Data Mining on Data of Catalytic Cracking Microactivity Reactors Using PCEM

Benjamin Moreno-Montiel, Carlos-Hiram Moreno-Montiel, Miriam-Noemi Moreno-Montiel, and René MacKinney-Romero

***Abstract*—Crude oil can have great uses and applications, to achieve this it must undergo a process of conversion of primary to secondary energy called refining. Refining is the set of processes that are applied to crude oil in order to separate its useful components and adapt its characteristics to the needs of society. Among these products obtained from the refining process is gasoline, which is obtained using various types of catalysts. In this paper, we propose to use the Parallel System of Classification based on the Ensemble of Mixture of Experts (PCEM) developed in C using MPI (Message Passing Interface) that guarantees the obtaining of results that reflect the performance of a set of evaluated catalysts and thus proceed to the election of one that meets the industrial requirements of this process or propose improvements to this based on their behavior in the process. To carry out this system, it is proposed to use the Data Mining process on a repository of data obtained from a Catalytic Cracking Microactivity Reactor. Within the process of Data Mining is the task of classification of data, which was selected to be the engine of operation of the system proposed in this paper. We implemented a series of classifiers to compare the operation of the PCEM, that can predict new data between three different types of gasoline grades, obtaining in all the tests that the PCEM high rates in the performance measures with respect to the traditional classifiers.**

***Index Terms*—Catalysts, catalytic cracking, classification, classifiers, computational simulations, data mining, computational simulations, microactivity reactor, refining.**

I. Introduction

Throughout the years, Mexico has consolidated as a major oil country, with this resource contributing approximately one third of the federal government's expenditure. With a daily production of 2.5 million barrels of oil by 2019, Mexico ranks 10th worldwide. Undoubtedly one of the arms of the oil industry with greater relevance due to its function and objectives is refining, since it oversees converting crude oil into commercially usable products.

In Mexico, the national refining system (NRS) has 6 refineries in operation, with a total processing capacity of 1,199,000 barrels of crude oil per day, which positions PEMEX as the 13th company worldwide [1].

One of the processes that acquires special relevance within refining and that has been evolving since its creation is the catalytic fluidized bed disintegration also known as Fluid Catalytic Cracking (FCC).

The FCC process generates high octane gas by contributing up to 40% of the gasoline in the refinery pool [2], depending on the orientation of its operation. The importance of this process lies in its great flexibility to operate with various types of cargo to transform high molecular weight hydrocarbons into smaller chains of greater commercial value; consequently, any small improvement or benefit in this process is very profitable.

For all the above the FCC process has been called as the "heart" of a modern refinery and has become the process with more research within the refining train [2]. Therefore, this paper is oriented towards this process of analyzing a considerable number of datasets to create patterns of behavior that allow us to propose alternatives in the characteristics of the catalyst or in the same operating conditions to optimize the process [2]. Patterns based on data classification will be particularly focused on the analysis of process variables, such as the types of load, types of catalyst used, product yield and surface phenomena (ie the adsorption, desorption and reaction of hydrocarbons in the active sites) that occur in the catalyst during the catalytic decay reactions.

Currently the Data Classification process is used in many areas of research such as Statistics, Bioinformatics, Medicine, Finance, Physics, Archeology, Criminology and Chemical Engineering to mention just a few of them. It is precisely in this last area of research that this project proposal arises, since there are currently large repositories of data which require analysis processes to search for behaviors that allow researchers to have greater control over the large number of processes that are simulated in this area of research.

In the Machine Learning there are many Classifiers, which allow us to perform data classification, among the most popular we have K - Nearest Neighbors, Naive Bayes and Decision Trees. These classifiers are the most used, since they offer good results for many models. However, when handling large amounts of data (more than 300,000 records), the performance of these classifiers tends to worsen the indexes of the operating averages, as well as increase the execution times.

The classifiers based on ensembles seem to be a good option for classification of large amounts of data, however when classification is performed on a large database, over 300,000 records or huge database, over 1,000,000 records, the size becomes a problem for these classifiers. In a previous

Manuscript received May 28, 2019; revised August 14, 2019.

Benjamin Moreno-Montiel and René MacKinney-Romero are with the Universidad Autónoma Metropolitana – Unidad Iztapalapa, Departamento de Ingeniería Eléctrica, Mexico (e-mail: bmm@xanum.uam.mx, rene@xanum.uam.mx).

Carlos-Hiram Moreno-Montiel, is with Universidad Tecnológica de México – Plantel Sur, Ingeniería en Sistemas Computacionales, Mexico (e-mail: hiramoreno@gmail.com).

Miriam Moreno-Montiel is with Instituto Politécnico Nacional, Departamento de Ingeniería Química Petrolera – ESIQIE, Mexico (e-mail: mimorenom@ipn.mx).

doi: 10.18178/ijesd.2019.10.11.1203

work we developed an ensemble model system called Hybrid Classifier with Genetic Weighting – HCGW [3]. This classifier uses a mixture of expert to construct the ensemble, using a genetic algorithm to assign the weights to weak learners, we could say that this classifier is the predecessor of system presented in this paper, as ID3 with respect to C4.5.

Although with HCGW we obtained better performance measures, it comes at a steep price (30 % increase on execution time). We can see that the main limitation of the ensembles is the execution time, which is approximately equal to sum of execution time of the weak learners (WeLe's) used. One possible solution for this issue is to apply parallel computing to reduce the execution time when used the set of WeLe's [3].

