The operators can be described as follows:

Feed The best regions on the aquarium where the fish can process the search are the ones where there is more food available. This way, fishes in better regions eat more getting fatter as consequence. Fishes can move freely in an independent way and, as results, each fish can gain or lose weight;

Swim a collection of operators (individual movement, collective-instinctive movement and collective-volitive movement) that are responsible for guiding the search process globally towards regions of the aquarium that are collectively sensed by all individual fishes as more promising;

Breeding responsible for refining the search performed.

A second version of the algorithm was proposed in order to simplify the swimming operator [4]. The old version is sensitive to the step value used. Also, the fitness functions is computed twice per fish. This way FSS-II presents some advantages: high exploitation capability, just one fitness evaluation per fish per iteration and easy implementation.

2.4 Clustering

The described algorithms can be used for the cluster discovering process. The particle from PSO, the fish from FSS or the food source in the ABC represent a set of N centroid vectors, where N is provided before the algorithm execution [12, 21, 23]. Any intra-cluster or extra-clustering validity measure can be used as fitness function. The most common option is to compute the intra-cluster distance among the sample, that is, the sum of the distances between every two samples inside a cluster [13]. However, this approach does not guarantee that samples inside a cluster wouldn't be better represented by an external centroid. This way, the Silhouette Index is a better suit for fitness function [19, 20].

Silhouette Index The silhouette index provides a measurement of consistency of the estimated clusters through the following equation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average dissimilarity between the $i-th$ sample from a given cluster and the other samples in the same cluster and $b(i)$ is the lowest average dissimilarity of the $i-th$ sample to the remaining clusters. The above expression can be rewritten as:

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

It can be seen that $-1 < s(i) < 1$. If the $i-th$ sample is well-represented in its cluster, then $s(i)$ is closer to 1. Otherwise, $s(i)$ is closer to -1. Data samples with silhouette values close to zero are in the border of two or more clusters. The average silhouette index, considering all data samples, is used to measure the performance of the clustering configuration.

3 Data set

The database used consists of 49 crude oil simple assays, characterized by 18 physical-chemical properties (see Table 1).

These physical-chemical properties consist of density, viscosity, the content of metals, asphaltene and heteroatoms compounds, acidity and the boiling point range. Some of the described properties denoted measurements at different temperatures, in special, the viscosities. It is readily understandable that these are highly correlated with each other and the same is valid for the true boiling points.

It is also comprehensible that other correlations can be observed among the properties as they are related to the chemical composition. For example, it is not expected density and viscosity being uncorrelated or finding high true boiling

Table 1. Crude oil assay properties used for creating the clustering models.

Property	Property	Property
Density	Nickel	True boiling point at 10%
Asphaltene	Vanadium	True boiling point at 30%
Sulfur	Kinematic viscosity at 20°C	True boiling point at 50%
Total acid number	Kinematic viscosity at 30°C	True boiling point at 70%
Total nitrogen	Kinematic viscosity at 40°C	True boiling point at 90%
Basic nitrogen	Kinematic viscosity at 50°C	Pour point

point values for low density (light) crude oils. Following the same reasoning, the high asphaltene content is associated with heavier crude oils since the *asphalt* constituents of petroleum are *high-molecular-weight* materials with high boiling points [22].

Petroleum density is a important information as it reflects, in average terms, the content of light and heavy fractions of crude oil. In fact, for a long time density has been used as a simple and direct form of crude oil classification, as it gives an indication of the relative proportions of light and heavy fractions. This classification can be seen in Table 2.

Table 2. Density-based crude oil classification [7].

Class	Minimum	Maximum
Light		0.870
Medium	0.870	0.920
Heavy	0.920	1.000

According this classification, based only on the density measurement, the database used in this research consists on 9 light, 22 medium and 18 heavy samples. However, all the intrinsic structure and patterns considering other properties are not taken in account. By disregarding this information and missing important nuances, it may occur that two crude oils, which should be considered separated classes, are characterized together erroneously, leading to a non optimal use of the crude oil as mentioned in the Section 1.

4 Methodology

The designed models are based on unsupervised learning algorithms, meaning that no *a priori* information concerning the data samples is applied.

