Quintero Gull, C., Aguilar Castro, J.L. & Rodriguez Moreno, M.D. 2021, "A semi-supervised learning approach to study the energy consumption in smart buildings", in 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 05-07 Dec. 2021.

Available at http://dx.doi.org/10.1109/SSCI50451.2021.9659911

A semi-supervised learning approach to study the energy consumption in smart buildings

Carlos Quintero Gull
Dpto de Ciencias Aplicadas y Humanísticas. Doctorado en Ciencias Aplicadas Facultad de Ingeniería
Universidad de Los Andes
Mérida 5101, Venezuela
carlgull@gmail.com

Jose Aguilar
Universidad de Alcalá, Escuela Politécnica Superior, ISG,
Alcalá de Henares, 28805, Spain;

CEMISID, Universidad de Los Andes,
Mérida, 5101, Venezuela;

GIDITIC, Universidad EAFIT,
Medellín, 50022, Colombia
jose.aguilar@uah.es

Maria D. R-Moreno
Universidad de Alcalá, Escuela Politécnica Superior, ISG,
Alcalá de Henares, 28805, Spain;

TNO, Intelligent Autonomous Systems Group (IAS),
The Hague, The Netherlands
malola.rmoreno@uah.es

***Abstract*—In this work, we use the semi-supervised LAMDA-HSCC algorithm for characterizing the energy consumption in smart buildings, which can work with labeled and unlabeled data. Particularly, it uses the LAMDA-RD approach for the clustering problem and the LAMDA-HAD approach for the classification problem. Additionally, this algorithm uses three submodels for merging, partition groups (classes/cluster) and migrating individuals from a group to another. For the performance evaluation, several datasets of energetic consumption are used, with different percent of labeled data, showing very encouraging results according to two metrics in the semi-supervised context.**

Keywords—Semi-supervised learning; Multivariate Data Analysis; LAMDA, Energetic Consumption

I. INTRODUCTION

In the literature, there are a large number of methods of machine learning (ML) techniques, which can be mainly classified into four branches: supervised, unsupervised, reinforcement learning and semi-supervised. Semi-supervised learning takes place when the algorithm learns with a data set that simultaneously contains labeled and unlabeled data, which introduces a greater degree of difficulty to the learning process [1, 2, 4, 6, 7, 15, 20]. This type of algorithm can work also separately, on classification or clustering tasks, which expands its usefulness [11, 12, 14].

In the particular case of the study of energy consumption in smart buildings, we can found several works based on ML techniques [5, 13, 22]. These models are mainly designed for forecasting energetic consumption, obtaining good results. However, in the particular case of the characterization of smart buildings according to their consumption energetic features in a semi-supervised context, there is a lack, for the creation of flexible models.

On the other hand, the LAMDA algorithm is a method based on fuzzy logic, which focuses on calculating the Global Adequacy Degree from an individual to a class (classification) or group (clustering) [3, 10, 16, 17, 18, 23, 24]. In this way, Quintero and Aguilar [21] developed a semisupervised algorithm, called LAMDA-HSCC, which involves tasks of classification and clustering to consider the following scenarios: i) Given a dataset with labeled samples, the algorithm assigns the input data to the corresponding classes; ii) Given a dataset with unlabeled samples, the algorithm group them into clusters based on the similarity of the descriptors within each cluster, iii) Given a dataset with labeled and unlabeled samples, the algorithm assigns them to classes or clusters according to each case. Additionally, [21] developed three strategies for the Merging Migration and Partition of class/clusters, and demonstrate that these strategies improve the efficiency of the previous algorithms of the LAMDA family.

In this work, we make a characterization of the energy consumption of smart buildings using the LAMDA-HSCC, considering labeled and unlabeled data. This work is organized as follows: Section 2 introduces the fundamentals of LAMDA and Section 3 LAMDA-HSCC technique. Section 4 analyses the energy consumption in smart buildings, defines a set of experiments in different scenarios. and presents the analysis of the results. Finally, Section 5 outlines the conclusions and the future works.

II. CONCEPTUAL FUNDAMENTS OF THE LAMDA ALGORITHM

A. LAMDA

LAMDA is a fuzzy algorithm that is based on the membership function that data can have to a class [10, 16, 17, 18, 24]. This algorithm usually works with a dataset of individuals with the next format $X = \{x_1, x_2, \ldots . x_j, \ldots x_m\}$, which is a vector with *m* descriptors/features, where x_j is the descriptor j of the individual X. Additionally, for labeled samples, X becomes $X = \{x_1, x_2, \ldots . x_j, \ldots . x_m, c_l\}$, where c_l is the label associated to the individual for l=1,…, k. In the unlabeled samples, it is not associated with any class.

