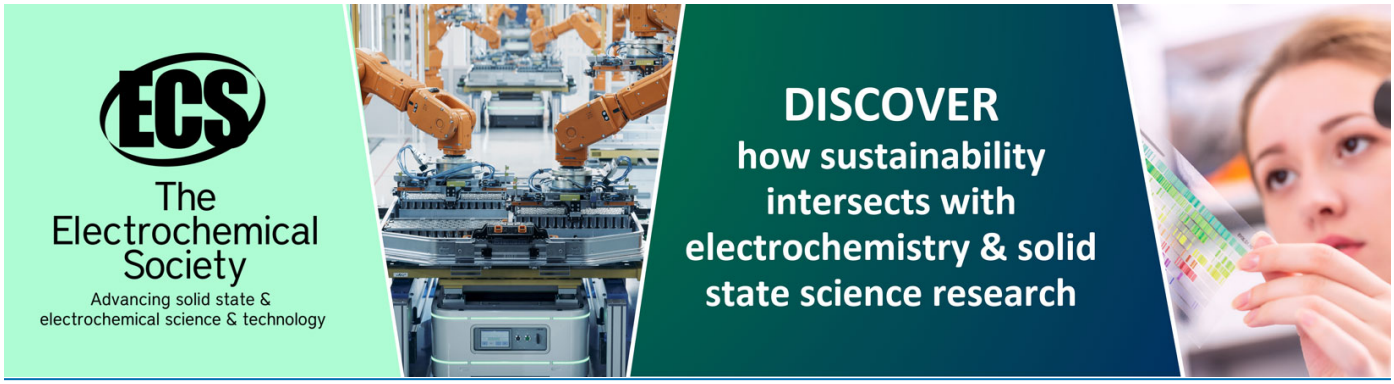
Explainable Clustering Using Hyper-Rectangles for Building Energy Simulation Data



Explainable Clustering Using Hyper-Rectangles for Building Energy Simulation Data

Aviruch Bhatia[1], Vishal Garg[1], Philip Haves[2] and Vikram Pudi[1]

[1]International Institute of Information Technology, Hyderabad, India
[2]Lawrence Berkeley National Laboratory, CA, USA

Abstract. Clustering has become a very popular machine learning technique for identifying groups of data points with common features in a set of data points. In several applications, there is a need to explain the clusters so that the user can understand the underlying commonalities. One such application is in the area of building energy simulation. There is a need to cluster solutions obtained by parametric energy simulation runs and explain the characteristics of each cluster for human consumption. This paper demonstrates how the axis-aligned hyper-rectangles based clustering, on building energy simulation data, can help identify clusters and describe the governing rules for each cluster. We are calling these rules design strategies. Instead of the distance-based clustering methods that are unable to extract simple rules from the underlying commonalities in each cluster, this method is able to overcome this limitation. This method is applied to identify design strategies from a parametric run of a simple five-zone rectangular building model. Based on a user-given threshold, low energy solutions are selected for clustering. Each axis-aligned hyper-rectangle cluster is a unique design strategy that can be easily communicated to the user.

1. Introduction

Parametric building energy simulations are helpful in the optimization of building design parameters. If the number of variables is large, then a large number of combinations will need to be simulated and analyzed. Analyzing a large output data and converting it to decisions is always a difficult task and requires expertise. Machine Learning (ML) techniques can be used to interpret a large amount of data. Researchers are using machine learning systems in a wide domain to generate rules to understand data. In medical science, Castellano et al. [1] have contributed to an approach to discover transparent fuzzy rules from data that are effective in medical diagnosis. Pazzani et al. [2] worked on two datasets for dementia and mental retardation. Authors studied a rule learning system with and without monotonicity constraints and argued that the former supports the results of learning becoming more acceptable to experts [2]. Some researchers apply machine learning techniques to automate rule generation in the construction of intelligent tutoring systems [3]. Machine learning has also been applied in games to generate rules. Ganzfried and Yusuf [4] used machine learning to generate several fundamental rules on poker strategy, which can be easily implemented by humans. Márquez-Chamorro et al. [5] used evolutionary decision rules for business process management systems.

There is always a need for rules that are easily explainable to humans. Narayanan et al. [6] investigated how humans understand explanations from a machine learning system. Kulesza et al. [7] researched about how intelligent agents explain themselves to users. Zeng et al. have focused on explainable artificial intelligence (AI) to explain decisions and conclusions from an AI system. Many researchers are using AI in the building science domain. Wang and Srinivasan [9] have provided a review of AI-based building energy prediction with a focus on ensemble prediction models. Wang et al. [10] used AI in building energy prediction.