Before any model is applied, data is normalized, in order to avoid the dominance of some properties over the others due to their dynamic range. While viscosity, for instance, is a property that range from 0 to 10^6, other input variables have a much smaller dynamic range. For that reason, the $\log 10$ function

is applied on viscosity properties. In the sequence, all properties are subtracted from their mean values and divided by their standard deviation, in order to become zero-mean input data with unity variance.

Thereafter, the swarm clustering algorithms were applied. The search-space is defined as a set of N centroids vectors, where N is the number of clusters, varying from 2 to 10. The algorithms' goal was to minimize the $J(x)$ fitness function, defined as:

$$J(x) = 1 - \text{silhouette index} + 10P$$

where P is a non-negative integer representing the number of empty clusters. As the silhouette index ranges from -1 to 1, the fitness function is limited to the range $[0, 10N + 2]$.

In total, four PSO variants described in its original paper [14] were used plus other two others described in [5] were used. The Table 3 shows all the variants and how they will referenced in the rest of this work. For all variations were used $\alpha = 0.7298$ and $\beta = \Phi/2$, where $\Phi = 2.9922$, as it was used in [5]. It was used 100 particles and 100 interactions.

Table 3. PSO variants used on this work

Variant	Description
PSO_1	PSO canonical (with inertia weight)
PSO_2	PSO canonical (with inertia weight and equal random weights of social and cognitive components)
PSO_3	PSO variant (with inertia weight same random number for all components.)
PSO_4	PSO variant (with inertia weight same random number for all components and equal weights of social and cognitive components
PSO_5	PSO canonical (with constriction factor)
PSO_6	Fully Informed Particle Swarm (FIPS)

ABC algorithm used 50 onlooker bees and 50 scout bees, during 100 interactions. It was also set the limit number of tries after which food source is dropped if not improved to 20 tries.

Concerning the algorithm based on fish-school behavior, the FSS-II was used. 100 fishes were used, with initial weights varying between 300 and 600. Both α and β were set to 0.1. The step value was also set to 0.1.

Finally, the consistency of the estimated clusters were measured using the silhouette index and the Clustering Validation index based on Nearest Neighbors (CVNN) [3]. Different from the most existing measures, the CVNN evaluates the inter-cluster separation based on objects that carry the geometrical information of each cluster. Sharing the same idea with K Nearest Neighbor consistency [3], CVNN uses dynamic multiple objects as representatives for different clusters in different situations when measuring the inter-cluster separation. If an object is located in the center of a cluster and is surrounded by objects in the same cluster, it is well separated from other clusters and thus contributes little to the inter-cluster separation. If an object is located at the edge of a cluster and is surrounded mostly by objects from other clusters, it connects to other clusters tightly and thus contributes a lot to the inter-cluster separation. CVNN

also employs the average pairwise distance between objects in the same cluster as the measurement of intra-cluster compactness. Finally, the CVNN index takes a form of the summation of the inter-cluster separation and the intra-cluster compactness after the normalization for both of them.

All the algorithms were ran inside the Pygmo parallelized framework provided by the European Space Agency [8].

5 Results

From 2 to 10 clusters have been estimated for each proposed approach. It is important to notice the models using only two clusters were created only for the matter of comparison as the samples used in this work are usually separated in at least three groups when using classic classification methods.

On average, the best results were obtained using canonical PSO with inertia weight and equal social and cognitive components on the model using 3 clusters as can be seen in the Figure 1.

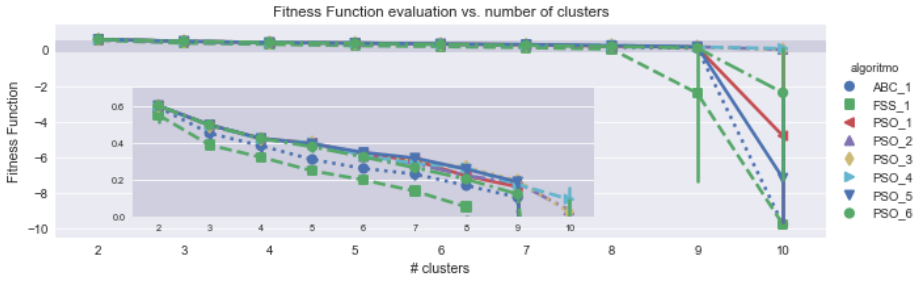

Fig. 1. Performance for the different algorithms according the to fitness function $J(x)$.

