GMM-PACO: Gaussian Mixture Models and Pareto-based Ant Colony Optimization for Multi-Objective Feature Selection

Anna Krysta, Inès Alaya, and Tristan Cazenave

LAMSADE, Université Paris Dauphine - PSL, Paris, France
anna.krysta@dauphine.eu
ines.alaya@parisnanterre.fr
tristan.cazenave@lamsade.dauphine.fr

Abstract. In this paper, we propose GMM-PACO, a novel Ant Colony Optimization (ACO) for multi-objective feature selection. This algorithm explicitly models each feature using Gaussian Mixture Models (GMMs) to capture distributional characteristics. Our heuristic function integrates mutual information with the Wasserstein distance between feature distributions to evaluate feature relevance and redundancy. GMM-PACO simultaneously optimizes classification accuracy, feature subset size, and redundancy using Pareto dominance. Extensive evaluations on multiple UCI datasets with k-nearest neighbors (KNN) classifier demonstrate improvements over existing methods in both subset compactness and classification accuracy.

1 Introduction

Feature selection (FS) aims to identify the most relevant features in order to improve model performance and reduce complexity. This study focuses on multi-objective FS, aiming to find feature subsets that balance three conflicting goals: maximizing classification accuracy, minimizing redundancy, and minimizing the number of selected features.
Ant Colony Optimization (ACO) [8] is used in this work for multi-objective FS. Existing ACO-based FS methods often rely on simplistic statistical assumptions and lack robust feature modeling or clear convergence criteria, which can lead to suboptimal subsets or unstable performance across datasets. To address these limitations, we propose a novel multi-objective FS framework that integrates Gaussian Mixture Models (GMMs) [3] and Pareto optimization into the ACO search process. Our contributions include:

- A novel multi-objective ACO-based FS framework using Gaussian Mixture Models (GMMs) to model feature distributions.
- An entropy-based stopping criterion that guides the selection process.
- A heuristic function combining mutual information (MI) and Wasserstein distance (WD) [23] to evaluate feature relevance and redundancy.
- A pheromone update rule based on accuracy and normalized by subset size.

The remainder of the paper is structured as follows. Section 2 reviews related work on ACO and FS methods. Section 3 introduces our proposed method, along with definitions and pseudocode. Section 4 presents the experiments and comparative results. Finally, Section 5 concludes with findings and future directions. The code is available at: https://github.com/annkry/GMM-PACO.

2 Related Work

Feature selection (FS) is a key dimensionality reduction technique that selects a subset of original features. Multi-objective FS typically balances competing goals such as minimizing subset size, prediction error, computational cost, or redundancy. Pareto-based approaches are widely used to identify a set of non-dominated solutions, forming the Pareto front, which represents optimal trade-offs among the objectives. These approaches have been successfully applied to classical combinatorial problems such as the multi-objective knapsack problem [2] and the traveling salesman problem with Time Windows [19]. For multi-objective FS problem, PMFS [15] combines relevance and redundancy with crowding distance to maintain diversity. PEFS [16] aggregates filter rankings and applies Pareto selection, while 2OMF [12] uses a two-step approach that combines mutual information with classifier stability.
Nature-inspired algorithms are widely applied in FS for their ability to search large search spaces. A multi-objective Artificial Bee Colony (MOABC) algorithm [14] explores both binary and continuous variants. ACO-based approaches such as ACOFS [17] select features probabilistically, with heuristic updates based on information gain, subset size, and accuracy. FSvACO [10] uses cosine similarity for transitions, while MLACO [21] extends ACO to multi-label FS using feature-class similarity and Pearson correlation. UFSACO [22] applies inverse cosine similarity and and another work [18] proposes a multi-objective ACO using non-dominated solutions for pheromone updates and crowding for diversity preservation. Many ACO-based FS methods rely on simple heuristic functions, which can limit their ability to handle complex data structures and dependencies among features. Gaussian Mixture Models (GMMs) [3] offer a way to better capture such complexity by modeling feature distributions as mixtures of Gaussian distributions. Prior work has applied GMMs in FS outside of metaheuristics. For example, [1] surveys FS techniques for GMMs and Hidden Markov Models, while [11] introduces a GMM-based FS method using a relevance index. However, most of these methods are single-objective and lack global search capabilities.

3 Proposed Method

In this section, we introduce our proposed method for solving the FS problem using the ACO algorithm and describe the pseudocode.