There is a lack of research that can provide explainable results in term of rules using clustering algorithms on parametric energy simulation data. In this paper, a method is proposed that is based on identifying clusters bound by axis-aligned hyper-rectangles (AAHR) because hyper-rectangle boundaries can be easily described with simple rules.

2. Problem Statement

From any parametric run of n variables with m values for each variable, we get $S=m^n$ number of simulations known as the design space. These n variables are also referred to as dimensions or features. For each simulation, there is an output known as E. Out of these S solutions, based on the value of E, we select R number of solutions, where $R<S$. We need to cluster R solutions into k clusters (k is not known) such that each cluster can be explained.

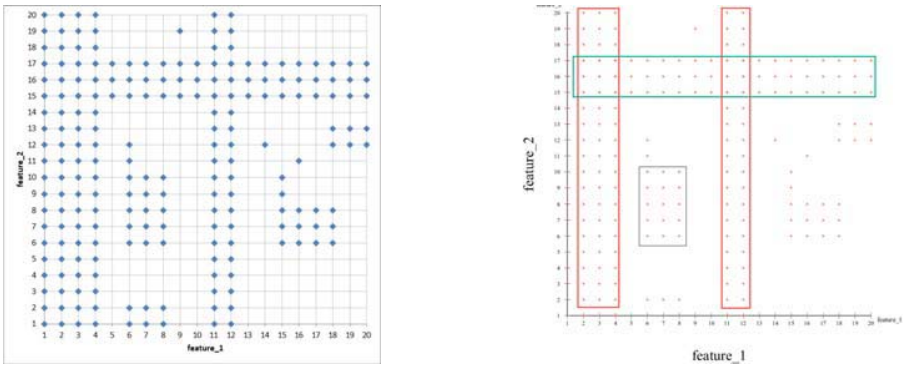
Suppose, we have data with n features, then we need to find clusters in the data such that:

1. They are easily explainable using the following types of expressions:
   $a_1 \leq feature_1 \leq b_1$, $a_2 \leq feature_2 \leq b_2$, and $a_n \leq feature_n \leq b_n$ (clusters are shown in Figure 1(b), where clustering is applied to data as shown in Figure 1(a)).
2. A priori, it is not known how many clusters are in space and need to be discovered.
3. The clusters can overlap.
4. Some data points may not belong to any cluster.
5. The clusters have to be ranked based on the size.
6. The clusters should be selected above a user given threshold.

There are two ways to calculate the size of the clusters:

- Based on the volume of the cluster.
- Based on the number of points in each cluster.

Figure 1(a) shows the example space (R), where clusters need to be identified.



(a) (b)

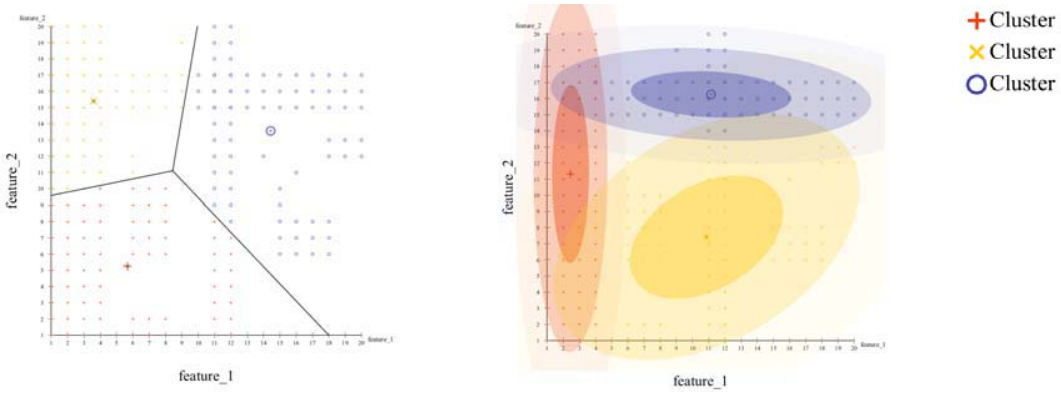
Figure 1(a) & (b). Data points in R; Describable Clusters in R data points

3. Conventional Clustering Algorithms

We applied conventional clustering algorithms to see if they satisfy the requirements given in the problem statement. The following sections give the clusters when K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Expectation Maximization (EM) clustering are applied. All these clustering techniques were applied to the same dataset, as shown in Figure 1(a).