Figure 2 shows that all three approaches have some samples that are poorly represented by their clusters, as their silhouette index are negative.

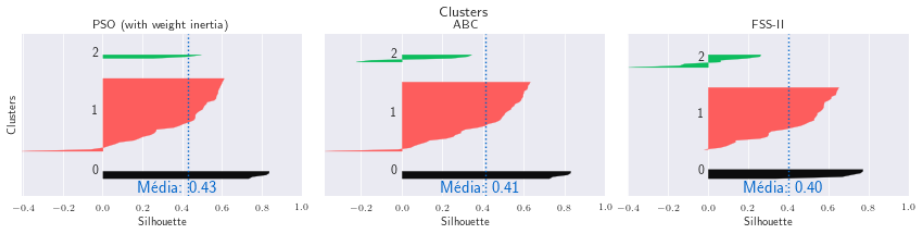

Fig. 2. The silhouette index performance for each sample for the best extracted models.

The average silhouette index for the PSO approach is the highest. This model seems to have achieved the best data sample representation in the clustering process. It can also be seen that the density of the clusters are more similar, when compared to other approaches, where cluster C0 and C2 has much less elements than cluster C1.

Figure 3 shows the cumulative distribution function (CDF) for the clusters of each approach, considering the density value. The density-based classification is also shown, which defines three classes: *Light*, *Medium* and *Heavy*. Notice that some of the clusters cover more than one class from the density-based classification (each amount is estimated from the CDF), shown as the shaded regions below and above the CDF curves. This behavior evinces that intrinsic patterns are laid aside when grouping only by the density information. For example, the FSS algorithm identified the oil with greater viscosity values as the heaviest ones, but not necessary the most dense ones.

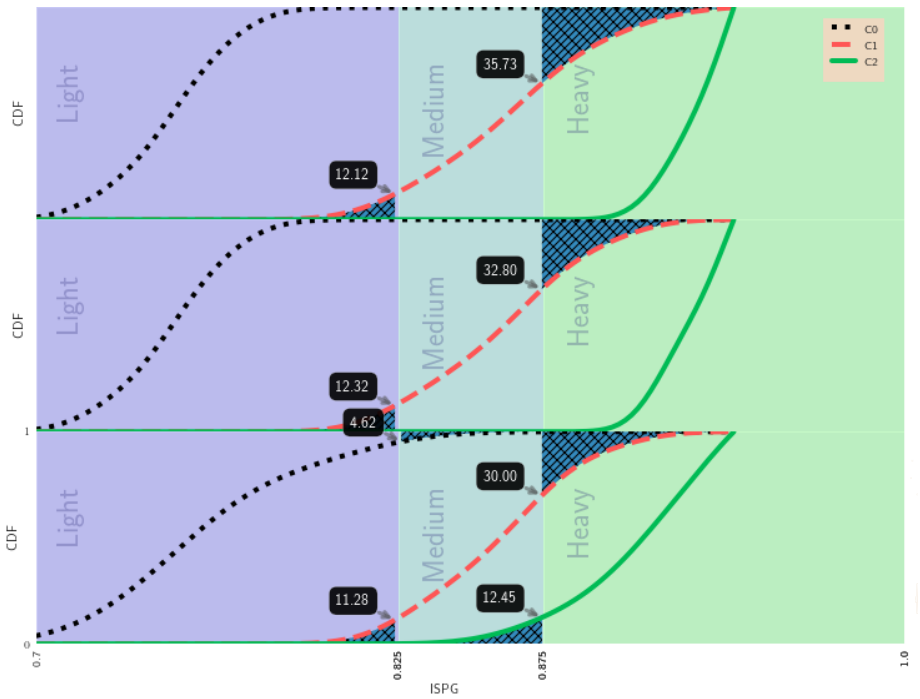

Fig. 3. CDF curves and the typical density-based classification over the density for the clusters found considering the PSO, ABC and FSS approaches.

6 Conclusion

This study used unsupervised data mining techniques in order to investigate intrinsic structures and patterns considering the basic physical-chemical properties of crude oil samples. These structures and patterns might be important for planning purposes in the allocation of crude oils among different refineries.

