Touch Processing for Terrain Classification with Feature Selection

Stephen D. Liang
Hewlett Packard Enterprise
6280 America Center Dr
San Jose, CA 95002
stephendliang@gmail.com

Abstract

In this paper, we study touch modality data collected by RHex robots in White Sands National Monument, New Mexico, United States. Inspired by the recent advances of partial information decomposition (PID), we make analysis of Interaction Information (II), and propose two new feature selection algorithms, namely Mutual Information and Interaction Information (MI^3) criterion and Mutual Information Difference (MID) criterion. We applied our MI^3 and MID algorithms to feature selection of the robots sensing data, and reduced 12 features to 7 features. Simulation results show that the selected 7 feature data could be successfully used for terrain classification using random forest classifier. Our MI^3 and MID feature selection algorithms perform better than the Mutual Information Maximization (MIM), Joint Mutual Information (JMI), and SVD-QR algorithms in terrain classification.

1 Introduction

Touch is an important sensor modality for robots, and touch sensing has become more prevalent in robotic systems. Robots are often used to explore areas where humans are not able to access, for example, areas with explosive or harmful materials such as nuclear plant, planets without oxygen such as Mars. In this paper, we are interested in studying terrain classification based on touch modality from robotic sensing. Machine learning has been widely applied to terrain classification. In [13], support vector machine (SVM) classifiers were applied to terrain classification for legged robots with passively compliant components, and a hexapod robot was operated in the surf zones of beaches. In [43], a deep learning approach was used for landslides classification based on hyperspectral images, which used a deep belief network and a logistic regression classifier. In [31], wavelet transform was applied to data fusion between LiDAR and QuickBird imagery, and landslide locations were differentiated using object-oriented classification method. In [27], three Indian geographical regions were studied based on Landsat7 satellite images using support vector machine and artificial neural network. In [1], a convolutional autoencoder was applied to PolSAR terrain classification using texture feature fusion.

The contributions of this paper include the following.

1. We propose two feature selection algorithms, Mutual Information and Interaction Information (MI^3) criterion and Mutual Information Difference (MID) criterion.

2. We apply the two criteria to robotic sensing data with touch modality for terrain classification.

NeurIPS 2024 Workshop on Touch Processing: From Data to Knowledge.

2 Touch Modality Data from Robotic Sensing

We used the touch modality data from robots reported in [34]. The robotic sensing data were collected in White Sands (WS) National Monument, New Mexico, United States. The RHex robot was used to sense and collect data in different terrains, for example, dry and loose gypsum sand at a barchan dune, or thick-crusted soil surface in the parabolic area. The datasets are located in [33][32], which consists of 27 files, of which 15 files have category barchan in a data directory WS-barchan [33] and 12 files have category parabolics in a data directory WS-parabolics [32].

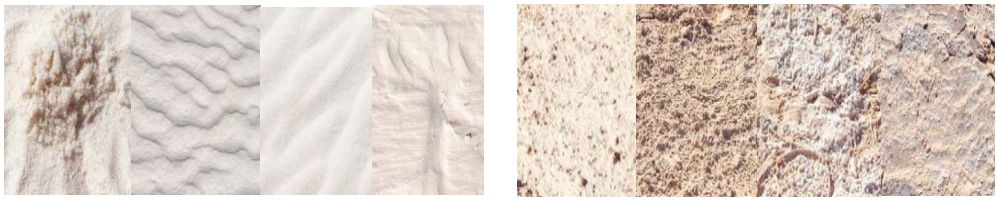
Each data file was named based on the date and time collected. For example, Test_03101430.txt in WS-barchan directory stands for a file of barchan terrain collected at 14:30 on March 10 in White Sands. If we open this file, it contains 2882 rows and 14 columns which means it has 2882 data samples and each data sample is represented by 14 variables. Each robot sensed data file has 14 columns, but has different number of rows which demonstrates the number of data samples of each file is different. The number of collected data samples for each individual barchan file is: {2882 2957 2940 2948 2957 2962 2911 3005 2910 2874 2902 2894 3001 2993 2854}; and the number of collected data samples for each individual parabolics file is: {2999 2862 2931 2894 2794 2841 2903 2833 2901 2894 2996 2978}. Observe that they are not equal. The total number of data samples for barchan is 43,990, and that for parabolics is 34826.

The 14 columns in each robot sensed data file include the following values,

- 1: elapsed time (on PC)
- 2: elapsed time (on robot)
- 3-6: desired toe position (polar – r, alpha; cartesian – x, y)
- 7-10: actual toe position (polar – r, alpha; cartesian – x, y)
- 11-14: force (polar – radial, tangential; cartesian – F_x, F_y)

Columns 11-14 are touch modality data.

We don't consider the first two columns (elapsed time) since they are not related to the terrain. We are interested in selecting features from columns 3-14 as predictors space, and subsequently apply the selected features to terrain classification using machine learning. In Fig. 1, we illustrate different terrains in White Sands presented in [34].