Also, it is necessary that the descriptors/features of the individual X be normalized based on their maximum and minimum values, as shown in Equation (1):

$$\bar{x}_j = \frac{x_j - x_{jmin}}{x_{jmax} - x_{jmin}} \quad (1)$$

Where: $\bar{x}_j$: normalized descriptor/feature, $\bar{x}_{jmin}$: Minimum value of descriptor j, $\bar{x}_{jmax}$: Maximum value of descriptor j.

Next, the bases of the LAMDA algorithm are presented:

Definition 1 Marginal Adequacy Degree (MAD). It determines how similar a descriptor is with respect to the same descriptor

in a given class. For MAD calculation, density functions are used, one the most common is Fuzzy Binomial function, showed in Equation (2).

$$MAD(\bar{x}_j/\rho_{k,j}) = \rho_{k,j}{}^{\bar{x}_j}\,(1-\rho_{kj})^{(1-\bar{x}_j)} \quad (2)$$

Where $\rho_{k,j}$: is the average value of the descriptor j that belongs to the class k, calculated using Equation (3):

$$\rho_{kj} = \frac{1}{n_{kj}}\sum_{t=1}^{n_{kj}} \bar{x}_j(t) \quad (3)$$

On the other hand, the degree of Marginal Adequacy can also be determined by the Gaussian function, through Equation 4:

$$MAD(\bar{x}_j/\rho_{kj}) = e^{-\frac{1}{2}\left(\frac{\bar{x}_j-\rho_{k,j}}{\sigma_{kj}}\right)^2} \quad (4)$$

Where, σ_{kj}:is the standard deviation of the descriptor j that belong to the class k:

$$\sigma_{kj}^2 = \frac{\sum_{t=1}^{n_{kj}}(\bar{x}_j(t)-\rho_{kj})^2}{n_{kj}} \quad (5)$$

Definition 2 Global Adequacy Degree (GAD). It determines the adequacy of an individual to each class. This value is based on the MAD, and can be determined according to Equation (6):

$$GAD_{k,\bar{X}} = (MAD_{k,1}, MAD_{k,2}, \ldots, MAD_{k,n}) = \alpha T(MAD_{k,1}, \ldots, MAD_{k,n}) + (1-\alpha)S(MAD_{k,1}, \ldots MAD_{k,n}) \quad (6)$$

Where, α is an exigency parameter (0 ≤α≤1) used to calibrate the fuzzy partition of the data, determining the type of classification of the data [16].

The assignment of an individual to a class is done by calculating the maximum GAD of all existing classes. In that sense, the index in Equation 7 corresponds to the number of the class where the individual will be assigned.

$$\text{index} = \max(\text{GAD}_{1\bar{X}}, \text{GAD}_{k\bar{X}}, \ldots, \text{GAD}_{m\bar{X}}, \text{GAD}_{\text{NIC}\bar{X}}) \quad (7)$$

Where: $GAD_{NIC\bar{X}}$: Marginal Adequacy Degree of the non-informative class that is formed when an individual is not assigned to existing classes. This happens when calculating the GAD'S, these probabilities are less than 0.5.

B. LAMDA –HAD Algorithm

The LAMDA-HAD (Highest Adequacy Degree) classification algorithm was developed by Morales et al. [16] as an improvement to the LAMDA algorithm because the threshold value for sending an individual to the Non-Informative Class (NIC) is constant for all classes, or also, because the GAD calculation is not very reliable. So, Morales et al. [16] improved the efficiency of this algorithm through the calculation of a variable NIC for each class, and through the calculation of the highest degree of adequacy (HAD), which gives greater robustness to the LAMDA algorithm. For that, the next concepts are defined [16]:

Definition 3. Let $p = \{1, \ldots, m\}$ the number of existing classes in a dataset. Let $MGAD_{k,p}$ the average of $GAD's$ of class p in the class k:

$$MGAD_{k,p} = \frac{1}{n_k}\sum_{t=1}^{t=n_k} GAD_{p,t} \quad (8)$$

Definition 4. Let GAD_{NIC_p} the GAD of the no informative class (NIC) of the class p calculated as the average of the $MGAD_{k,p}$:

$$GAD_{NIC_p}: \frac{1}{m}\sum_{p=1}^{p=m} MGAD_{k,p} \quad (9)$$

Definition 5. Let $AD_{GAD_{k,p,\bar{x}_r}}$ be a parameter that allows one to compare the similarity between the GAD of an individual and each $MGAD_{k,p}$

$$\boldsymbol{AD_{GAD_{k,p,\bar{X}_r}} = MGAD_{k,p}{}^{GAD_{p,\bar{X}_r}}\,(1 - MGAD_{k,p})^{(1-GAD_{p,\bar{X}_r})}} \quad (10)$$