3.1. K-Means

K-means clustering [11] aims at dividing n observations into k clusters, where each observation belongs to the cluster with the nearest mean. Assign each observation to the cluster whose mean has the least squared Euclidean distance - this is intuitively the "nearest" mean.



(a) K-Means clusters (b) EM clusters

Figure 2(a) & (b). Clustering with K-Means and EM method

3.2. EM Clustering

Expectation maximization (EM) [12] is an iterative process that begins with a "naive" or random initialization and then alternates between the expectation and maximization steps until the algorithm reaches convergence. The EM clustering algorithm computes probabilities of cluster memberships based probability distribution. This clustering algorithm maximizes the overall probability or likelihood of the data, given the (final) clusters.

3.3. DBSCAN

Density-based spatial clustering of applications with noise [13] finds core samples of high density and expands clusters from them. This algorithm is good for data with clusters of similar density.

4. Explainable clustering

An explainable AI is an artificial intelligence whose actions can be easily understood by humans [14]. An essential criterion for their explanation is that they must be interpretable. There are several AI methods that can generate rules that are humanly describable, such as Decision Tree [7], Association rules [15], Dirichlet multinomial mixture [16], evolutionary algorithms [17]. One of the methods for explainable clustering is identifying axis aligned hyper rectangles (AAHR) in the data [18]. Generally, hyper-rectangle boundaries can be easily described with simple rules and are easily understandable.
AAHR clustering is applied to the same dataset, as shown in Figure 1(a), and clusters generated are shown in Figure 3(b).

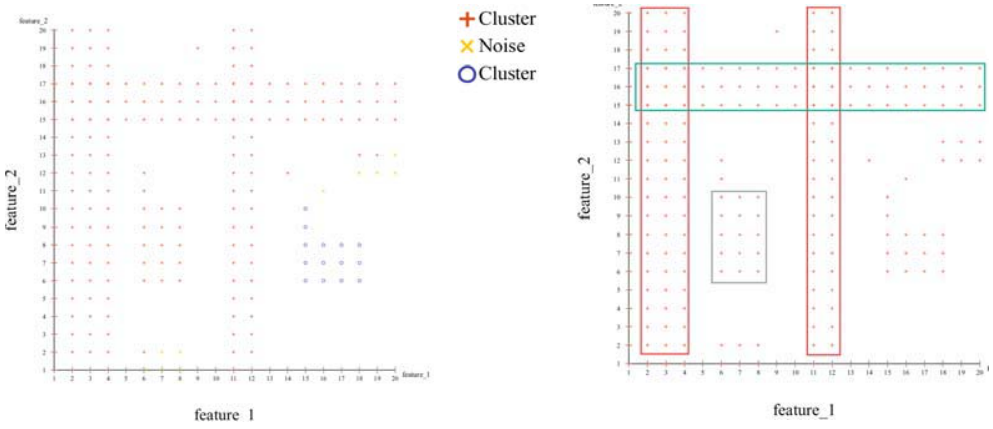
Rules generated from this method are as follows:
One-variable rules

1. $feature_1$: range 2 to 4; size: 54 square units
2. $feature_1$: range 11 to 12; size: 36 square units
3. $feature_2$: range 15 to 17; size: 54 square units

Two variable rules

4. *feature*$_1$: range 6 to 8 and *feature*$_2$: range 6 to 10; size: 8 square units

Furthermore, rules 1 and 2 can be combined as follows:
feature$_1$: range 2 to 7 and 11 to 12



(a) DBSCAN clusters (b) Explainable clusters

Figure 3(a) & (b). Clustering with DBSCAN and Explainable Clusters

5. Comparison of clustering techniques

It can be seen from Figures 2 and 3 that conventional clustering algorithms are unable in providing clusters that are easily explainable. Distance-based clustering methods are unable to extract simple rules from the underlying commonalities in each cluster, whereas explainable clustering can be used where cluster commonalities are easily communicated to users as simple rules. The comparisons of different clustering techniques are shown in Table 1.