3.1 Overview of GMM-PACO Framework

Algorithm 1 GMM-PACO pseudocode

Input: dataset
Output: Pareto front

```
Preprocessing features
Compute pairwise Wasserstein distances between feature GMMs
Initialize pheromone levels τ_j
while max iterations not reached do
    for each ant k do
        Start at a random feature
        while stopping probability far from uniform do
            Ant k selects a feature with transition probability P_j^k
        end while
    end for
    Evaluate objectives: classification accuracy, subset size, redundancy
    Update Pareto set and pheromone levels
end while
return final Pareto set
```

Algorithm 1 outlines our proposed ACO method. In preprocessing, feature values are discretized using equal-width binning and MI is computed. For each feature, GMM parameters are estimated via the Expectation–Maximization (EM) algorithm [7] until likelihood convergence and pairwise Wasserstein-1 distances between GMMs are approximated using the Sinkhorn algorithm [6]. Feature pheromones are initialized using normalized maximum cosine similarity to class labels. During optimization, each ant incrementally constructs a feature subset until transition probabilities become nearly uniform, guided by a heuristic combining mutual information and Wasserstein-based redundancy. The Pareto set is updated each iteration: the Chebyshev score [25] selects the best solution in the first iteration and Pareto dominance is applied thereafter. Pheromone levels are finally updated using our accuracy-based rule normalized by subset size.

3.2 Gaussian Mixture Models for Features Modeling

We define GMM associated with each feature f_j within a dataset as

$$p_j(x) = \sum_{m=1}^{M} \pi_{jm}\, \mathcal{N}(x \mid \mu_{jm}, \sigma_{jm}^2),$$

where π_{jm} are the mixture weights, $\sum_m \pi_{jm} = 1$, M is the number of mixtures, and

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

is the Gaussian probability density function. Real-world datasets often contain features that have complex and multimodal patterns, making simplistic single-distribution assumptions inadequate [20]. The GMM allows capturing these intricate structures by representing the distribution of each feature through a weighted sum of several Gaussian components.

3.3 Heuristic Function

The heuristic function we define aims at guiding ants towards promising feature selections by balancing two critical aspects: relevance and diversity. Formally, we express it as follows

$$\eta_j(t) = \lambda \cdot \frac{\mathrm{MI}(f_j, Y)}{\max_k \mathrm{MI}(f_k, Y)} + (1-\lambda) \cdot \frac{\mathrm{WWD}(f_j, S(t))}{\max_j \mathrm{WWD}(f_j, S(t))},$$

where λ balances the importance of mutual information, that is, feature relevancy, with feature redundancy, that is, the Wasserstein distance between feature GMM distributions. The heuristic combines two components. The first component assesses the relevance of a feature f_j to the target class label random variables Y by measuring the mutual information (MI) and is defined as $\mathrm{MI}(X,Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$, where $p(x,y)$ is the joint probability of observing $X = x$ and $Y = y$, and $p(x)$ and $p(y)$ are the marginal probabilities of X and Y, respectively. High MI values indicate strong relationships between the feature and the class labels. To ensure numerical stability and fair comparison, MI is normalized by the highest MI observed across all features.
The second component encourages feature diversity within the currently built solution $S(t)$, ensuring that newly selected features are distinct from those already chosen. This diversity is measured using a weighted Wasserstein distance (WWD), defined as $\mathrm{WWD}(f_j, S(t)) = \sum_{k=1}^{|S(t)|} \tilde{w}_k \cdot \mathrm{WD}(p_{t_k}, p_j)$,

where $\mathrm{WD}(p_{t_k}, p_j)$ is the entropic-regularized Wasserstein-1 distance between the GMM of previously selected feature f_{t_k} and the candidate feature f_j. The weighting scheme $\tilde{w}_k$ exponentially decays, giving more importance to recently selected features

$$\tilde{w}_k = \frac{\exp\left(-\gamma_w(|S(t)| - k)\right)}{\sum_{s=1}^{|S(t)|} \exp\left(-\gamma_w(|S(t)| - s)\right)}$$

with $\gamma_w > 0$ controlling the decay rate. A higher γ_w gives greater weight to recent selections, dynamically adapting feature diversity.