Definition 6. Let $AD_{GAD_{k,p,\bar{x}_r}}$ be the Highest Degree of Adequacy of an individual to a class, $(HAD_{k,\bar{X}_r})$ is formed by adding all the $AD_{GAD_{k,p,\bar{x}_r}}$ in class p:

$$\boldsymbol{HAD_{k,\bar{X}_r} = \sum_{p=1}^{p=m} AD_{GAD_{k,p,\bar{x}_r}}} \quad (11)$$

Let E_I be the class that the individual has the highest probability of belonging:

$$\boldsymbol{E_I = max(HAD_{1,\bar{X}_r}, HAD_{2\,\bar{X}_r}, \ldots HAD_{k,\bar{X}_r}, \ldots . HAD_{m,\bar{X}_r})} \quad (12)$$

Definition 7. Let index the value that identifies the class that an individual will be assigned, which is obtained by comparing the maximum value between E_I and the $GAD_{NIC_{E_I}}$:

$$\boldsymbol{index = max\left(GAD_{E_I,\bar{X}_r}, GAD_{NIC_{E_I}}\right)} \quad (13)$$

Once the LAMDA-HAD algorithm is finished, the result will be the number of the class to which the individual is assigned; or otherwise, the individual is sent to the non-informative class (NIC).

C. LAMDA –RD

The LAMDA-RD algorithm was developed by Morales and Aguilar [18], in order to improve the LAMDA algorithm in the context of clustering's tasks, since this algorithm LAMDA has the tendency to create clusters that do not correspond to the desired number of clusters. To mitigate this inconvenience, they determined a more robust metric of clustering, and also, develop a strategy of automatic cluster fusion. Next, the conceptual bases of this algorithm.

Definition 8 Cauchy Marginal Adequacy Degree (CMAD). It corresponds to the Marginal Adequacy Degree, but using the Fuzzy Cauchy Function:

$$CMAD = \frac{1}{1+dist(\bar{x}_j,\rho_{kj})} \quad (14)$$

Where: $\text{dist}(\bar{x}_j, \rho_{kj})$: Distance between the individual $\bar{x}_j$ and the average ρ_{kj}.

Definition 9 Robust Marginal Adequacy Degree (RMAD). It corresponds to the Marginal Adequacy Degree, but now is accompanied by a factor that penalizes each cluster, and is determined by:

$$\boldsymbol{RMAD = k_{\bar{x}k} * CMAD} \quad (15)$$

For the calculation of $k_{\bar{x}k}$ two parameters are required, the first is the average distance of the individual between the clusters, which is calculated like:

$$\boldsymbol{d_{k,\bar{X}_r} = dist(\bar{x}_j , \rho_{k,j}) = \frac{1}{n}\sum_{j=1}^{n}|\bar{x}_j - \rho_{k,j}|} \quad (16)$$

The second parameter is the average distance between neighbor clusters $(d_{n,b})$, $d_{n,b} \in [0 , 1]$, which is a parameter that describes the average distance between clusters, and is obtained through calibration.

Definition 10. Let $d_t \in [0 , 1]$ be a threshold of the density of the cluster, which is obtained through a calibration process.

Definition 11. The penalty factor is calculated as shown in Equation (17), in case of the distance $d_{k,\bar{X}_r}$ is greater than $d_{n,b}$

$$k_{\bar{x}k} = \frac{d_{n,b}}{d_{n,b}+dist\left(d_{k,\bar{X}_r} , d_{n,b}\right)} \quad (17)$$

Definition 12. Global Adequacy Degree is a linear combination of the Robust Marginal Adequacy Degree (RMAD), where $\alpha \in [0 , 1]$ is the exigency parameter; T, and S are the linear operators for a clustering type.

$$\mathrm{GAD}_{k,\bar{X}} = \left(\mathrm{RMAD}_{k,1} , \mathrm{RMAD}_{k,2} , \ldots . , \mathrm{R\ MAD}_{k,n}\right) = \alpha T\left(RMAD_{k,1} , \ldots , RMAD_{k,n} ,\right) + (1 - \alpha) S\left(RMAD_{k,1} , \ldots RMAD_{k,n}\right) \quad (18)$$

Once the calculations were defined based on the Degree of Adequacy of Distance of Cauchy, Morales et al. [17] also developed a strategy for the automatic fusion of clusters.