The main objective of this work is to show the use of the Parallel System of Classification based on an Ensemble of Mixture of Experts – PCEM [4], on a set of data obtained from a Catalytic Cracking Microactivity Reactor, obtaining higher rates of performance and low execution times on large amounts of data. PCEM developed using parallel schemes and a set of classifiers whit a parallel genetic algorithm to assign weights to each classifier.

In PCEM a weighted voting criterion is used, in which a weight is assigned to each WeLe's by a parallel genetic algorithm. With the parallel genetic algorithm, the best combination of weights is searched, thus obtaining higher rates in performance measures against a traditional approach.

In addition, PCEM handles a parallel scheme based on the MIMD architecture (Multiple Instruction and Multiple Data Stream); it implements parallel schemes for each WeLe and also the genetic algorithm. Through this parallel scheme we were able to handle large amounts of data for classification; obtaining low execution times in each of the tests we performed with PCEM.

A parallel process of data mining will be carried out on a data repository of a Catalytic Cracking Microactivity Reactor to obtain patterns, a set of attributes that define a special behavior, that will bring benefits to the simulation process. Considering that the classifiers present problems in the means of operation and high execution times when classifying large amounts of data, the decision was made to use the PCEM, given the advantages it offers with respect to traditional classification schemes.

This paper is organized as follows: in Section II we will discuss previous work on classifiers based on ensemble, also some concepts on chemical engineering. In Section III we describe in detail each component of the PCEM. In Section IV we will show how the data of the Catalytic Cracking Microactivity Reactor were transformed to apply the data classification on these. In Section V we present the results of our tests, compared the PCEM with other classifiers. Finally, we will present some conclusions and future work.

.

II. Previous Work

A. Introduction

In this section we review some of the work reported in the literature on classifiers based on ensembles (CBE), these are Bagging and Boosting, and we review the two ensembles which inspired this work, which are Stacked Generalization and Mixture of Experts. For parallel schemes of traditional classifiers, we review the most relevant for this work.

Also, in this section we will review some important concepts of Chemical Engineering such as: the refinery process, the properties of gasoline and the different processes for obtaining gasolines. We will finish this section by reviewing each of the properties present in the different types of gasolines, which will help us understand what was the crude state of these data that was used to develop this work, which will be reviewed in more detail in Section 3.

B. Classifiers Based on Ensembles

As we mentioned earlier, we describe the four types of CBE; Bagging, Boosting, Staked Generalization [5], and Mixture of Experts [6]

The ensemble Bagging was introduced by Breiman [7], integrating the concept of reinforced aggregation. The implementation of this method is simple, the ensemble consists of choosing a single type of classifier (usually decision trees are chosen), dividing the set of training patterns to generate disjoint subsets of a single type of classifier. We generate n decision trees for each of the subsets of the training set. Subsequently once they have different subclassifies, they classify the test set to obtain a set of individual classifications. Once the system gets the individual classification for the database, these are combined using a majority voting criterion to obtain the final classes of the test set.

In the 90's this classifier based on ensembles was developed by students of Shapire [8]. In this classifier they proved that if they selected a group of weak learners, training them with different training sets and combining with an appropriate criterion their results, they could generate strong learners. Thus, obtaining the Boosting algorithm, considered one of the seminal classification algorithms.

For such classifiers based on ensembles, usually different models of decision trees or neural networks are generated; in this case we only described the case for decision trees. The operation begins with the generation of the decision trees using different subsets of the training set. In contrast to Bagging, the construction of trees is not performed individually, for Boosting whenever we generate a tree i, this will provide the information about hardly classified examples (HCE) and the examples that were easily classified.

Once the tree i is constructed, a new tree takes the HCE to improve the accuracy of the new tree, so on and so forth to generate a defined number of trees. This procedure is performed iteratively until a stopping criterion used. Once the first phase of Boosting operation is concluded, it proceeds to assign a series of weights for each tree generated according to the error obtained in the previous phase. Finally, once each tree obtains individual classifications, we proceed to implement a weighted voting criterion, to get the individual classification of the test set.

For the case of Staked Generalization, the main difference to Bagging and Boosting, is that in this ensemble different types of classifiers are used, choosing the class using a majority voting criterion. Stacked Generalization was introduced by Wolpert, the general method using a set of classifiers denoted by $C1$, $C2$, $C3$, … , CT which are trained first, so that an individual classification for each of them is obtained, which are called the First Level Base Classifiers.

After obtaining these individual classifications, a majority voting criterion is selected, thus constructing the final classifier, this phase is called Second Level Meta Classifier. One example of this type of ensemble is the work of Sun [9], in this paper the Autor propose a CBE of Stacked Generalization (CBE-SG), using local within-class accuracies for weighting individual classifiers to fuse them.

Where distance metric learning is adopted to search for within-class nearest neighbours, called W-LWCA. In this ensemble more than two types of classifiers which are combined using a weighted voting criterion are applied directly to the training set, to create better learnings. In the tests conducted with the UC Irvine repository in some cases the method proposed in this paper, shows improvements over ensembles.

For the case of Mixture of Experts [10], is like Stacked Generalization; it considers a set of classifiers denoted by $C1, C2, C3, \ldots, CT$, to perform first level base classifiers, later a classifier CT+1 combines the individual classifications of each one obtaining the final classification.

This model considers a phase in which weights are assigned to each classifier Ci, $\forall\ i = 1,2, \ldots, T$, to finally apply a criterion of weighted majority voting. Usually this part of the model is performed by a neural network, called the gating network. One difference with Stacked Generalization, since the voting criteria in mixture of experts is a weighted voting criterion uses a neural network to assign the weights of each parallel classifier considered.