3.4 Pheromone Update

Another component of the ACO algorithm is the pheromone levels. Initially, we compute for each feature a similarity score

$$s_j = \max_c \frac{\mathbf{x}_j^\top \mathbf{y}^{(c)}}{\|\mathbf{x}_j\|_2 \, \|\mathbf{y}^{(c)}\|_2} = \max_c \frac{\sum_{i=1}^n x_{ij}\, \delta_{y_i,c}}{\sqrt{\sum_{i=1}^n x_{ij}^2}\, \sqrt{n_c}},$$

where x_{ij} is the value of feature j in sample i, $\delta_{y_i,c} = 1$ if $y_i = c$, and n_c is the number of samples in class c. These scores are then linearly normalized to initialize the pheromones. In each iteration of the ACO algorithm, we update

the pheromone levels and apart from the typical ρ-declined part, we define it to be an accuracy deposit normalized by a number of selected features. Formally, the pheromone update rule is defined as

$\tau_j \leftarrow (1-\rho)\cdot\tau_j + \mathbf{1}_{\{\text{feature } j \in \text{sol}_k\}}\frac{\text{acc}_k}{|\text{sol}_k|}$,

where $\mathbf{1}_{\{A\}} = 1$ if A is true, and $\mathbf{1}_{\{A\}} = 0$ otherwise, ρ is the evaporation rate ($0 < \rho < 1$), sol_k is the set of selected features of k-th ant that had the highest accuracy in a given iteration, $|\text{sol}_k|$ is the number of selected features and acc_k is the accuracy of a classifier using only the selected features for k-th ant.

3.5 Feature Selection Stopping Criteria

In every iteration, an ant chooses a subset of features until the selection probabilities are close to becoming uniform. The motivation is that when all the features have the same probability of being selected, they would not introduce any new information to the currently chosen feature subset, since no feature is significantly different from the remaining features. To monitor this, we measure the entropy of the transition probabilities and define a probabilistic stopping condition based on the level of uniformity. First, we define the normalized entropy for the k-th ant as

$$U_t^k = \frac{-\sum\limits_{j\in R_t} P_j^k(t)\log P_j^k(t)}{\log|R_t|},$$

where R_t is the set of remaining candidate features at a given time t, and $P_j^k(t) = \frac{[\tau_j(t)]^\alpha\cdot[\eta_j(t)]^\beta}{\sum_{l\in N^k(t)}[\tau_l(t)]^\alpha\cdot[\eta_l(t)]^\beta}$ is the probability of the k-th ant selecting feature j, where $N^k(t)$ is the set of features not yet selected at time t and the parameters (α, β) control the influence for pheromone and heuristic information. Now, the probabilistic stopping probability is as follows

$$P_{\text{stop}}^k(t) = \frac{1}{1+\exp\big(-\gamma(U_t^k-\theta)\big)},$$

where $\theta \in [0,1]$ denotes the entropy threshold controlling when the stopping probability becomes significant and $\gamma > 0$ is a scaling parameter that controls the steepness of the stopping probability function. An ant would stop adding new features when a random variable r sampled from a uniform distribution $\mathcal{U}(0,1)$ is smaller than the stopping probability $P_{\text{stop}}^k(t)$.

3.6 Pareto Optimization

In multi-objective feature selection, we consider three objectives: minimizing 1−accuracy after classification with a given feature subset, minimizing the number of selected features, and minimizing redundancy. Redundancy is computed for a subset $S = \{f_{t_1}, f_{t_2}, \ldots, f_{t_m}\}, m = |S|$, as the negative average pairwise

Table 1. Comparison of the average best f-score values of the KNN model on different datasets before and after feature selection using our proposed approach.

	Before feature selection		After feature selection		
Dataset	**# of features**	**f-score**	**# of features**	**f-score**	
WDBC	30	0.907	5.6	**0.949**	(+)
Dermatology	34	0.694	7.8	**0.923**	(+)
Ionosphere	34	0.747	4.8	**0.933**	(+)
Arrhythmia	279	0.122	6.2	**0.302**	(+)
Wine	13	0.742	4.4	**1.000**	(+)
Hepatitis	19	0.395	7	**0.859**	(+)
Spambase	57	0.411	7	**0.459**	(+)
Madelon	500	0.664	11.6	**0.899**	(+)

WD between feature GMMs:

$$\text{subset\_redundancy}(S) = -\frac{1}{\binom{m}{2}} \sum_{1 \le a < b \le m} \text{WD}(p_{t_a},\, p_{t_b}).$$

Our goal is to find a Pareto set of feature subsets. In the first iteration, when the Pareto set is empty, the Chebyshev score [25] is computed as

$\max_i \frac{f_{a,i} - z_i^*}{z_i^{\text{nad}} - z_i^* + \varepsilon}$, where $f_{a,i}$ is the value of objective i for solution a, z_i^* and z_i^{nad} are the ideal and nadir values, and ε ensures numerical stability. The solution with the smallest Chebyshev score is added to the Pareto set. In subsequent iterations, ant solutions are evaluated individually using Pareto domination: a solution is added if it is non-dominated with respect to the current set and any dominated solutions are removed. Consequently, multiple non-dominated solutions may be added in one iteration, while dominated ones are pruned.