(a) (b)

Figure 1: Different terrains in White Sands [34]. (a) Four examples of barchan terrain; (b) four examples of parabolic terrain.

Given the 12 columns of robotic sensing data, We are interested in studying feature selection and subsequently applying the selected columns to terrain classification.

3 Proposed Feature Selection Algorithms

Our criteria for feature selection incorporates mutual information and interaction information which can be represented as

$$J(\mathcal{S}) = \max_{\mathcal{S}=\mathcal{S}_1 \cup \mathcal{S}_2} \left[\sum_{i \in \mathcal{S}_1} I(X_i; Y) + \sum_{i,j \in \mathcal{S}_2} II(X_i, X_j; Y) \right] \tag{1}$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ are subsets selected for mutual information and interaction information, respectively. If we use $|\mathcal{S}|$ to stand for the number of features in the selected subset $\mathcal{S}$, then $|\mathcal{S}_1| + |\mathcal{S}_2| \geq |\mathcal{S}|$ because there may be some overlaps between the two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$. The mutual information reflects the correlation between a predictor X_i and the target variable Y; the interaction information measures the correlation between two predictors (X_i, X_j) and Y. We call this criterion as Mutual Information and Interaction Information (MI^3) criterion. The details are in the Appendix.

Based on (1), our criteria for feature selection could be simplified further as

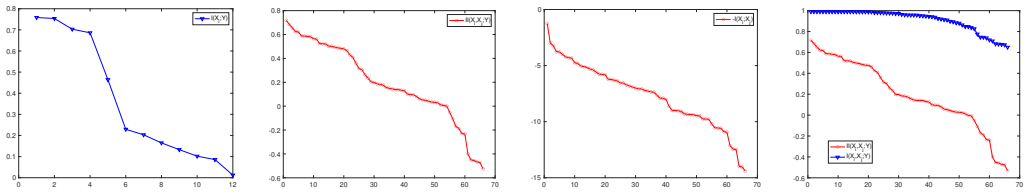
$$J(\mathcal{S}) = \max_{\mathcal{S}=\mathcal{S}_1 \cup \mathcal{S}_2} \left[\sum_{i \in \mathcal{S}_1} I(X_i; Y) - \sum_{i,j \in \mathcal{S}_2} I(X_i; X_j) \right] \tag{2}$$

We call this criterion as Mutual Information Difference (MID) criterion. The result in (2) shows that redundancy $I(X_i; X_j)$ should be removed. This justifies why some existing criteria tried to minimize redundancy, for example, minimal-redundancy-maximal-relevance criterion [30], and MRIDFS method [46]. The details are in the Appendix.

4 Simulation and Performance Analysis

We applied our MI^3 and MID to feature selection in binary terrain classification using Random Forest Classifier. For binary terrain classification, we assign the target variable $Y = 0$ for barchan terrain, and $Y = 1$ for parabolic terrain. For the 27 files of barchan and parabolic terrains, there are totally 78816 tuples, of which 70% tuples (55171 tuples) are used for training, and 30% tuples (23645 tuples) are used for test. We used train_test_split from sklearn in Python, so the training and test datasets are randomly selected.

For our MI^3 algorithm in (1), we need the mutual information $I(X_i; Y)$ and interaction information $II(X_i, X_j; Y)$. In Fig. 2ab, we summarize $I(X_i; Y)$ and $II(X_i, X_j; Y)$ respectively. Since we have 12 features, so $II(X_i, X_j; Y)$ has $\binom{12}{2} = 66$ values. Observe that $II(X_i, X_j; Y)$ goes down slowly, and $I(X_i; Y)$ drops quickly after the 4th value. We choose the top 6 values of $I(X_i; Y)$ and $II(X_i, X_j; Y)$, of which 4 values are from $I(X_i; Y)$, and 2 values are from $II(X_i, X_j; Y)$. For $\mathcal{S}_1$, the selected features (corresponding to the top 4 $I(X_i; Y)$ are $\{14, 11, 13, 6\}$. For $\mathcal{S}_2$, the selected feature pairs (corresponding to the top 2 $II(X_i, X_j; Y)$) are $\{(3, 9), (7, 9)\}$, so the effective features are $\{3, 9, 7\}$ for $\mathcal{S}_2$. Take the union of $\mathcal{S}_1$ and $\mathcal{S}_2$, $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 = \{14, 11, 13, 6, 3, 9, 7\}$. For our MID algorithm in (2), we need to have $I(X_i; Y)$ and $-I(X_i; X_j)$. In Fig. 2c, we summarize the sorted $-I(X_i; X_j)$. Similar to MI^3, we select the top 4 features from $I(X_i; Y)$ and top 3 features from $-I(X_i; X_j)$, and obtain the selected features, $\{14, 11, 13, 6, 5, 9, 10\}$.