III. OUR SEMI-SUPERVISED LAMDA ALGORITHM (LAMDA-HSCC)

Let X be a sample represented by Equation (19).

$$X = \left((x_1, y_1), \ldots (x_j, y_j) \ldots , (x_n, y_n)\right) \quad (19)$$

Where: x_j descriptor j of X, and y_i defined by Equation (21):

$$y_i \begin{Bmatrix} 0 \ if \ x_i \ descriptor \ unlabeled \\ l_i \ if \ x_i \ descriptor \ labeled \end{Bmatrix} (21)$$

Where l_i is the value of the class that descriptor x_i belongs

Definition 13. Objects grouping. Let P_r a grouping object, described by the Equation

$$\boldsymbol{P_r = (\rho_r , \bar{X}_r , type , name)} \quad (22)$$

Where ρ_r: Centroid of group P_r; $type$: Boolean variable defined by Equation (23)

$$type = \begin{Bmatrix} Cluster \ if \ y_i = 0 \\ Class \ \ if \ y_i = l_i \end{Bmatrix} \quad (23)$$

Where $name$: group identifier's. if $type = class$ then $name = l_r$, otherwise $type = Cluster$ and then $name = id_{Cluster}$ where $id_{Cluster}$ cluster identifier's, P_r is type class or cluster.

Definition 14. The centroid (ρ_r:) of the group (P_r) is the most representative value of the group. In our proposal, the centroid will be the average:

$$\boldsymbol{\rho_r = \frac{\sum x_i}{n}} \qquad \boldsymbol{\forall x_i \in P_r} \quad (24)$$

Where n: number of elements of (P_r)

The macro algorithm of our semi-supervised algorithm, called LAMDA-HSCC, is shown below.

0. Input(x_i; y_i)
1. if $y_i = 0$ then
   a. Run LAMDA-RD
   b. Make group $P_r = (\rho_r , \bar{X}_r , type = cluster , name = id_{cluster})$

   else
   a. Run LAMDA-HAD
   b. Make group $P_r = (\rho_r , \bar{X}_r , type = clase , name = l_i)$
2. Determine neighbour P_{nb} of each group P_r
3. Determine the process for adjusting groups/classes settings
   a. Merger Analysis of groups/class
   b. Analysis of the migration of individuals
   c. Analysis of the División of groups/class

The fundamental idea is the creation of groups P_r conditioned to the type of data, that is, if there are labeled data, then it runs the LAMDA-HAD algorithm (see step 3.2) and then create group P_r (Definition 16); otherwise, it runs LAMDA-RD algorithm (see step 3.3) and then create group P_r.

Next, the last three steps of the algorithm are detailed.

Definition 15. Neighbor Group. The neighbor group P_{nb} of the group P_r is the one that has the second largest value of the Degree of Global Adequacy (GAD_{k,X_r}) for a given variable x_r, where k: Group identifier.

A. Merging of groups/class

The merging process of groups/classes will take place when the following scenarios are met:

- There is an area of overlap between the group P_r and its neighbor group P_{nb}
- The distance between the two groups is less than a user-defined threshold.

The LAMDA approach, and the extended versions used in this work, allow cluster-cluster, class-cluster or class-class mergers. For this, we will use the following definitions:

Definition 16: Overlapping Regions. For the detection of regions of overlap between classes. [20] proposes a method for their detection based on the conditional probabilities of occurrence of the classes, as shown in Equation (25):

$$R(\theta) = \{x_i: P(P_k/x_i) - P(P_j/x_i) < \theta)\}) \quad \forall k \neq j \tag{25}$$

Where: $R(\theta)$: Overlapping region; $P(P_k/x_i)$: Conditional probability of the group of class k given the individual x_i; (θ) Overlapping threshold.

The meaning of Equation (33) is that there is an area of overlap between two groups k and j if the difference between the probabilities of occurrence is less than a threshold, called the overlap threshold (θ).

The key to detecting the overlap region is to find a value for the overlap threshold (θ), which can be determined by a calibration process. Furthermore, in the particular case of the LAMDA algorithm, the probability of occurrence of a particular class is determined by the GAD. Thus, in our case, $P(P_k/x_i) = GAD_{k,i}$, consequently, the determination of the overlap region between the groups P_r and its closest neighbor P_{nb}. In the context of the LAMDA algorithm is expressed as:

$$R(\theta) = \{x_i: (GAD_{r,i}) - (GAD_{nb,i}) < \theta)\}) \quad \forall i \neq j; \forall i \in X_r \tag{26}$$

Definition 17 Number of individuals in the overlapping area. The number of individuals that belong to the area of overlap between the groups P_r and P_{nb}, are those whose distance is less than the probability threshold (θ), which are determined by the Equation:

$$N_l = |R(\theta)| \tag{27}$$

Where: N_l: Cardinality of the set $R(\theta)$

Definition 18 Density of overlapping area. The density D_{k-nb} of the area of overlap between two groups P_r and P_{nb} is determined by the Equation:

$$D_{r-nb} = \frac{N_l}{n_{nb} + n_r} \tag{28}$$

Where: n_{nb}: Number of individuals in the group P_{nb}; n_r: Number of individuals in the group P_r.