One example of this ensemble is Hybrid Classifier with Genetic Weighting (HCGW). HCGW considered a set of different classifiers and weighted voting criteria to find the better way to combine the individual classification of each classifiers [3]. To find the better form to weigh each classifier, in the HCGW a genetic algorithm was used.

With this algorithm we realize that ensemble classifiers are a good option to classifying large amounts of data. After an analysis of execution times an issue was detected in the HCGW, which is that the running time is approximately equal to the sum of the individual times of classifiers, performed the classification of a datasets with 300 thousand records in a time of 2 hours and 16 minutes on average, one of the main problems of this ensemble.

C. Oil Refinery Process

The oil refinery process begins with the exploration and production phase by drilling onshore and offshore, extracting crude oil with an infrastructure of 30,000 wells and 300 platforms. Once the crude oil is extracted, the logistics phase transports it through a pipeline system of more than 5,000 km. to one of the six existing refineries in Mexico.

At this moment the industrial transformation phase processes the crude to obtain gasolines, diesel and other petroleum and petrochemical products. Subsequently, the fuels are transported to the 77 Storage and Distribution Terminals (SDT), through a fleet of 525 tank cars and the network of pipelines of more than 8,800 km.

When arriving at the TADs, the gasolines are added with an exclusive active detergent dispersant to achieve an optimum quality. At the end of the process, they are distributed to all Pemex gas stations through 1,485 auto-tanks. Once at Pemex gas stations, gasoline and diesel are monitored in quality and quantity of dispatch through 77 mobile laboratories, in Fig. 1 shows the scheme oil refinery process.

Fig. 1. The scheme oil refinery process.

Gasoline is a liquid composed of a mixture of hydrocarbons, obtained in the process of refining oil, whose main function is as fuel. Hydrocarbons are a family of chemical compounds that, as their name suggests, are composed of hydrogen (H) and carbon (C) exclusively. They are the basic compounds of petroleum, in which we find a great variety of them, according to the number and spatial organization of their carbons [11], this we can see in Fig. 2.

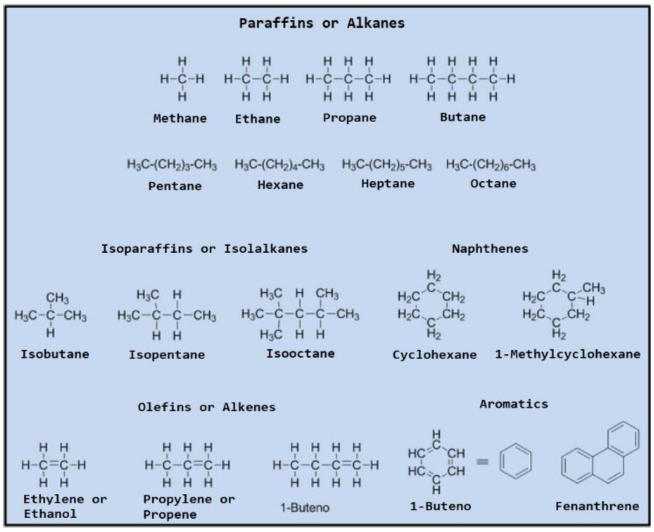


Fig. 2. Different types of hydrocarbons according to their structure and number of carbons.

Hydrocarbons differ in whether they form linear chains (paraffins), branched (isoparaffins), if they are cyclic (naphthenes and aromatics) or if they contain double bonds (olefins). In addition, the number of carbons also represents an important differentiation between them. All these structural differences affect properties such as, for example, their state of aggregation at room temperature (solid, liquid or gas) or their melting and boiling point. Thus, in general, the more carbons you have, the higher your boiling point will be, although the ramifications and the formation of cycles can cause it to vary.

Thus, gasoline is composed of hydrocarbons whose boiling point is approximately between 40 and 150 °C, which are usually between 5 and 9 carbons. That of all the hydrocarbons that make up the oil, they make up gasoline. The process to separate several hydrocarbons from a mixture is called Refined and consists of four stages [2]:

Fractionation: it consists of a fractional distillation of the "crude oil" (the oil resulting from the removal of impurities from gases, water and solids after extraction from land or marine deposits), in which the different hydrocarbons that make up the mixture are separated. lower to higher boiling point. That is, the oil slowly warms up, and in this way the different hydrocarbons evaporate as it reaches its boiling point. When they evaporate, they can be "trapped" and separated into different compartments.

Cracking: once all the fractions are obtained, some can be transformed into others through chemical reactions. Normally the light fractions (natural gas, gasoline) are usually of greater industrial interest, so that a fraction of hydrocarbons of 10-14 carbons, can be broken chemically for example in fractions of 5-7 carbons.

Reformed: this process is already more oriented to the light fractions of gasoline, aimed at improving its quality as a fuel. To do this, chemical modifications are made in specific hydrocarbons (hydrogenate, dehydrogenate, isomerize ... are to modify the chemical structure) to improve the Octane Index of gasoline.

Purification: finally, the purpose of the purification process is to rid gasoline of undesirable compounds, such as sulfur derivatives (that is, containing sulfur). Gasoline is obtained from oil in a refinery. In general, it is obtained from direct distillation naphtha, which is the lightest liquid fraction of oil (except gases). Naphtha is also obtained from the conversion of heavy fractions of oil (vacuum gas oil) into process units called FCC.

After these four stages, gasoline is already suitable for its purpose. The quality as fuel for gasoline cars is measured by its Octane Index (IO). Technically, the IO measures the ability of a gasoline to withstand high pressures and temperatures inside the engine cylinder, without spontaneous detonation.