4 Experiments

This section presents numerical results and comparisons with existing methods.

4.1 Experimental Setup

We use high-dimensional datasets from the Physics and Chemistry, and Health and Medicine domains of the UCI Machine Learning Repository [9]. Experiments ran for 1000 ACO iterations with 25 ants and parameters $\alpha = 1.0$, $\beta = 4.0$ and $\rho = 0.01$ chosen via a small grid search, while remaining parameters were set through preliminary experiments. Feature subsets were evaluated using a k-nearest neighbors (KNN) classifier [5] with $k = 5$ to compute classification accuracy.

F-score Evaluation We evaluate the average best f-scores of GMM-PACO with KNN before and after feature selection, using a 2/3 training and 1/3 testing split with 5 runs per dataset, as the low variance across runs indicates stable

Table 2. Comparison of GMM-PACO using the KNN model. Each cell reports accuracy followed by the average number of selected features, separated by a semicolon, except for the first column. The second column shows KNN accuracy using all features.

Dataset	5-NN acc	ABC-ER	ABC-Fit2_{2C}	LFS	GSBS	GMM-PACO
Ionosphere	83.02	92.12; 12	91.74; 12	90.48; 6	89.52; 29	**94.15**; **3**
Madelon	64.67	72.91; 252	72.20; 248	71.03; **7**	74.88; 250	**87.17**; 11
Musk 1	76.22	**83.11**; 83	82.32; 81	80.71; 12	82.86; 124	79.09; **11**
Musk 2	96.11	81.52; 82	81.54; 81	82.87; **8**	80.24; 122	**94.44**; 9
Opt Digits	98.70	98.10; 41	98.22; 37	97.86; 32	**98.75**; 38	93.03; **15**
Hill-Valley	54.52	54.13; 48	54.92; 45	55.49; 9	54.40; 95	**56.45**; **6**

performance. Table 1 reports the average best achieved f-scores, with the highest values in bold. GMM-PACO consistently selects significantly smaller feature subsets. Wilcoxon signed-rank tests [24] were applied at two levels: (1) per dataset, comparing best f-scores across runs before and after feature selection (with + indicating significance at $p < 0.05$) and (2) across datasets using average best f-scores. GMM-PACO improved f-scores, with a statistically significant cross-dataset result ($p = 0.008 < 0.05$).

Comparison with Other Methods (Cross-Validation) In Table 2, we compare GMM-PACO with ABC-ER and ABC-Fit2_{2C} [14], LFS [13], and GSBS [4]. They were chosen to represent diverse single-objective feature selection methods, covering both metaheuristic and greedy strategies. The experiments consisted of 30 runs following a 70/30 train-test split, where feature subsets were selected based on the average 10-fold cross-validation accuracy on the training set and test accuracy was reported afterward. GMM-PACO achieves comparable or higher accuracy, while selecting significantly fewer features. These results can be attributed to GMM-PACO's ability to better capture feature redundancy via distributional similarity, which allows the selection of compact yet highly discriminative feature subsets.

5 Conclusion

We proposed GMM-PACO, a feature selection method combining a probabilistic stopping rule, a mutual information– and Wasserstein-based heuristic and a pheromone update favoring high-performing subsets. It was evaluated on multiple UCI datasets using a KNN classifier with hold-out and cross-validation, achieving competitive or superior accuracy with fewer features and statistically significant improvements by Wilcoxon tests. Future work includes exploring other classifiers, refining the heuristic and analyzing dataset conditions where the method performs best or is outperformed.

Disclosure of Interests The authors have no competing interests to declare.