(a) (b) (c) (d)

Figure 2: (a) $I(X_i; Y)$, (b) $II(X_i, X_j; Y)$, (c) $-I(X_i; X_j)$; (d) $I(X_i, X_j; Y)$ and $II(X_i, X_j; Y)$.

We compare our approaches with Joint Mutual Information (JMI) [41][26][4] in feature selection. In Fig. 2d, we plot the joint mutual information $I(X_i, X_j; Y)$ and interaction information $II(X_i, X_j; Y)$. Observe that $I(X_i, X_j; Y)$ is very smooth, and doesn't have much difference. The interaction information $II(X_i, X_j; Y)$ drops quickly, which means it is a good indicator for feature selection. We select the top 7 values of JMI $I(X_i, X_j; Y)$, which results in the selected features $\{13, 14, 6, 11, 7, 3, 10\}$. We also compare with Mutual Information Maximization (MIM) [10][17] in feature selection. The mutual information values $I(X_i; Y)$ are plotted in Fig. 2a. Based on the top 7 values of $I(X_i; Y)$, we obtain the selected features $\{14, 11, 13, 6, 12, 4, 8\}$ for MIM.

Further, we compare our approaches with an unsupervised feature selection approach, SVD-QR [12][21]. Based on the training data, we construct a matrix X where each column is from

all sensed data of one feature. We apply SVD to find the maximal singular values of $\mathbf{X}$, and QR to identify the corresponding columns [20]. Based on SVD-QR, we selected important features from the robot sensed information, and the selected features are $\{8, 13, 14, 11, 4, 12, 9\}$.

In Fig. 3, we plot the confusion matrix for random forest classifiers with different feature selections. Observe that MI^3 performs the best, and it achieves almost the same performance as the case when all 12 features are kept.

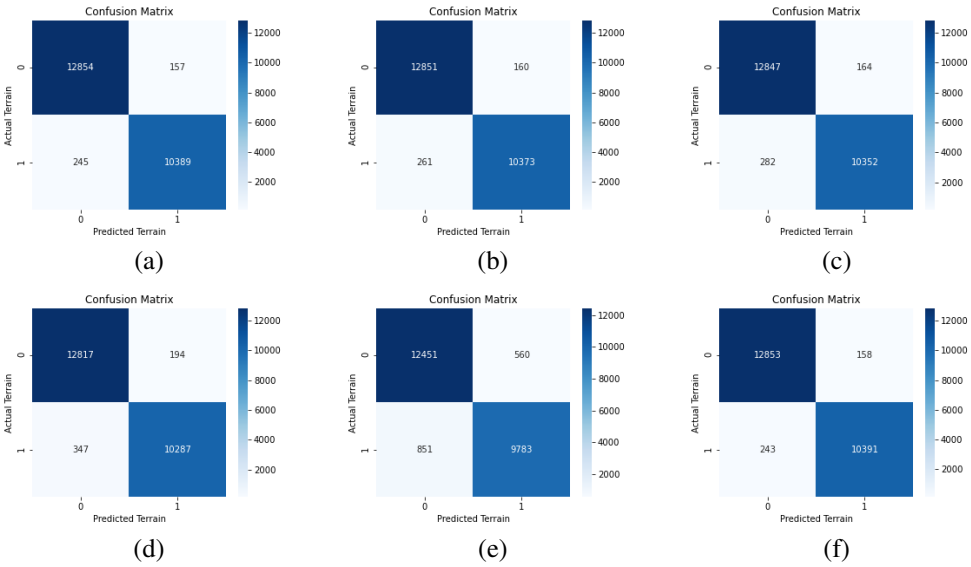


Figure 3: The confusion matrix for random forest classifiers with different feature selections. (a) MI^3, (b) MID, (c) JMI, (d) MIM, (e) SVD-QR, (f) with all 12 features.

In Table 1, we summarize the feature selection results and performance for the six classifiers.

Table 1: The selected features and performance of Random Forest Classifier (RFC) in binary terrain classification.

	Selected Features	RFC Errors
MI^3	$\{14, 11, 13, 6, 3, 9, 7\}$	402
MID	$\{14, 11, 13, 6, 5, 9, 10\}$	421
JMI	$\{13, 14, 6, 11, 7, 3, 10\}$	447
MIM	$\{14, 11, 13, 6, 12, 4, 8\}$	541
SVD-QR	$\{8, 13, 14, 11, 4, 12, 9\}$	1411
All Features	All 12 features	401

5 Conclusions

In this paper, we have presented two feature selection algorithms, MI^3 and MID. The MI^3 is originated from the PID, which consists of four parts, synergetic information, redundant information, and two unique information. Our MI^3 is based on interaction information which is the difference between synergetic information and redundant information. We further simplify the MI^3 and obtain its lower bound. Our MID algorithm is based on the lower bound of MI^3 and is obtained when the synergetic information is zero. We apply MI^3 and MID to terrain classification based on robotic sensing data with touch modality collected in White Sands National Monument, New Mexico, United States. The data has 12 features, and we have reduced it to 7 features using MI^3 and MID algorithms. In terrain classification using random forest classifier, MI^3 could achieve similar performance as the ideal case when all 12 features were kept for classification. They performed much better than the existing approaches, JMI, MIM, and SVD-QR for binary terrain classification.