Clustering techniques based on bio-inspired algorithms are applied to the data samples in order to extract structured patterns from data. Three algorithms were used: PSO, FSS and ABC. The fitness function was based on the silhouette index. The algorithm based on particles and bees had greater clustering validation values compared to the fish. Then, the results have been compared to the standard density-based classification and it was shown that other properties actually contributed for clustering these samples.

All algorithms pointed out the best models using 3 clusters, however, results were not adherent to the typical classification. In other words, several crude oils weren't classified in the same way it is when only considering the density measurements. Therefore, by using intelligence computing approach it is possible to profile crude oils in a more complete manner and this results in a better use of them.

Acknowledge

The authors would like to thank PETROBRAS, FAPERJ, CAPES and CNPq (Brazil) for their support during this work.

References

1. Ab Wahab, M.N., Nefti-Meziani, S., Atyabi, A.: A comprehensive review of swarm optimization algorithms. PLoS ONE 10(5) (2015)
2. Abraham, A., Das, S., Roy, S.: Swarm intelligence algorithms for data clustering. Soft Computing for Knowledge Discovery and Data Mining pp. 279–313 (2008)
3. et al, Y.L.: Understanding and enhancement of internal clustering validation measures. IEEE Transactions on Cybernetics (2013)
4. Bastos-Filho, C.J.A., Nascimento, D.O.: An enhanced fish school search algorithm. Proceedings - 1st BRICS Countries Congress on Computational Intelligence, BRICS-CCI 2013 (2), 152–157 (2013)
5. Blackwell, T.M., Kennedy, J., Poli, R., et al.: Particle swarm optimization. Swarm intelligence 1(1), 33–57 (2007)
6. Filho, C.J.a.B., Neto, F.B.D.L., Lins, A.J.C.C., Nascimento, A.I.S., Lima, M.P.: A novel search algorithm based on fish school behavior. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics pp. 2646–2651 (2008)
7. Groysman, A.: Corrosion Problems and Solutions in Oil Refining and Petrochemical Industry. Springer International Publishing (2017)
8. Izzo, D.: PyGMO and PyKEP: open source tools for massively parallel optimization in astrodynamics (the case of interplanetary trajectory optimization) (2012)

9. Kar, A.K.: Bio inspired computing - A review of algorithms and scope of applications. Expert Systems with Applications 59, 20–32 (2016)
10. Karaboga, D., Akay, B.: A survey: algorithms simulating bee swarm intelligence. Artificial Intelligence Review 31(1-4), 61–85 (2009)
11. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of Global Optimization 39(3), 459–471 (2007)
12. Karaboga, D., Ozturk, C.: A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied Soft Computing 11(1), 652–657 (2011)
13. Karthi, R., Arumugam, S., Rameshkumar, K.: Comparative evaluation of Particle Swarm Optimization Algorithms for Data Clustering using real world data sets. 8(1) (2008)
14. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. Engineering and Technology pp. 1942–1948 (1995)
15. Kennedy, James: Encyclopedia of Machine Learning, chap. Particle Swarm Optimization, pp. 760–766. Springer US, Boston, MA (2010)
16. Kumar, V. (ed.): Data Clustering Algorithms and Applications. Taylor & Francis Group, LLC, Minneapolis, Minnesota, E.U.A (2014)
17. Martens, D., Baesens, B., Fawcett, T.: Editorial survey: Swarm intelligence for data mining (2011)
18. Parpinelli, R., Lopes, H.: New inspirations in swarm intelligence: a survey. International Journal of Bio-Inspired Computation 3(1), 1 (2011)
19. Petrovic, S.: A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In: Proceedings of the 11th Nordic Workshop of Secure IT Systems. pp. 53–64 (2006)
20. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Computational and Applied Mathematics 20, 53–65 (1987)
21. Serapião, A.B., Corrêa, G.S., Gonçalves, F.B., Carvalho, V.O.: Combining K-Means and K-Harmonic with Fish School Search Algorithm for data clustering task on graphics processing units. Applied Soft Computing 41, 290–304 (2016)
22. Speight, J.G.: The chemistry and technology of petroleum. CRC Press (2014)
23. Van Der Merwe, D., Engelbrecht, A.: Data Clustering using Particle Swarm Optimization. IEEE pp. 215–220 (2003)