The gasoline must withstand the pressure and temperature caused by the engine without exploding before the spark in the spark plug occurs. In practice, the IO of a gasoline sample is measured.

According to EU regulations, the minimum IO that a gasoline must have is 95. To increase the IO of a gasoline to 95 (Reforming process) chemical reactions are carried out on the hydrocarbons that make up this mixture to obtain different hydrocarbons with a structure that, according to established guidelines, improve the IO (more branches, lower molecular weight, greater presence of naphthene’s and aromatics ...). Another way to increase IO is using additives, specifically substances that act as antiknockants.

In the mid-twentieth century, tetraethyl lead, present in the extinct leaded gasoline of 97, began to be used as antiknock. But due to its polluting properties (it was expelled into the atmosphere with the exhaust gases) and its accumulation inside the engine , in the mid-70s he began to question its use, reducing it to 0.6 g / l, then to 0.15 g / l in the 90's, until completely removing it from gasoline at present, that is why all Gasolines are now "unleaded".

The properties of gasoline are listed below:

- Cetane Index indicates the ease with which diesel begins to burn. The higher the number, the easier it is for the fuel to start combustion.
- Aromatics: they can be liquid or solid, they are insoluble in water, but soluble in most organic solvents and have a pleasant characteristic odor.
- Elemental sulfur: It is a chemical element of atomic number 16 and symbol S. It is used mainly as a fertilizer but also in the manufacture of gunpowder, laxatives, phosphors and insecticides.
- Carbon waste: is formed after evaporation and decomposition by heating of an oil product
- Ashes: It is the product of the combustion of some material, composed of non-combustible inorganic substances. Part remains as waste in the form of dust.
- Flammability point: The set of conditions of pressure, temperature, gas mixture in which a combustible substance produces enough vapors that, when mixed with air, would ignite when applying a source of heat.

III. Parallel Classifier Based on Mixture of Experts (PCEM)

A. Parallel Architecture of PCEM

PCEM is a heterogeneous ensemble classifier. Using Parallel Computing we can solve many problems (time reduction, saving memory, handling large amounts of data, maximum use of computing power, sharing processing), using a system of parallel computing (cluster, grid, gpu’s and multiprocessor) for the implementation of a PWeLe’s [12]. For the implementation of PCEM we use the programming language C, this programming language provides a framework for parallel programming. In this tool we can make use of parallel computing using processes that communicate through messages using the MPI library.

In the PCEM we also have a parallel scheme of a genetic algorithm (PGA) [13], which is a novel approach since its usually done with a Neural Network (NN), to help us find the right weights for each classifier. With the PGA we obtain

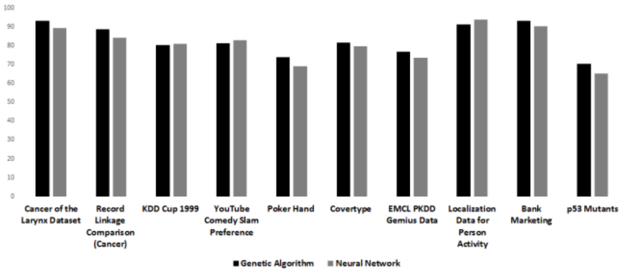


Fig. 3. Comparison of accuracy of GA and NN.

In Fig. 3, in seven of ten datasets the PGA have better performance against NN, this result and the simplicity to implement the parallel scheme of Genetic Algorithm is the reason why a PGA was used to find the right weights for each classifier.

Finally, a parallel scheme is implemented for the weighted voting, therefore we have global parallelization for each component of PCEM. First, we describe the quantity and type of classifiers we selected for the construction of PCEM.

The structure we selected for the ensemble was a heterogeneous mixture of experts. The PCEM uses the following criteria for the selection of classifiers:

1) The implementation of the parallel scheme to any classifier had to be simple, because the main objective of mixture of experts is combining a set of WeLe's for constructing a strong learner.
2) Select some parallel classifiers reported in the literature, to have a theoretical basis for its correct operation. In these criteria we selected five WeLe's, this number we considered adequate, because on our tests with this number PWeLe's, we obtained improvements in performance measures. But as we increase further the number of classifiers this improvement was no longer significant, and the execution times are increased. So, we settled for use only five PWeLe's for this first approach, since in the results show that there is no significant improvement in using more PWeLe's.
3) The parallel classifiers selected, must support large amounts of data.
4) Finally, we selected five WeLe that meets these criteria. We selected four supervised (k-NN, Na¨ıve Bayes, Decision Tables, and C4.5) and one unsupervised (K-means).

To use k-means as a classification algorithm we iterate until we find that the number of clusters equals the number of classes we know exist. Then we test on unseen data based on the proximity to the clusters found by the algorithm.

B. Operation of PCEM

To develop the PCEM we chose the MIMD architecture of the Flynn taxonomy [13]. Each PWeLe has training and a testing set, which obtains the individual classification of test set in parallel. In the case of parallel genetic algorithm (PGA), we use a subset to find the best weights for each PWeLe [14]. Finally, we have a coordinator process to compile this information of PWeLe and PGA for applying the weighted voting criterion, in the scheme of Fig. 4 we show the architecture of PCEM.