References

[1] Jiaqiu Ai, Feifan Wang, Yuxiang Mao, Qiwu Luo, Baidong Yao, He Yan, Mengdao Xing, and Yanlan Wu. A fine polsar terrain classification algorithm using the texture feature fusion-based improved convolutional autoencoder. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

[2] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994.

[3] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

[4] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 13(1):27–66, 2012.

[5] Peter Bugata and Peter Drotar. Feature selection based on a sparse neural-network layer with normalizing constraints. *IEEE transactions on cybernetics*, 53(1):161–172, 2021.

[6] Bilian Chen, Jiewen Guan, and Zhening Li. Unsupervised feature selection via graph regularized nonnegative cp decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2582–2594, 2022.

[7] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[8] Behrouz Zamani Dadaneh, Hossein Yeganeh Markid, and Ali Zakerolhosseini. Unsupervised probabilistic feature selection using ant colony optimization. *Expert Systems with Applications*, 53:27–42, 2016.

[9] Gauthier Doquire and Michel Verleysen. Mutual information-based feature selection for multilabel classification. *Neurocomputing*, 122:148–155, 2013.

[10] Włodzisław Duch, I Guyon, M Nikravesh, and S Gunn. Feature extraction: foundations and applications. *Studies in Fuzziness and Soft Computing. Springer: Berlin Heidelberg New York*, pages 89–117, 2006.

[11] Sanghamitra Dutta, Praveen Venkatesh, and Pulkit Grover. Quantifying feature contributions to overall disparity using information theory. *arXiv preprint arXiv:2206.08454*, 2022.

[12] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

[13] John Grezmak, Nicole Graf, Alexander Behr, and Kathryn Daltorio. Terrain classification based on sensed leg compliance for amphibious crab robot. *IEEE Sensors Journal*, 21(20):23308–23316, 2021.

[14] Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In *Guided self-organization: inception*, pages 159–190. Springer, 2014.

[15] G James, D Witten, T Hastie, and R Tibshirani. An introduction to statistical learning, edited by: Casella, g., fienberg, s., and olkin, i. *Springer, doi*, 10:978–1, 2017.

[16] Amol Avinash Joshi and Rabia Musheer Aziz. A two-phase cuckoo search based approach for gene selection and deep learning classification of cancer disease using gene expression data with a novel fitness function. *Multimedia Tools and Applications*, pages 1–32, 2024.

[17] David D Lewis. Feature selection and feature extraction for text categorization. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[18] Zhengxin Li, Feiping Nie, Jintang Bian, Danyang Wu, and Xuelong Li. Sparse pca via p-norm regularization for unsupervised feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5322–5328, 2021.

[19] Jessica E Liang. Partial information decomposition for causal discovery with application to internet of things. *IEEE Internet of Things Journal*, 11(13):24289–24299, 2024.

[20] Stephen D Liang. Optimization for deep convolutional neural networks: How slim can it go? *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(2):171–179, 2018.

[21] Stephen D Liang. Smart and fast data processing for deep learning in internet of things: Less is more. *IEEE Internet of Things Journal*, 6(4):5981–5989, 2018.

[22] Stephen D Liang. Variational autoencoder for data analytics in internet of things based on transfer entropy. *IEEE Internet of Things Journal*, 8(20):15267–15275, 2021.

[23] Stephen D Liang and Jessica E Liang. Interaction information for feature selection with application to terrain classification. In *the 39th Annual AAAI Conference on Artificial Intelligence (AAAI)*, February 2025. Submitted.

[24] Stephen D Liang and Jerry M Mendel. Multimodal transformer for parallel concatenated variational autoencoders. *arXiv preprint arXiv:2210.16174*, 2022.

[25] Ying Liu, Zhen Xu, and Chen Zhang. Distributed semisupervised partial label learning over networks. *IEEE Transactions on Artificial Intelligence*, 3(3):414–425, 2022.

[26] Patrick Emmanuel Meyer, Colas Schretter, and Gianluca Bontempi. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):261–274, 2008.

[27] Amritendu Mukherjee, Arjun Anil Kumar, and Parthasarathy Ramachandran. Development of new index-based methodology for extraction of built-up area from landsat7 imagery: Comparison of performance with svm, ann, and existing indices. *IEEE Transactions on Geoscience and Remote Sensing*, 59(2):1592–1603, 2020.

[28] Xueyan Niu and Christopher J Quinn. A measure of synergy, redundancy, and unique information using information geometry. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 3127–3131. IEEE, 2019.