Table *1.* Comparison of clustering techniques

Feature	**K-Means**	**DBSCAN**	**EM**	**AAHR**
Need to pre-define the number of clusters	Yes	No	Yes	No
Output clusters overlap	No	No	Yes	Yes
Each point belongs to some cluster	Yes	No	Yes	No
Humanly describable	No	No	No	Yes

6. Algorithms for finding hyper rectangles

Axis-parallel hyper-rectangles provide interpretable generalizations for multi-dimensional data points with numerical attributes [18]. Gao proposed [18] a novel rectangle-based rule learning approach that finds rule sets with small cardinality. For high dimensional data, a faster algorithm is required. Ordonez et al. [19] presented a clustering algorithm to discover low and high-density hyper-rectangles in subspaces of multidimensional data for data mining applications. Eckstein [20] presented an algorithm to find maximum box (containing a maximum number of points) and showed its application in data science. Dumitrescu and Jiang [21] provided an algorithm for finding the size of the maximal empty hyper-rectangles. Lemley et al. [22] provided an algorithm for finding hyper-rectangles in high dimensional data that runs in polynomial time with respect to the number of dimensions. This algorithm discovers large empty holes in the dataset, and the same can be utilized to find AAHR. We have used the algorithm developed by Lemley et al. [22] in this paper, which can provide easily understandable rules from the parametric building energy simulation data.

7. Case studies

A building energy simulation model is prepared and simulated for four locations: London, New Delhi, San Francisco, Singapore, and Abu Dhabi. Building simulation model fixed parameters are provided

in Table 4. Energy simulations have been performed in EnergyPlus v8.6. Input variable building parameters for the case study are provided in Table 2.

Table *2.* Input variable parameters for the case study

Parameter	**Min**	**Max**	**No. of values**
Window to Wall Ratio (WWR) (%)	5	80	16
Wall Insulation (W_Insu) thickness (mm)	1	125	6
Overhang Profile Angle (OPA) (Degree)	2	45	9
Glass ID (G_ID) (Refer Table 3)	1	6	6

Table 3. Glazing parameters for the case study

G_ID	**No. of Panes**	**U-value ($W/m^2.K$)**	**SHGC**	**VLT**
1	Single	6.2	0.80	0.80
2	Single	5.6	0.66	0.65
3	Triple	0.8	0.62	0.73
4	Single	5.6	0.40	0.33
5	Double	1.5	0.27	0.49
6	Double	1.8	0.18	0.32

Table 4. Building model fixed parameters

Component	**Value**
Building Dimensions	20 m x 20 m five zones
Roof U-factor	3.911 W/m^2-K
Lighting Power Density (LPD)	9 W/m^2
Daylight controls	In all perimeter spaces
Equipment Power Density (EPD)	10 W/m^2
Occupancy	10 m^2 /person
HVAC type	IdealLoadsAirSystem
Cooling set point	24 ℃
Heating set point	20 ℃
Schedule	Office, 9 AM to 6 PM

Locations selected are in different climatic zones [23]—Singapore (0A, extremely hot and humid), Abu Dhabi (0B, extremely hot and dry), New Delhi (1B, very hot and dry), San Francisco (3C, warm marine), and London (4A, mixed humid).

Equation (1) is used to find the threshold energy value to find low energy solutions.

$$\text{Cut off Energy} = \text{Min Energy} + (\text{Max Energy} - \text{Min Energy}) \times 20\% \quad \ldots(1)$$

8. Results and discussions

Strategies that were identified by the algorithm for all the mentioned cities are shown in Table 5-7. We can see that clusters are now describable in terms of rules; for example, single-variable rule chooses WWR 5% to 20% for San Francisco. Such rules are easily understandable.

It was found that single-variable rules have been identified for San Francisco only, which has a warm marine climate. For all the other cities, there are no single- variable rules for the given set of inputs, and restriction on at least two variables needs consideration.

Table 5. Strategies identified for San Francisco

Restriction	**Design Freedom**
WWR 5% to 20%	G_ID: Any, OPA: Any, W_Insu: Any
G_ID 6	WWR: Any, OPA: Any, W_Insu: Any
WWR 5% to 35% **AND** G_ID 4 to 6	OPA: Any, W_Insu: Any
WWR 5% to 50% **AND** G_ID 5 to 6	
WWR 5% to 30% **AND** OPA 30 to 45	G_ID: Any, W_Insu: Any
WWR 5% to 35% **AND** OPA 40 to 45	
G_ID 4 to 6 **AND** OPA 30 to 45	WWR: Any, W_Insu: Any

The results for San Francisco are shown in Table 5. It can be seen that a low WWR is one of the strategies for designing an efficient building. If WWR is restricted to 20%, all other studied parameters can be chosen in any range, and the building energy consumption will still lie in the lowest 20% range. Another variable strategy concerns choosing a high-performance glass (G_ID 6). Two variable strategies are low to medium WWR with a high-performance glass and low to medium WWR with a large overhang. Also, for G_ID 4 to 6, large overhang is the restriction.