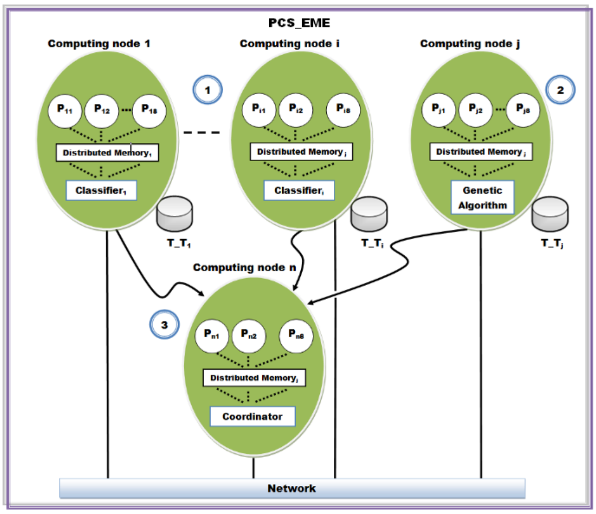


Fig. 4. Parallel Architecture of PCEM.

Fig. 4 shows following elements:

- $PWeLei$ $\forall$ i = 1, 2, ..., n; are all PWeLe of PCEM.
- $Computing\ node$ $\forall$ i = 1,2, ..., n, are the nodes available in the Multicomputer System
- $Distributed\ memoryjk$ $\forall$ j = 1, 2, ..., n and k = 1, 2, …, n; are j distributed memories for each k computing nodes.
- Pij $\forall$ i = 1,2, ..., n and j = 1, 2, . . ., n, are the j processes executed in each i computing nodes.
- T_Ti $\forall$ i = 1, 2, ..., n, are the training and test sets used for each classifier and the GA of the PCEM.

Once any PWeLe receives its training and test set by the General Coordinator (GeCo), it executes its training stage in parallel. Whenever some classifier completes its training phase, it proceeds to find the individual classification of the test set and sends a message to GeCo. When the GeCo gets all individual classifications, we proceed with the next stage of PCEM [9].

The next stage is the allocation of weights to the PWeLe's. As we can see the problem of assigning weights is a complex problem because of the possible solutions that can be had, which is why to give a possible solution to this problem optimization methods are used. For this work, we decided to use the PGA, to find the weights for each PWeLe, because its better performance as was discussed in Section 3.1, we can see its benefits over traditional scheme of ensemble of mixture of experts. In this case the genetic algorithm process receives a message from the GeCo, with its training and test set.

Once that the genetic algorithm obtains the best combination of weights for each classifier, this sends a message to the GeCo with these combinations of weights, collecting the information generated in the second stage.

Once the GeCo receives all individual classifications and the weights generated by the genetic algorithm in the second stage, we proceed to implement the weighted voting criterion to obtain the final classification of each elements of the test set. To describe the communication in PCEM we can see in the scheme of Fig. 5.

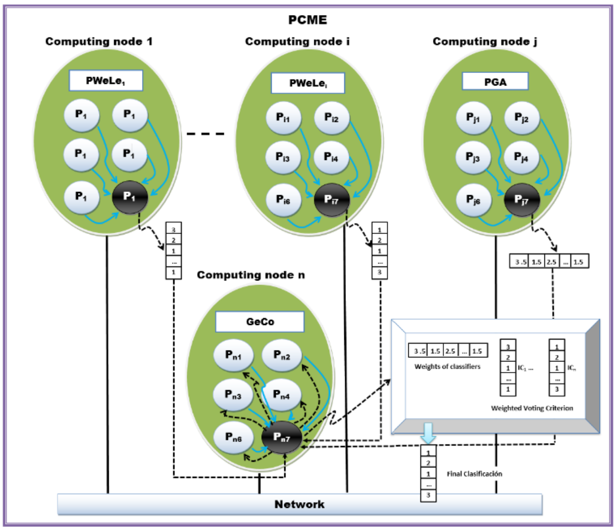

PCME
Computing node 1
Computing node i
Computing node j
PWeLe$_1$
PWeLe$_i$
PGA
P_1
P_{i1} P_{i2} P_{i3} P_{i4} P_{i6} P_{i7}
P_{j1} P_{j2} P_{j3} P_{j4} P_{j6} P_{j7}
3.5 1.5 2.5 ... 1.5
Computing node n
GeCo
P_{n1} P_{n2} P_{n3} P_{n4} P_{n6} P_{n7}
Weights of classifiers
IC$_1$ ...
IC$_n$
Weighted Voting Criterion
Final Clasificación
Network

Fig. 5. Communication of PCEM.

Fig. 5 has the following information:

- $Classifieri$$\forall$ i = 1, 2, …, n, are all PWeLe of PCEM.

- PGA is the Parallel Genetic Algorithm
- $Computer\ nodei$ ∀ i = 1, 2, …, m, are the m nodes available in the Multicomputer System
- $GeCo$ is the General Coordinator.
- The Processes of blue colour in each $Computer\ node$, we use to perform the operation of each $PWeLe$'s, the PGA and the $GeCo$, we called $P\_Builders$.
- The Processes of black colour in each $Computer\ node$, we use them as sub coordinator of each component of PCEM to gather the information in each component for example individual classification, weights of each classifiers and combination of individual classification with the weighted voting criterion. This process we called $CP\_ga\quad herers$.
- ICi ∀ i = 1, 2, …, n, are the individual classification generated for each classifier.

To illustrate communication in the PCEM, let's assume we have i classifiers, the genetic algorithm and the Coordinator, which execute 8 processes per computer node. In Fig. 5 we see that in each of the computer nodes, local communication is performed which is represented by the blue lines. Through this local communication the $P\_Builders$ send their respective portion of the information generated to CP gatherers.