[29] Xueyan Niu and Christopher J Quinn. Information flow in markov chains. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3442–3447. IEEE, 2021.

[30] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[31] Biswajeet Pradhan, Mustafa Neamah Jebur, Helmi Zulhaidi Mohd Shafri, and Mahyat Shafapour Tehrany. Data fusion technique using wavelet transform and taguchi methods for automatic landslide detection from airborne laser scanning data and quickbird satellite imagery. *IEEE Transactions on Geoscience and remote sensing*, 54(3):1610–1622, 2015.

[32] F. Qian, D. Lee, G. Nikolich, D. Koditschek, and D. Jerolmack. Shear strength data (whitesands - parabolics). `https://doi.org/10.6084/m9.figshare.7056353.v2`, 2018.

[33] Feifei Qian, Dylan Lee, George Nikolich, Daniel Koditschek, and Douglas Jerolmack. Shear strength data (whitesands - barchan). figshare, 2018.

[34] Feifei Qian, Dylan Lee, George Nikolich, Daniel Koditschek, and Douglas Jerolmack. Rapid in situ characterization of soil erodibility with a field deployable robot. *Journal of Geophysical Research: Earth Surface*, 124(5):1261–1280, 2019.

[35] Akash Saxena, Siddharth Singh Chouhan, Rabia Musheer Aziz, and Vani Agarwal. A comprehensive evaluation of marine predator chaotic algorithm for feature selection of covid-19. *Evolving Systems*, pages 1–14, 2024.

[36] José Martínez Sotoca and Filiberto Pla. Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition*, 43(6):2068–2081, 2010.

[37] Li-Min Wang, Peng Chen, Musa Mammadov, Yang Liu, and Si-Yuan Wu. Alleviating the independence assumptions of averaged one-dependence estimators by model weighting. *Intelligent Data Analysis*, 25(6):1431–1451, 2021.

[38] Qianlong Wang, Yifan Guo, Lixing Yu, Xuhui Chen, and Pan Li. Deep q-network-based feature selection for multisourced data cleaning. *IEEE Internet of Things Journal*, 8(21):16153–16164, 2020.

[39] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

[40] Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *Journal of Machine Learning Research*, 24(131):1–44, 2023.

[41] Howard Yang and John Moody. Data visualization and feature selection: New algorithms for nongaussian data. *Advances in neural information processing systems*, 12, 1999.

[42] Abrar Yaqoob, Navneet Kumar Verma, Rabia Musheer Aziz, and Akash Saxena. Enhancing feature selection through metaheuristic hybrid cuckoo search and harris hawks optimization for cancer classification. *Metaheuristics for Machine Learning: Algorithms and Applications*, pages 95–134, 2024.

[43] Chengming Ye, Yao Li, Peng Cui, Li Liang, Saeid Pirasteh, José Marcato, Wesley Nunes Goncalves, and Jonathan Li. Landslide detection of hyperspectral remote sensing data based on deep learning with constrains. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(12):5047–5060, 2019.

[44] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[45] Xunzheng Zhang, Alex Mavromatis, Antonis Vafeas, Reza Nejabati, and Dimitra Simeonidou. Federated feature selection for horizontal federated learning in iot networks. *IEEE Internet of Things Journal*, 10(11):10095–10112, 2023.

[46] HongFang Zhou and Jing Wen. Dynamic feature selection method with minimum redundancy information for linear data. *Applied Intelligence*, 50(11):3660–3677, 2020.

A Related Work

Feature selection is a process that identifies a subset of features to act as predictor variables for machine learning classifiers. Reducing the feature set can simplify classifier design, leading to less computational demand and a simpler model structure. This approach is beneficial for data analytics as it conserves memory and storage.

Feature selection methods are broadly classified into supervised and unsupervised techniques. Supervised methods make use of the target category in selecting features, while unsupervised methods operate solely on the feature data, independent of target categories. The majority of feature selection techniques are supervised. In [44], redundancy in features was defined, and explicit redundancy analysis was introduced for feature selection, employing a correlation-based method to analyze relevance and redundancy. In [5], a neural network approach for feature selection was enhanced with two constraints, resulting in a sparse selection layer. Information-theoretic approaches have also been widely used in feature selection, with methods like Mutual Information Maximization (MIM) [17], which relies on mutual information between each predictor and a target variable, and Mutual Information Feature Selection (MIFS) [2], a generalization of MIM. Joint Mutual Information (JMI) was employed in feature selection in [41] and [26]. In [38], conditional mutual information and entropy formed the basis of a supervised similarity measure for feature selection, while a minimal-redundancy-maximal-relevance criterion was developed in [30] to simplify the implementation of the maximal dependency condition. Conditional mutual information-based clustering methods were proposed in [36] to aid in feature selection, and conditional probability was used in [4] to formulate feature selection as a heuristic for mutual information criteria. A multivariate mutual information criterion with a pruning strategy was proposed by Doquire and Verleysen [9] for improved feature selection. Zhou and Wen [46] developed the Dynamic Feature Selection Method with Minimum Redundancy Information (MRIDFS) by using conditional mutual information to evaluate redundancy among features. All these approaches depend on the target category for analysis.