Definition 19. Merging groups. To merge two groups, [7] proposed that two groups P_r and P_{nb} are merged if:

$$D_{r-nb} \geq D_t \tag{29}$$

Where, $D_t \in [0, 1]$ is a neighborhood threshold, which can be obtained through a calibration process. A high value of D_t implies a large area of the density of individuals in the area of overlap. Now, in our proposal, a value of D_t is established as the average of D_r and D_{nb}, based on the statistical fact that the average is an unbiased estimator of the true population average. Additionally, for this particular case, the average between two elements will always be the midpoint between both, consequently, this way of calculating the D_t value could allow us to obtain an accurate estimate of the neighborhood threshold, which will be calculated through Equation.

$$D_t = \frac{(D_r + D_{nb})}{2} \tag{30}$$

Definition 20. New merged group. The resulting group after the merging process is determined by the tuple:

$$\boldsymbol{P_{new} = P_r \cup P_{nb} = \{\rho_{new}, \bar{X}_r \cup \bar{X}_{nb}, type, name\}}$$

B. Migration of individuals from one group to another

Similar to the case of merging group, the migration of individuals from one group to another can happen between cluster-cluster, classes-cluster or classes-classes. The migration of individuals from a group occurs when it is determined that an individual does not belong to the provisional assigned group, consequently, it belongs to the closest group. To define this process the following definitions are necessary

Definition 21: High similarity. The individual, $x_j \in P_m$ can have a high similarity with other individuals of the group P_v if:

$$\forall x_j \in P_m; \quad d(x_j; x_i) < \theta_0 \; ; \forall x_i \in P_v \tag{31}$$

Where, θ_0: Threshold of the probability of migration of individuals, defined by the user or determined through a calibration process. From there, we can calculate:

$$\boldsymbol{PGAD_V = \frac{\sum_{x_l \in P_v \cap d(x_l, x_i) < \theta_0} |GAD_{vx_i} - GAD_{mx_l}|}{n_v}} \tag{32}$$

$$\boldsymbol{PGAD_m = \frac{\sum_{x_k \in P_m \cap d(x_k, x_i) < \theta_0} |GAD_{mx_k} - GAD_{vx_l}|}{n_m}}$$

Where: n_v: Number of individuals satisfying the Equation (32), n_m: Number of individuals satisfying the Equation (32).

Definition 22: Migration of an individual. The migration of individuals from a group P_m to a neighboring group P_v happens if:

$$PGAD_v < PGAD_m \cap n_v \geq n_m \tag{33}$$

Definition 23: New group formed. The groups resulting from the migration process P_{mnew} and P_{mnew} are defined by the equations:

$$P_{vnew} = (\rho_{vnew}, \bar{X}_v + x_i, type = cluster, name = cluster_{id})$$
$$P_{mnew} = (\rho_{mnew}, \bar{X}_m - X_i, type = cluster, name = cluster_{id}) \tag{34}$$

Where ρ_{mnew} , ρ_{vnew} are updated according to the updated values of the individuals in each reconfigured group.

C. Division of Groups

Similar to the case of merging groups and migration of individuals, the division of groups, can be executed regardless of the nature of the grouping object, that is, it can be executed

in groups or classes. The division of groups seeks that they are as compact as possible, starting from the premise that within them they are as homogeneous as possible, and between them, they are sufficiently heterogeneous.

Definition 24: Measure of compactness of a group. It is a metric used as a measure of cohesion of a generated group, such that small values indicate better groups. Although there are various techniques that determine the compaction of the group, in our algorithm the sum of squares of the group will be used since it is considered as an unbiased estimator of the group variability, and will be estimated according to Equation (46):

$$SS_r = \frac{1}{N_r}\sum_{x_i \in P_r} |x_i - \rho_r| \quad (35)$$

Let SS_r be the measure of the compactness of group P_r, and SS_{prom} the average compaction of all groups, determined by the Equation:

$$SS_{prom} = \frac{SS_1 + SS_2 + \cdots \ldots + SS_m}{m} \quad (36)$$

Where SS_m:Sum of the square of group m; m: Número de groups formed.