Once that CP gatherers get the IC and the weights of classifiers, these processes send a message (dotted lines) to $CP\_gatherers$ of Coordinator. With this information the Coordinator apply the Weighted Voting Criteria with a set of P Builders to find the Final Classification. When the Coordinator gets the final classification, this calculates the performance measures to evaluate the performance of PCEM, which is the last stage of the operation of the PCEM.

In the following sections we present the dataset of Catalytic Cracking Microactivity Reactor used for performing the experiment to evaluate the results against of a traditional and parallel classifier to obtain the main contributions this work.

IV. The Dataset of Catalytic Cracking Microactivity Reactor

A. Transformation of the Database

To obtain our final database, the data was collected from a monitoring of physical and chemical properties of Gasoline and Diesel commercialized in the metropolitan area of the Valley of Mexico, carried out by the Secretariat of Extension and Social Integration of the National Polytechnic Institute.

The raw data are found in a series of documents, in which the results of fuel characterization are reported, taking into consideration three types of gasoline [3], called Regular or Magna Gasoline (87 octane), Premium Gasoline (92 octane)) and

Diesel (52 octane).

These documents were generated during the month of November 2016 and we can find 4100 samples of Regular Gasoline, 4100 samples of Premium Gasoline and 4100 samples of Diesel. The format of one register for the Regular Gasoline can be seen in Table I.

- RD is the Residue from Distillation
- vol is the unit of volume
- ppm is parts-per-million, 10−6, and Mppm is the Maximum parts-per-million
- MV is the Maximum Volume
- M is the Maximum ℃
- m is the minimum value of Octanes
- SPEC is the specification.

TABLE I: Format of the Regular, Premium and Diesel Gas Data

Property	ASTM	Value	Unity	SPEC
Octane	D6729	89.6	-	87 m
Aromatics	D6729	23.6	%vol	25% MV
Olefins	D6729	6.20	%vol	10% MV
Oxygenated	D6729	2.04	% w	2.7% MW
Elemental sulfur	D7039	14.01	ppm	80 Mppm
Steam pressure at 100 °F	D323	7.6	lb/pul^3	7.8% lb/pul^3
% DV	D86			
IBT		38.4	℃	Report
At 10% R		55.8	℃	70 ℃ M
At 50% R		92.5	℃	[77, 121] °C M
At 90% R		159.6	℃	190 ℃ M
FBT		198.1	℃	225 ℃ M
RD		0.8	%vol	"% M

Table I has the following information:

- ASTM is the American Society for Testing and Materials or also called ASTM International [15]
- DV is the Distilled Volume
- IBT is the Initial Boiling Temperature
- R is the is the percentage of the Volume Recovered
- FBT is the Final Boiling Temperature

Since the characteristics that describe Regular and Premium Gasoline are not the same as the characteristics that describe Diesel, it was done to eliminate the value of Ashes, in the Diesel data, this because in most of the records the value is shows as Less than 0.001. With this modification all data could be merged into a single table. Since all the data were of real type, we opted to convert them to integers since most of the PWeLe's that make up the PCEM's operation engine have better operation with the integer type data. The last modification that was made to the data to finish the formation of our database was that an attribute was added in the last row, which indicates to what type of Gasoline each instance belongs. The values assigned to each class were 1 for Regular Gasoline, 2 for Premium Gasoline and 3 for Diesel, thus having a problem of multi-class classification.

To obtain more test data, we considered the possibility of combining the data we had in order to obtain a larger set of the initial base. The base that we obtained by combining the values of each of the attributes with each class gave us a total of 760, 008 examples, which has an equitable distribution, that is, we have the same number of examples for each class (253,336). In the next Section we will show the experiments and results that were obtained when using the PCEM with this database of Catalytic Cracking Microactivity Reactor.

V. Experimental and Results

Within Machine Learning there are different performance measures to evaluate the task of classification; in this paper we only show the results for Accuracy. Accuracy is the percentage of examples classified correctly in the test set. 2.

We have a parallel scheme Na ïve Bayes, Decision Tables, C4.5, k-NN and used KMeans. For PGA, a roulette is used for selection whit 50% for crossover a mutation rate of 12% is used.

In the experiments we use 10-fold cross-validation, for each iteration this validation uses one of the subgroups for the test set and the rest of the subgroups for the training set. This is done 10 times reporting the average of each measures. Now, we will present the results found with all datasets, comparing different, traditional and parallel classifiers, against the PCME.

The Cluster used have 7 nodes in total. Each node has 2 quad core processor with 8 cores in total and 8 GB in RAM. We used 2, 4, 6, 8, 10, 12 and 36 processes for each node, ie, each component of PCME, the PWeLe's, the PGA and the GeCo were tested with this number of processes So, in a Cluster we can have a maximum number of 252 processes.

The distribution of each of the dataset is performed as follows. At first the GeCo produces 10 random subsets of the total number of records. Ten iterations are performed to test all examples of datasets used. Therefore, in each iteration the GeCo sends the training set (a subset) and the test set (nine subsets) to each PWeLe and PGA.

To define the appropriate percentage of the training and test set of the genetic algorithm, that for every dataset used is different, we use a statistical test based on the variance of a test sample. Considering a confidence level of 0.95, with a maximum error of 0.1 obtaining a variance of 154.5, according to the sample random simple calculation of the training set will be calculated by the GeCo, to assign the training set for the genetic algorithm. After a series of calculations with the datasets used, for example if we have a set of training records 200,000, the value recommended for the correct operation of a genetic algorithm is the 27,049 records, which is equivalent to 13.5% of the size of the training set.