Unsupervised feature selection methods, in contrast, do not require the target category. An algorithm using ant colony optimization was proposed for unsupervised feature selection in [8]. A nonnegative tensor CP (CANDECOMP/PARAFAC) decomposition-based model for feature selection was presented in [6]. An unsupervised method employing sparse PCA with L_2 and p-norm regularization was introduced in [18]. In [25], partial label learning was studied to address data with ambiguous or absent labels. Two novel feature selection methodologies, SMOCS and CSSMO, which combine Cuckoo Search (CS) with Spider Monkey Optimization (SMO), were introduced in [16]. The Marine Predator Chaotic Algorithm was proposed in [35] for feature selection tasks in COVID-19 data analysis. Metaheuristic Hybrid Cuckoo Search and Harris Hawks Optimization were applied to cancer classification in [42]. Singular Value Decomposition (SVD) and QR methods were employed in principal component analysis in [21] for feature selection. In this approach, SVD identifies the highest singular values of $\mathbf{X}$, while QR selects the corresponding columns, making SVD-QR an unsupervised approach. Recently, unsupervised federated feature selection was explored in horizontal federated learning for IoT networks [45]. This paper also introduces Partial Information Decomposition (PID) in feature selection, presenting two new supervised feature selection algorithms that utilize target category information for training.

PID has been proposed to assess the mutual information between a fixed target variable and multiple predictor variables [39][3]. PID decomposes mutual information into unique, redundant, and synergistic components [39]. Griffith and Koch described synergistic information as the difference between the total information and the combined contributions of its parts [14]. A mutual information decomposition for two sources and one target variable was proposed in [28], but it applied only to variables with exponential distributions. PID was used to analyze information transfer in Markov chains in [29]. Causal discovery using PID in IoT contexts was proposed in [19]. PID was further employed in [11] to assess the contribution of each feature to observed disparities, such as university admission decisions. These PID methods largely emphasize theoretical analysis or simple scenarios like logical operations and admissions decisions. PID was also applied to multimodal transformers as a loss function in [24]. Recently, PID was used in feature selection [40] to maximize relevance and minimize redundancy via conditional mutual information. This paper advances PID analysis, introducing interaction information and deriving a theoretical lower bound to make PID applicable in complex tasks. With this approach, we propose the Mutual Information and Interaction Information

(MI^3) criterion, and by using its lower bound, we introduce the Mutual Information Difference (MID) criterion.

B Partial Information Decomposition: An Introduction

The mutual information between two variables X and Y is defined as [22][7],

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

It is well known that $I(X;Y)$ is always nonnegative.

In recent years, PID was introduced to evaluate the mutual information between multiple variables [39][28][11]. PID can decompose mutual information to redundant information, synergetic information, and unique information. For variables (X_i, X_j) and variable Y, their mutual information can be decomposed as [39]

$$\begin{aligned} I(X_i, X_j; Y) &= R(X_i, X_j; Y) + S(X_i, X_j; Y) \\ &\quad + U(X_i; Y|X_j) + U(X_j; Y|X_i) \end{aligned} \tag{4}$$

where $R(X_i, X_j; Y)$ denotes the redundant information of Y present in both X_i and X_j; $S(X_i, X_j; Y)$ is the synergetic information of Y that is not present in X_i or X_j individually, but present in (X_i, X_j) jointly; $U(X_i; Y|X_j)$ denotes the unique information of Y present only in X_i and not in X_j; and $U(X_j; Y|X_i)$ denotes the unique information of Y present only in X_j and not in X_i. In Fig. 4, we illustrate the relations of these four parts [39][24]. The whole region is $I(X_i, X_j; Y)$; the red region is redundant information $R(X_i; X_j)$; the blue region is synergetic information $S(X_i, X_j; Y)$; the orange and green regions are the unique information $U(X_i; Y|X_j)$ and $U(X_j; Y|X_i)$. Mathematically, they can be represented as [39]

$$R(X_i, X_j; Y) = I(X_i; X_j) \tag{5}$$

$$\begin{aligned} S(X_i, X_j; Y) &= I(X_i, X_j; Y) - I(X_i; Y|X_j) \\ &\quad - I(X_j; Y|X_i) - I(X_i; X_j) \end{aligned} \tag{6}$$