Definition 25 Standard deviation of mean compactness: Given the average compactness, the standard deviation is defined as the average distance between the compaction of each group SS_r $(\forall\, r = 1, .. m)$ with respect to the average SS_{prom} and is calculated by Equation:

$$\sigma_{SS_{prom}} = \sqrt{\frac{\sum_{r=1}^{m} (SS_r - SS_{prom})^2}{m-1}} \quad (37)$$

Now, assuming normality in the distribution of SS_{prom}, in our proposal, we will consider that a group have atypical compactness when the compactness of the group is less than 2 standard deviation from the average compactness

Definition 26: candidate Group to divide. Once the average compactness of all the groups has been calculated, the division of groups should occur in those groups whose compaction has a low atypical behavior to the rest of the compactions. That is, a group (P_r) with compaction SS_r and centroid ρ_r will be divided in two if:

$$SS_r < 2\sigma_{SS_{prom}} \quad (38)$$

Then, once the candidate group to be divided has been determined, the conformation of the individuals for the two new groups will be through the K-means clustering algorithm, (for k = 2).

IV. Experiments

A. Datasets for Classification tasks

With the aim of implementing the LAMDA-HSCC algorithm in semi-supervised tasks in the context of energy consumption of smart buildings, Table 1 shows the datasets to use according to consumption energetic context. All these datasets can be download from [8, 9].

These datasets originally are not ready for machine learning applications, so, we carried out a data preprocessing that consists in the next steps [19]: i) Detection and elimination of missing values; ii) Detection and elimination of outliers, iii) Utilization of Principal Components Analysis (PCA) for conformation of classes in the datasets, iv) Labeled of records according the PCA results. In this way, we formed the initial classes in the datasets, after the steps above mentioned.

Table 1. Dataset used in the Semi-supervised Task

Dataset	Size	Number of descriptors	Characteristics
Chicago-Usage	102	13	Characterization of smart buildings (7 Classes) from the Chicago city according to 13 descriptors of energy consumption
NY-Power	78	2	Characterization of smart buildings from New York City (11 Classes) according to 2 descriptors of energy consumption

B. Metrics to evaluate Semi-supervised learning process

In the work [7] is proposed a metric for semi-supervised tasks that is a linear combination of two measures, one of them for the classification task (accuracy) and the other for the clustering task, which uses cohesion and separation metrics. Therefore, this hybrid metric is a weighted sum:

$$H = \rho\, x\, acc + \beta\, x\, CS \quad (39)$$

Where: ρ: Ratio of classes formed by the system, calculated by Equation (40):

$$\rho = \frac{N_c}{N_c + N_{cl}} \quad (40)$$

Where: N_c: Number of classes formed, N_{cl}: Number of clusters formed, *acc* is the accuracy of the algorithm, *CS* is a relationship between average cohesion and average separation of all clusters in the system, given by Equation (41).

$$CS = 1 - \frac{prom_{j=1 \ldots m}\{Co_j\}}{prom_{j=1 \ldots m}\{Sep_j\}} \quad (41)$$

Similarly, we can define β as the fraction of clusters created by the system, determined by Equation (42).

$$\beta = \frac{N_{cl}}{N_c + N_{cl}} \quad (42)$$

Additionally, [21] proposed another metric to help improve the decision criteria in the semi-supervised algorithms, which is called the SSC Semi-supervised Criterion, expressed by Equation (43):

$$SSC = acc + WB \quad (43)$$

Where (WB) is the proportion between the sum of squares within the cluster, with respect to the sum of squares between clusters, expressed by the Equation (44):

$$WB = \frac{N_{cl}\, x\, SSW}{SSB} \quad (44)$$

When *WB* is closer to zero, the better the clustering, since the sum of squares within the cluster is close to zero, while the intra-cluster sum is big.

For the elaboration of a performance criterion of the proposed metric in the semi-supervised context, we make the following assumptions:

1. Acceptable performance in the classification context will be one whose accuracy is greater than 0.70
2. Acceptable performance in the context of clustering will be one whose wb index is less than 0.3
3. Acceptable performance in the context of semi-supervised will be one whose CSS value is greater or equal to 0.7 due to acc.

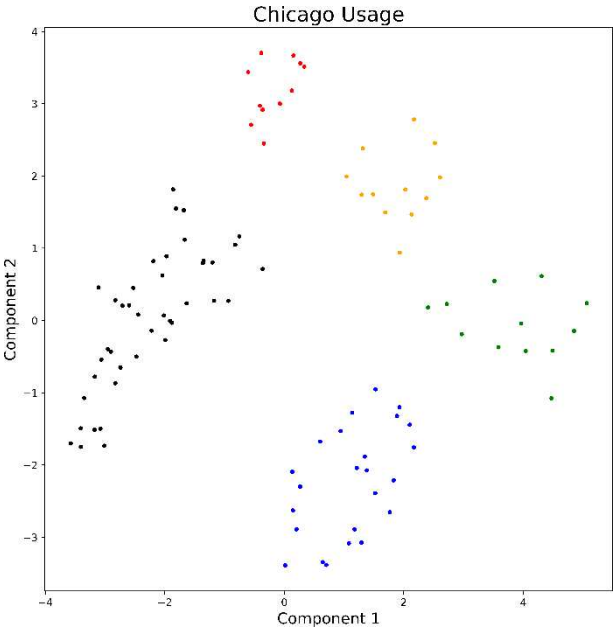
80% of the observations from each dataset will be chosen for training the algorithm and 20% will be used later for validation. Also, the cross-validation technique is used.