Now, we will present the results found when per-forming these tasks with all datasets, comparing different traditional and parallel classifiers against the PCEM, showing the accuracy results obtained, these results we can see in Table II.

TABLE II: COMPARISON OF RESULTS OF ACCURACY

Classifier	k	Acurracy	Time (min)
k-NN	21	82.5 %	11.1
P*k*-NN	21	82.5 %	3.5
NB	-	77.8 %	21.3
PNB	-	77.8 %	5.9
C4.5	-	83.2 %	27.4
PC4.5	-	85.7 %	9.8
DT	-	81.6 %	31.7
ParalTabs	-	86.2 %	6.9
K-Means	-	73.4 %	13.8
P*K*-Means	-	78.7 %	3.7
Bagging	21	83.3 %	54.2
Boosting	-	85.7 %	43.5
SG	7	87.5 %	61.6
HCGW	21	89.2 %	63.3
PCEM	-	94.3 %	12.5

A. Final Version of SIMGRAPH of the DSBM

Table I has the following information:

k-NN is the sequential scheme of classifier called k - nearest neighbors

Pk-NN is the parallel scheme of k - nearest neighbors

NB is the sequential scheme of classifier called Na ïve Bayes.

PNB is the parallel scheme of classifier called Na ïve Bayes.

C4.5 and PC4.5 is the sequential and parallel scheme of classifier used to generate a decision tree developed by Ross Quinlan.

DT is the sequential scheme of classifier called Decision Tables.

ParalTabs is the parallel scheme of decision tables [12]

Bagging and Boosting are two types of classifiers based on ensembles and we described in Section 2.2

SG is the classifier called Staked Generalization and we described in Section 2.2

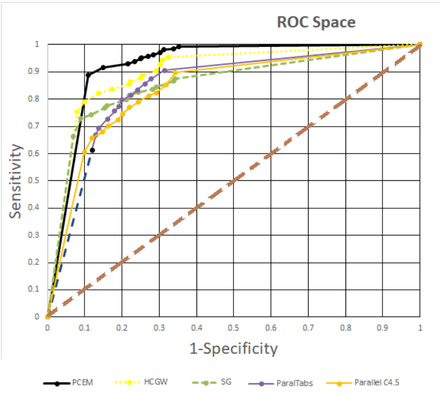


Fig. 5. ROC curve.

HCGW is the Hybrid Classifier with Genetic Weighting

We can see from the results of Table II, the PCEM gets the better results; the increases with PCEM for the dataset of Catalytic Cracking Microactivity Reactor is more than 10% against to the traditional classifiers. This improvement is due to the PCEM that implemented parallel schemes to each WeLe, which, compared with their implementation in the HCGW, in some cases the WeLe of PCEM obtained best accuracy rates. Comparing the results of the PCEM and HCGW an improvement of over 5% accuracy is obtained.

In Table II we can see that PCEM obtains better execution times compared to sequential classifiers including the sequential version of HCGW. In the case of HCGW, which is like PCEM because the two are based on an ensemble of a type Mixture of Ex-pert, the execution time obtained by PCEM represents the 19% of the execution time was obtained by HCGW, representing a large reduction in execution times. With these results we can see the advantage to use a classifier based on ensembles with respect to traditional classifiers.

In Fig. 6 shows the ROC Curve and PCEM has better results with respect to HCGW, SG, parallel scheme of C4.5 and ParalTabs.

The areas under the ROC convex hull curve of each classifier used in the experiments are shown in Table III. In Table III, we can see that for the PCEM we obtained a better value of the ROCCH AUC with respect to other classifiers used for these experiments.

TABLE III: THE AREA UNDER THE ROC CONVEX HULL

Classifier	ROCCH AUC
PCEM	0.9523
HCGW	0.8679
SG	0.7950
ParalTabs	0.7762
PC4.5	0.7551

VI. CONCLUSION AND FUTURE WORK

In the first stage of this work the process of Knowledge Discovery in Databases (KDD) was carried out to generate a database with the different types of gasoline and its properties. Additionally, the data mining process was carried out on the databases of the Catalytic Cracking Microactivity Reactor with the five PWeLe's.

From Table II, it can be highlighted that the best result is obtained with the ParalTabs classifier, obtaining an 86.2% average accuracy. We identify that using a traditional classifier even if its parallel scheme is not enough since the indices in the operating measures are low. We infer this by comparing the result of ParalTabs with respect to the HCGW classifier, from which we can see that there is an increase of 3%. Since we saw this advantage when using classifiers based on ensembles, we decided to use the PCEM as the operating engine of this application.

Given each one of the statistical tests we could confirm that the PCEM classifier is an excellent option to manage databases of Catalytic Cracking Microactivity Reactors, since as we present in Fig. 6 and Table III, of the ROC analysis, this classifier obtains the best area on the ROC curve.

The main objective of this multidisciplinary project was fulfilled, since we were able to implement Data Mining on the raw data of a Catalytic Cracking Microactivity Reactor, the question would be, what is the real use that can be given to the results obtained in this work? We could verify that based on the results of Table II, the PCEM obtains a result of 94.3% of average effectiveness, which is an excellent index for this type of data. Taking this percentage of accuracy, we can think of an application that can be given to our work to attack the problem of fuel traffic that afflicts the country of Mexico in the last 5 years, commonly called as Huachicol.