$$U(X_i; Y|X_j) = I(X_i; Y|X_j) \tag{7}$$

$$U(X_j; Y|X_i) = I(X_j; Y|X_i) \tag{8}$$

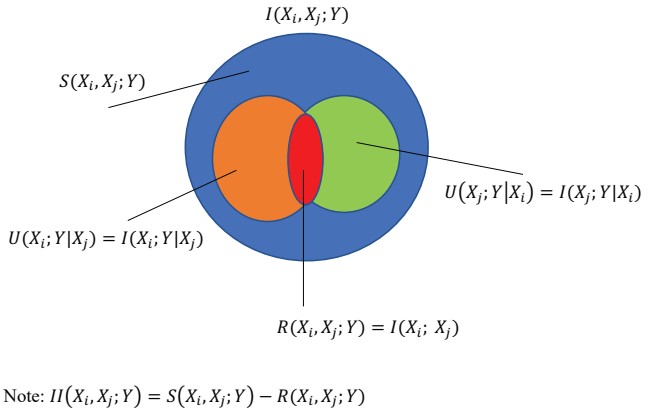


Figure 4: A diagram of PID which consists of redundancy information $I(X_i; X_j)$, synergetic information $S(X_i, X_j; Y)$, and unique information $I(X_i; Y|X_j)$ and $I(X_j; Y|X_i)$ [39]. The Interaction Information $II(X_i, X_j; Y)$ is denoted.

C Proposed Feature Selection Algorithms

C.1 Mutual Information and Interaction Information Criterion

Interaction Information (II) is from PID and is defined as [39][37],

$$\begin{aligned} II(X_i, X_j; Y) &= I(X_i; Y|X_j) - I(X_i; Y) && (9)\\ &= S(X_i, X_j; Y) - I(X_i; X_j) && (10)\\ &= S(X_i, X_j; Y) - R(X_i, X_j; Y) && (11) \end{aligned}$$

The Interaction Information $II(X_i, X_j; Y)$ is denoted in Fig. 4. The expression in (11) demonstrates that $II(X_i, X_j; Y)$ is the difference between Synergetic Information and Redundant Information. It could have the following two cases [39]:

1. $II(X_i, X_j; Y) > 0$, which means the joint knowledge of X_j and X_j enhances the correlation of X_i, X_j, and Y, so X_j and X_i have synergetic information, which means X_i and X_j should be included into the predictors.
2. $II(X_i, X_j; Y) < 0$, which means the joint knowledge of X_i and X_j reduces the correlation of X_i, X_j, and Y, so X_i or X_j is redundant, and X_j or X_j should not be included into the predictors.

We use these two properties to perform feature selection.

Our criteria for feature selection incorporates mutual information and interaction information which can be represented as

$$J(\mathcal{S}) = \max_{\mathcal{S}=\mathcal{S}_1 \cup \mathcal{S}_2} \left[\sum_{i \in \mathcal{S}_1} I(X_i; Y) + \sum_{i,j \in \mathcal{S}_2} II(X_i, X_j; Y) \right] \tag{12}$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ are subsets selected for mutual information and interaction information, respectively. If we use $|\mathcal{S}|$ to stand for the number of features in the selected subset $\mathcal{S}$, then $|\mathcal{S}_1| + |\mathcal{S}_2| \geq |\mathcal{S}|$ because there may be some overlaps between the two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$. The mutual information reflects the correlation between a predictor X_i and the target variable Y; the interaction information measures the correlation between two predictors (X_i, X_j) and Y. We call this criterion as Mutual Information and Interaction Information (MI^3) criterion.

The procedure to obtain the subset $\mathcal{S}$ based on MI^3 can be summarized as follows.

1. Compute the mutual information $I(X_i; Y)$ and the interaction information $II(X_i, X_j; Y)$ for all features.
2. Sort all mutual information and interaction information in descending order.
3. Choose the top M values of all sorted mutual information and interaction information. A selected mutual information is associated with one feature for $\mathcal{S}_1$, and a selected interaction information is associated with two features for $\mathcal{S}_2$.
4. Take the union of subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, and we can get $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$. Then $\mathcal{S}$ is the subset with the selected features.

The above procedure is a hybrid approach which combines mutual information and interaction information. The mutual information involves each individual feature, and the interaction information explores the interaction between features. Selecting features based on their interaction information can lead to better-performing models. By focusing on features that provide better feature interactions can avoid overfitting and improve generalization to new data. This makes MI^3 is very promising to perform better.

C.2 Mutual Information Difference Criterion

We can simplify the interaction information further to obtain a lower bound based on the following theorem.

Theorem 1. *The interaction information, $II(X_i, X_j; Y)$ has a lower bound,*

$$II(X_i, X_j; Y) \geq -I(X_i; X_j) \tag{13}$$

The proof of this Theorem is in [23]. This theorem is easy to understand because when the synergetic information satisfies the following condition,

$$S(X_i, X_j; Y) = 0 \tag{14}$$

the interaction information in (11) has minimum value $-I(X_i; X_j)$.