C. Results Analysis

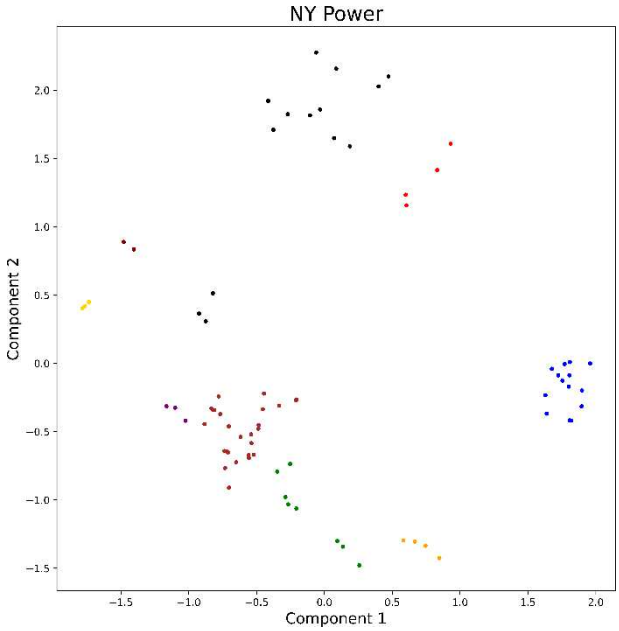
In Fig 1 can be observed the perceptual maps, after the application of PCA, for the datasets studied in this work. For the Chicago Usage dataset (see Fig. 1.a) there are 5 classes. Similarly, in the NY Power dataset there are 10 classes (see Fig. 1.b), but in this case, the classes are not separated clearly, and also there are classes with very few points.

To show the performance in the implementation of the LAMDA-HSCC algorithm in the energy consumption context, in this section, the results of the algorithm are shown. Table 2 shows the results for the datasets in the study, for the 25%, 50% and 75% of labeled data. We can observe that the LAMDA-HSCC algorithm has good performance for the Chicago Usage and NY Power Data, datasets. This can be explained because, with the labeled data, all the original classes in the datasets (Chicago and NY-Power Data) are identified by the algorithm. Then, the merging, partition and migration strategies made the final readjusted of the groups

We can note that in both datasets, according to the SSC criterion, the performance of our proposal is better when the labeled data increase of 25 to 50%. It can be explained because our proposal in the supervised context has a good performance.



(a)



(b)

Fig. 1. Perceptual Map of datasets studied Chicago Usage (a), NY Power (b)

Table 2 SSC and H Index for several datasets in semi supervised learning

Datasets/ métric	25% Labeled		50% Labeled		75% Labeled	
	SSC	Criterion H Index	SSC	Criterion H Index	SSC	Criterion H Index
Chicago Usage	0,724	0,60794	0,7722	0,671712	0,7722	0,70719603

NY- Power Data	0,745	0,72	0,7739	0,748	0,7739	0,748

V. Conclusion

In this work, we implement the semi-supervised LAMDA-HSCC algorithm for several datasets refers to the energetical consumption. The performance of LAMDA-HSCC in the context of smart buildings, the algorithm can identify a lot of situations according to the consumption features. Particularly, it has a good performance with datasets with high dimensionality of groups and features.

This implementation is important because has applications in the real world, where unlabeled data in the energetical context is common, or the process for obtaining the labels is too expensive. Additionally, it allows the design of energy policies based on consumption features.

For future works, it is necessary to design a methodology for calibrating the parameters, which initially are introduced by the user, with the aim to improve the performance of the LAMDA-HSCC algorithm

Acknowledgment

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska Curie Grant Agreement No. 754382. M.D. R-Moreno is supported by the JCLM project SBPLY/19/180501/000024 and the Spanish Ministry of Science and Innovation project PID2019-109891RB-I00, both under the European Regional Development Fund (FEDER).

Disclaimer

The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the authors.

References

[1] Aguilar J., Cerrada M., Mousalli G., Rivas F., Hidrobo F. “A Multiagent Model for Intelligent Distributed Control Systems”. Lecture Notes in Computer Science, vol 3681, pp. 191-197, 2005.