It is planned to create a mobile intelligent system that can offer the service to the federal and municipal patrols so that they can deduce in real time what type of gasoline is confiscated and have a greater scope of deducing from what type of oil well it comes from, having a very significant contribution to the attack of this problem that has caused serious problems and economic instability in the country of Mexico.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Benjamín Moreno-Montiel carried out the initial approach of the case study, the writing of the paper, as well as the implementation of all modules of PCEM, analysis and interpretation of the results of the ROC process. Carlos-Hiram Moreno-Montiel supported the writing of the paper and analysis of the data obtained from this paper. René MacKinney-Romero contributed the analysis of the classification results presented in this paper, as well as the analysis and interpretation of the results of the ROC process. Miriam-Noemi Moreno-Montiel supported in most of the theoretical aspects of the Chemical Reaction Engineering and in the writing of the paper.

REFERENCES

[1] PEMEX Refinación, *Especificación No. 107/2008*, Hoja técnica de especificaciones, 2008.

[2] P. Caballero-Mata *et al*., "Análisis de las propiedades fisicoquímicas de gasolina y diesel mexicanos reformulados con Etanol," *Ingeniería Investi-gación y Tecnología*, vol. XIII, 2012, pp. 293-306.

[3] B. Moreno-Montiel and R. MacKinney-Romero, "A hibrid classifier with genetic weighting," in *Proc. the of the Sixth International Conference on Software and Data Technologies*, pp. 359-364, Sevilla, España, 2011.

[4] B. Moreno-Montiel and R. MacKinney-Romero, "Parallel classification system based on ensemble of mixture of experts," in *Proc. the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2014, Angers, Francia 2014.

[5] E. Menahem, L. Rokach, and Y. Elovici, "Troika — An improved stacking schema for classification tasks," *Inf. Sci*., pp. 4097–4122. 2009.

[6] D. J. Miller and H. S. Uyar, "A mixture of experts classifier with learning based on both labeled and unlabeled data," *Neural Information Processing Systems*, no. 9, pp. 571-577, 1997.

[7] L. Breiman, *Bagging predictors. Machine Learning*, no. 25, pp. 123–140, 1996.

[8] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, pp. 119–139, 1997.

[9] Sun, "Local within-class accuracies for weighting individual outputs in multiple classifier systems," *Pattern Recognition Letters*, vol. 31, no. 2, pp. 119â-124, 2010.

[10] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput*., vol. 3, no. 1, pp. 79–87, 1991.

[11] T. J. Bruno *et al*., "Composition-explicit distillation curves for mixtures of gasolina with four-carbon alco-hols (butanols)," *Energy & Fuels*, vol. 23, 2009.

[12] C. H. Moreno-Montiel, "Detection of genes in individual associated with laryngeal cancer using paraltabs," *IETE Journal of Research (IETE)*, 2017, vol. 63

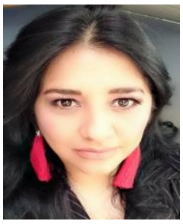
[13] M. Flynn, "Some computer organizations and their effectiveness," *IEEE Trans. Comput.*, 1972.

[14] D. F. Oliveira, A. M. P. Canuto, and M. C. P. Souto, "Use of multi-objective genetic algorithms to investigate the diversity/accuracy dilemma in heterogeneous ensem-bles," in *Proc. International Joint Conference on Neural Networks*, 2009, pp. 2339–2346.

[15] American Petroleum Institute, Alcohols and Ethers: A Technical Assessment of Their Application as Fuels and Fuel Components, 3rd ed., API publication 4251, Washington, DC, 2001.

Benjamin Moreno-Montiel received the BSc and MSc and PhD degrees in science and technologies of information from Universidad Autónoma Metropolitana, Unidad Iztapalapa, Mexico city, México, in 2007 2009 and 2017, respectively. He has been a professor of computer sciences since 2015 in Universidad Autónoma Metropolitana where he took his undergraduate courses in computer sciences. He teaches artificial intelligence and machine learning for undergrads and postgrads and has some publications in international conferences and journals.

Carlos-Hiram Moreno-Montiel received the BSc and the MSc degrees in science and technologies of information from Universidad Autónoma Metropolitana - Unidad Iztapalapa, Mexico City, México, in 2006 and 2010, respectively. He is cur-rently working towards the PhD degree at the Posgrado en Ciencias y Tecnologías de la Información in Universidad Autónoma Metropolitana – Unidad Iztapalapa, México City, México. His research interests are parallel computing, software engineering, and artificial Intelligence.



Miriam-Noemi Moreno-Montiel received the BSc and MSc and PhD degrees in science chemical engineering from Universidad Autónoma Metropolitana – Unidad Iztapalapa, Mexico City, México, in 1998 2000 and 2004, respectively. She has been a professor of computer sciences since 2005 in Instituto Politécnico Nacional, of Departamento de Ingeniería Química Petrolera – ESIQUIE, where she took her undergraduate courses in computer sciences. She teaches Modelling and simulation in chemical engineering, Characterization of Petroleum Fractions, Chemical Reaction Engineering, Transport Phenomena for undergrads and postgrads and has some publications in international conferences and journals.

René MacKinney Romero received the BSc degree from Universidad Autónoma Metropolitana, Unidad Iztapalapa, Mexico City, Mexico, in 1993, the MSc degree in computation from the Univer-sity of Oxford, England, in 1994, and the doctorate degree in computer sciences from the University of Bristol, England, in 2002. He has been a professor of computer sciences since 1990 in Universidad Autónoma Met-ropolitana where he took his undergraduate courses in com-puter sciences. He teaches artificial intelligence and machine learning for undergrads and postgrads and has many publica-tions in international conferences and journals.