Table 6. Strategies identified for New Delhi and Singapore

Restrictions in New Delhi	**Restrictions in Singapore**	**Design Freedom**
WWR 5% to 60% **AND** G_ID 6	WWR 15% to 30% **AND** G_ID 5 to 6	OPA: Any W_Insu: Any
WWR 5% to 35% **AND** G_ID 5 to 6	WWR 15% to 55% **AND** G_ID 6	
WWR 5% to 10% **AND** G_ID 3 to 6		
WWR 5% to 15% **AND** OPA 40 to 45	WWR 5% to 10% **AND** OPA 20 to 45	G_ID: Any W_Insu: Any
	WWR 15% **AND** OPA 45	
G_ID 5 to 6 **AND** OPA 35 to 45	G_ID 6 **AND** OPA 15 to 45	WWR: Any W_Insu: Any
	G_ID 5 **AND** OPA 40 to 45	
WWR 5% to 15% **AND** W_Insu 50 mm to 125 mm	WWR 5% to 10% **AND** W_Insu 50 mm to 125 mm	OPA: Any G_ID: Any
	WWR 15% **AND** W_Insu 75 mm to 125 mm	

Table 7. Strategies identified for Abu Dhabi and London

Restriction for Abu Dhabi	**Restriction for London**	**Design Freedom**
WWR 10% to 30% **AND** G_ID 5 to 6	Rules were not found	OPA: Any W_Insu: Any
WWR 5% to 20% **AND** G_ID 3		
WWR 5% to 15% **AND** OPA 40 to 45	Rules were not found	G_ID: Any, W_Insu: Any
G_ID 6 **AND** OPA 25 to 45	Rules were not found	WWR: Any, W_Insu: Any
G_ID 5 **AND** OPA 40 to 45		
WWR 5% to 15% **AND** W_Insu 50 mm to 125 mm	WWR 10% to 15% **AND** W_Insu 100 mm to 125 mm	OPA: Any, G_ID: Any
	WWR 25% **AND** W_Insu 125 mm	

For New Delhi and Singapore, the results are shown in Table 6. It can be seen that for New Delhi, strategies are low to medium WWR with a high-performance glass and low to medium WWR with a large overhang. Also, for WWR in a 5%–15% range, there is need of a 50 mm to 125 mm insulation to keep the building energy consumption in its lowest 20% range. For Singapore, the strategies are low to medium WWR with high-performance glass and low to medium WWR with a large overhang.

Also, for a WWR in the 5%–15% range, there is need of a 50 mm to 125 mm insulation to keep the building energy consumption in its lowest 20% range. The results for London are shown in Table 7, and it can be seen that if WWR is kept between 10% and 15%, then there is a need for a 100 mm to 125 mm external wall insulation. If the WWR requirements in the building are 25%, then the external wall insulation needs to be 125 mm. It can be seen from Table 7 that strategies for Abu Dhabi are low to medium WWR with a high-performance glass and low to medium WWR with a large overhang. Also, for WWR in a 5%–15% range, there is need of a 50 mm to 125 mm insulation to keep the building energy consumption in its lowest 20% range.

The results presented in Tables 5-7 illustrate the method presented in the paper. However, in practice, the allowed ranges for key variables need to be chosen to suit the type of building being defined. For example WWR<20% is only applicable to particular building types, e.g. big box retail, warehouses and prisons.

9. Conclusions

We have used AAHR to find out strategies from parametric building energy simulation data. These strategies are easy to understand, as they have been written down in form of rules. A building energy simulation model has been developed and simulated for five cities in different climates. The strategies identified for all the five cities have been discussed in the paper. We have shown that AAHR is the effective clustering technique for building energy simulation data to produce *humanly explainable rules*—design strategies. There are possibly other clustering methods that are need of further investigation and research.

Acknowledgments

The U.S. Department of Energy (DOE), and the Department of Science and Technology (DST), Government of India (GOI) provided joint funding for work under the U.S. – India Partnership to Advance Clean Energy Research (PACE-R) program's "U.S. – India Joint Center for Building Energy Research and Development" (CBERD) project. The Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technology, State and Community Programs, of the U.S. DOE under Contract No. DE-AC02-05CH11231 supports the U.S. CBERD activity. The DST, GOI, administered by Indo-U.S. Science and Technology Forum, supports the Indian CBERD activity.