[2] Aguilar J., Jerez M., ExpositoE. Villemur T., "CARMiCLOC: Context Awareness Middleware in Cloud Computing," Proceedings Latin American Computing Conference, 2015.

[3] Aguilar-Martín J. and López De Mantaras R. “The process of classification and learning the meaning of linguistic descriptors of concepts, in Approximate reasoning in decision analysis”, North-Holland, pp. 165–175, 1982.

[4] Araujo M., Aguilar J., Aponte H. “Fault detection system in gas lift well based on artificial immune system”, Proceedings International Joint Conference on Neural Networks, pp. 1673-1677, 2003.

[5] Bourdeau, M., qiang Zhai, X., Nefzaoui, E., Guo, X., & Chatellier, P. “Modeling and forecasting building energy consumption: A review of data-driven techniques”. Sustainable Cities and Society, vol. 48, 2019.

[6] Burrell, J. “How the machine ‘thinks’: Understanding opacity in machine learning algorithms”. Big Data & Society, vol. 3, 2016.

[7] Cerrada, M., Aguilar, J., Altamiranda, J., & Sánchez, R. “A hybrid heuristic algorithm for evolving models in simultaneous scenarios of classification and clustering. Knowledge and Information Systems”, vol. 61, no. 2, 755-798, 2019.

[8] Kaggle.com.NY Power Authority (NYPA) Electric Supply. Available at: <https://www.kaggle.com/new-york-state/ny-power-authority-nypa-electric-supply> [Accessed 21 May 2021].

[9] Kaggle.com. Chicago Energy Usage 2010. Available at: <https://www.kaggle.com/chicago/chicago-energy-usage-2010> [Accessed 21 May 2021].

[10] Kempowsky, T., Subias, A., & Aguilar-Martin, J. ”Process situation assessment: From a fuzzy partition to a finite state machine”. Engineering Applications of Artificial Intelligence,vol. 19, pp. 461-477, 2006.

[11] Lai, W. S., Huang, J. B., & Yang, M. H. “Semi-supervised learning for optical flow with generative adversarial networks”. Proceedings 31st International Conference on Neural Information Processing Systems, pp. 353-363, 2017.

[12] Ligthart A., Catal C., Tekinerdogan B. “Analyzing the effectiveness of semisupervised learning approaches for opinion spam classification”, Applied Soft Computing, vol. 101, 2021.

[13] Liu, Z., Wu, D., Liu, Y., Han, Z., Lun, L., Gao, J., ... & Cao, G. “Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction”. Energy Exploration & Exploitation, vol. 37, pp. 1426-1451, 2019.

[14] Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. “Predicting secondary school students' performance utilizing a semi-supervised learning approach”. Journal of educational computing research, vol. 57, pp. 448-470, 2019.

[15] Marsland, S. “Machine learning: an algorithmic perspective”. Chapman and Hall/CRC, 2014.

[16] Morales, L., Aguilar, J., Chavez, D., & Isaza, C. “LAMDA-HAD, an extension to the Lamda classifier in the context of supervised learning”. International Journal of Information Technology & Decision Making, 2018.

[17] Morales, L., Ouedraogo, C. A., Aguilar, J., Chassot, C., Medjiah, S., & Drira, K. “Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform”. Service Oriented Computing and Applications, vol. 13, pp. 199-219, 2019.

[18] Morales, L., Aguilar, J., “An Automatic Merge Technique to Improve the Clustering Quality Performed by LAMDA”, IEEE Access, vol 8, 2020, 162917-162944.

[19] Pacheco F., Rangel C., Aguilar C., Cerrada M., Altamiranda J., "Methodological framework for data processing based on the Data Science paradigm," Proceedings 2014 XL Latin American Computing Conference, 2014,

[20] Portugal, I., Alencar, P., & Cowan, D. “The use of machine learning algorithms in recommender systems: A systematic review”. Expert Systems with Applications, vol. 97, pp. 205-227, 2018.

[21] Quintero, C., Aguilar, J., “LAMDA-HSCC: a semi-supervised Learning Algorithm based on a Multivariate Data Analysis”. (Under Review), 2021.

[22] Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. “Machine learning approaches for estimating commercial building energy consumption”. Applied energy, vol. 208, pp. 889-904, 2017.

[23] Ruiz, F. A., Isaza, C. V., Agudelo, A. F., & Agudelo, J. R. “A new criterion to validate and improve the classification process of LAMDA algorithm applied to diesel engines”. Engineering Applications of Artificial Intelligence, vol 60, pp. 117-127, 2017.

[24] Waissman J., Sarrate R., Escobet T., Aguilar J., Dahhou B., "Wastewater treatment process supervision by means of a fuzzy automaton model," Proceedings IEEE International Symposium on Intelligent Control, pp. 163-168. 2000.