References

1. Adams, S., Beling, P.: A survey of feature selection methods for gaussian mixture models and hidden markov models. Artificial Intelligence Review **52** (10 2019). https://doi.org/10.1007/s10462-017-9581-3
2. Alaya, I., Solnon, C., Ghedira, K.: Ant colony optimization for multi-objective optimization problems. In: 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007). vol. 1, pp. 450–457 (2007). https://doi.org/10.1109/ICTAI.2007.108
3. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)
4. Caruana, R., Freitag, D.: Greedy attribute selection. In: Proceedings of the Eleventh International Conference on International Conference on Machine Learning. p. 28–36. ICML'94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994)
5. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13**(1), 21–27 (1967). https://doi.org/10.1109/TIT.1967.1053964
6. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013), https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B **39**, 1–38 (1977), http://web.mit.edu/6.435/www/Dempster77.pdf
8. Dorigo, M., Stützle, T.: Ant Colony Optimization. Bradford Company, USA (2004)
9. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
10. Eroglu, D.Y., Akcan, U.: An adapted ant colony optimization for feature selection. Applied Artificial Intelligence **38**(1), 2335098 (2024). https://doi.org/10.1080/08839514.2024.2335098
11. Fu, Y., Liu, X., Sarkar, S., Wu, T.: Gaussian mixture model with feature selection: An embedded approach. Computers Industrial Engineering **152**, 107000 (2021). https://doi.org/https://doi.org/10.1016/j.cie.2020.107000, https://www.sciencedirect.com/science/article/pii/S0360835220306707
12. Grandchamp, E., Abadi, M., Alata, O.: A pareto front approach for feature selection. In: Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods. p. 334–342. ICPRAM 2016, SCITEPRESS - Science and Technology Publications, Lda, Setubal, PRT (2016). https://doi.org/10.5220/0005752603340342, https://doi.org/10.5220/0005752603340342
13. Gütlein, M., Frank, E., Hall, M., Karwath, A.: Large-scale attribute selection using wrappers. pp. 332 – 339 (05 2009). https://doi.org/10.1109/CIDM.2009.4938668
14. Hancer, E., Xue, B., Zhang, M., Karaboga, D., Akay, B.: Pareto front feature selection based on artificial bee colony optimization. Inf. Sci. **422**(C), 462–479 (Jan 2018). https://doi.org/10.1016/j.ins.2017.09.028, https://doi.org/10.1016/j.ins.2017.09.028
15. Hashemi, A., Bagher Dowlatshahi, M., Nezamabadi-pour, H.: An efficient pareto-based feature selection algorithm for multi-label classification. Information Sciences **581**, 428–447 (2021). https://doi.org/https://doi.org/10.1016/j.ins.2021.09.052, https://www.sciencedirect.com/science/article/pii/S002002552100983X

16. Hashemi, A., Bagher Dowlatshahi, M., Nezamabadi-pour, H.: A pareto-based ensemble of feature selection algorithms. Expert Systems with Applications **180**, 115130 (2021). https://doi.org/https://doi.org/10.1016/j.eswa.2021.115130, https://www.sciencedirect.com/science/article/pii/S0957417421005716
17. Kabir, M.M., Shahjahan, M., Murase, K.: A new hybrid ant colony optimization algorithm for feature selection. Expert Systems with Applications **39**(3), 3747–3763 (2012). https://doi.org/https://doi.org/10.1016/j.eswa.2011.09.073, https://www.sciencedirect.com/science/article/pii/S0957417411013960
18. Ke, L., Feng, Z., Xu, Z., Shang, K., Wang, Y.: A multiobjective aco algorithm for rough feature selection. In: 2010 Second Pacific-Asia Conference on Circuits, Communications and System. vol. 1, pp. 207–210 (2010). https://doi.org/10.1109/PACCS.2010.5627071
19. Lallouet, N., Cazenave, T., Enderli, C.: Pareto-nrpa: A novel monte-carlo search algorithm for multi-objective optimization (07 2025). https://doi.org/10.48550/arXiv.2507.19109
20. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. The MIT Press (2012)
21. Paniri, M., Dowlatshahi, M.B., Nezamabadi-pour, H.: Mlaco: A multi-label feature selection algorithm based on ant colony optimization. Knowledge-Based Systems **192**, 105285 (2020). https://doi.org/https://doi.org/10.1016/j.knosys.2019.105285, https://www.sciencedirect.com/science/article/pii/S0950705119305805
22. Tabakhi, S., Moradi, P., Akhlaghian, F.: An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence **32**, 112–123 (2014). https://doi.org/https://doi.org/10.1016/j.engappai.2014.03.007, https://www.sciencedirect.com/science/article/pii/S0952197614000621
23. Villani, C.: Topics in Optimal Transportation. Graduate studies in mathematics, American Mathematical Society (2003), https://books.google.pl/books?id=idyFAwAAQBAJ
24. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945), http://www.jstor.org/stable/3001968
25. Zhang, Q., Li, H.: Moea/d: A multiobjective evolutionary algorithm based on decomposition. IEEE Transactions on Evolutionary Computation **11**(6), 712–731 (2007). https://doi.org/10.1109/TEVC.2007.892759