References

[1] G. Castellano, A. M. Fanelli, and C. Mencar, "Discovering Human Understandable Fuzzy Diagnostic Rules from Medical Data," *Proc. Third Eur. Symp. Intell. Technol. (ESIT 2003)*, vol. 2, pp. 227–233, 2003.

[2] M. J. Pazzani, S. Mani, and W. R. Shankle, "Acceptance of rules generated by machine learning among medical experts," *Methods Inf. Med.*, vol. 40, no. 5, pp. 380–385, 2001.

[3] M. Jarvis, G. Nuzzo-Jones, and N. T. Heffernan, "Applying Machine Learning Techniques to Rule Generation in Intelligent Tutoring Systems," *Intell. Tutoring Syst. SE - 51*, vol. 3220, pp. 541–553, 2004.

[4] S. Ganzfried and F. Yusuf, "Computing Human-Understandable Strategies: Deducing Fundamental Rules of Poker Strategy," *Games*, vol. 8, no. 4, p. 49, 2017.

[5] A. E. Márquez-Chamorro, M. Resinas, A. Ruiz-Cortés, and M. Toro, "Run-time prediction of business process indicators using evolutionary decision rules," *Expert Syst. Appl.*, vol. 87, pp. 1–14, 2017.

[6] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation," pp. 1–16, 2018.

[7] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models," in *IEEE Symposium on Visual Languages and Human-Centric Computing*, 2013, pp. 3–10.

[8] Z. Zeng, C. Miao, C. Leung, and C. J. Jih, "Building More Explainable Artificial Intelligence with Argumentation," in *The Twenty-Third AAAI/SIGAI Doctoral Consortium Building*, 2015, no. 2014, pp. 1–2.

[9] Zeyu Wang and R. S. Srinivasan, "A review of artificial intelligence based building energy prediction with a focus on ensemble prediction models," in *Proceedings of the 2015 Winter Simulation Conference*, 2015, pp. 3438–3448.

[10] Z. Wang, Y. Wang, and R. S. Srinivasan, "A novel ensemble learning approach to support building energy use prediction," *Energy Build.*, vol. 159, pp. 109–122, 2018.

[11] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. 1, no. 233, pp. 281–297.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[13] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, pp. 226–231, 1996.

[14] "Explainable Artificial Intelligence based Verification & Validation for Increasingly Autonomous Aviation Systems." [Online]. Available: https://sbir.gsfc.nasa.gov/SBIR/abstracts/18/sbir/phase1/SBIR-18-1-A3.02-8802.html. [Accessed: 29-Sep-2018].

[15] D. R. Duling, "Use Machine Learning to Discover Your Rules," 2017, pp. 1–10.

[16] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 2014, pp. 233–242.

[17] R. G. Carter and J. Levine, "An investigation into tournament poker strategy using evolutionary algorithms," *Proc. 2007 IEEE Symp. Comput. Intell. Games, CIG 2007*, pp. 117–124, 2007.

[18] J. B. Gao, "Hyper-rectangle-based discriminative data generalization and applications in data mining," 2007.

[19] C. Ordonez, E. R. Omiecinski, S. B. Navathe, and N. F. Ezquerra, "A Clustering Algorithm to Discover Low and High Density Hyper-Rectangles in Subspaces of Multidimensional Data.," 1999, pp. 1–14.

[20] J. Eckstein, P. L. Hammer, Y. Liu, M. Nediak, and B. Simeone, "The maximum box problem and its application to data analysis," *Comput. Optim. Appl.*, vol. 23, no. 3, pp. 285–298, 2002.

[21] A. Dumitrescu and M. Jiang, "On the Number of Maximum Empty Boxes Amidst n Points," in *32nd International Symposium on Computational Geometry*, 2016, vol. 59, no. 3, pp. 742–756.

[22] J. Lemley, F. Jagodzinski, and R. Andonie, "Big Holes in Big Data: A Monte Carlo Algorithm for Detecting Large Hyper-Rectangles in High Dimensional Data," in *Proceedings - International Computer Software and Applications Conference*, 2017, vol. 1, pp. 563–571.

[23] ASHRAE 90.1, *Energy Standard for Buildings Except Low- Rise Residential Buildings*. Atlanta: American Society of Heating, Refrigerating and Air Conditioning Engineers), 2016.