Based on this theorem, our criteria for feature selection could be simplified as

$$J(\mathcal{S}) = \max_{\mathcal{S}=\mathcal{S}_1 \cup \mathcal{S}_2} \left[\sum_{i \in \mathcal{S}_1} I(X_i; Y) - \sum_{i,j \in \mathcal{S}_2} I(X_i; X_j) \right] \tag{15}$$

We call this criterion as Mutual Information Difference (MID) criterion. The result in (2) shows that redundancy $I(X_i; X_j)$ should be removed. This justifies why some existing criteria tried to minimize redundancy, for example, minimal-redundancy-maximal-relevance criterion [30], and MRIDFS method [46].

Observe (2), $\max \left[-\sum_{i,j \in \mathcal{S}_2} I(X_i; X_j) \right]$ is equivalent to $\min \left[\sum_{i,j \in \mathcal{S}_2} I(X_i; X_j) \right]$. The procedure to obtain the subset $\mathcal{S}$ based on MID can be summarized as the following.

1. Compute the mutual information $I(X_i; Y)$ and the mutual information $I(X_i, X_j)$ for all features.
2. Sort all mutual information $I(X_i; Y)$ in descending order, and sort all $I(X_i; X_j)$ in ascending order.
3. Choose the maximum N_1 features for $\mathcal{S}_1$ based on the sorted mutual information $I(X_i; Y)$, and each selected X_i is associated with one feature; choose the minimum N_2 features for $\mathcal{S}_2$ based on the sorted mutual information $I(X_i; X_j)$, and a selected mutual information is associated with two features.
4. Take the union of subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, and we can get $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$. Then $\mathcal{S}$ is the subset with the selected features.

D More Experiment Results

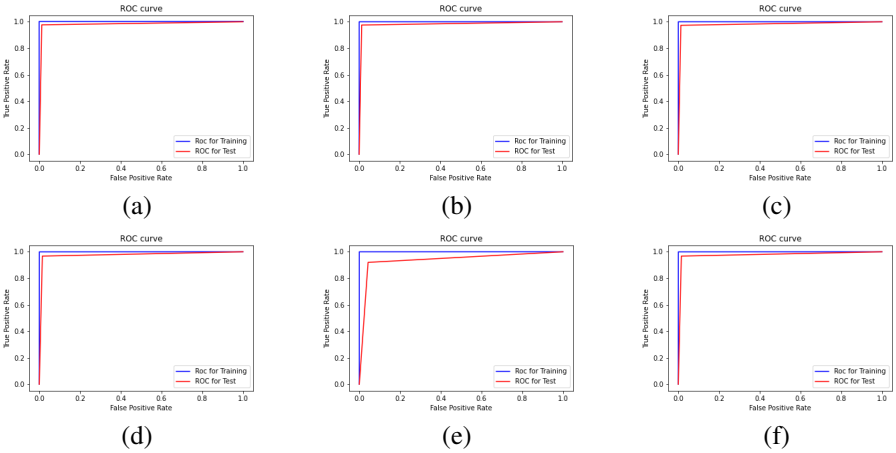
In Fig. 3, we display the confusion matrices for random forest classifiers using various feature selection techniques. Notably, MI^3 demonstrates the highest performance, achieving results nearly identical to those obtained when all 12 features are included. Out of the 23,645 test samples, only $157 + 245 = 402$ errors occur with MI^3, while the scenario with no feature reduction results in $158 + 243 = 401$ misclassifications.

For other methods, MID results in $160 + 261 = 421$ errors; JMI shows $164 + 283 = 447$ errors; and MIM has $194 + 347 = 541$ errors. As an unsupervised method, SVD-QR produces $560 + 851 = 1411$ misclassifications. Compared to JMI, MI^3 reduces errors by $\frac{447-402}{447} = 10\%$, and against MIM, it shows a reduction of $\frac{541-402}{541} = 26\%$. The confusion matrix breaks down the results into True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), which are useful for calculating metrics like accuracy, precision, recall, and F1 score [15].

In Fig. 3a, the FP is 245 and FN is 157, indicating 245 cases where Parabolic terrain was misclassified as Barchan, and 157 cases where Barchan terrain was predicted as Parabolic. This error analysis can help identify shared characteristics among misclassified instances, providing insights into potential model improvements.

Additionally, Fig. 5 illustrates the ROC (Receiver Operating Characteristic) curves for random forest classifiers under different feature selection methods. The ROC curve shows the relationship between the true positive rate (TPR) and false positive rate (FPR) [15]. Ideally, a classifier achieves a TPR of 100% with no false positives. It is observed that, for training data, all ROC curves resemble an ideal classifier, and for test data, the MI^3 and MID methods closely approximate this ideal performance.

Fig. 5 provides AUC (Area Under Curve) values for each ROC test curve. AUC serves as a single metric summarizing model performance across all classification thresholds by measuring the area beneath the ROC curve. This metric considers both the True Positive Rate and False Positive Rate, making it a reliable measure even for imbalanced classes. A model with AUC = 1 represents perfect sensitivity and specificity. Among the five classifiers with feature selection, MI^3 exhibits the highest AUC for the ROC test curve, indicating superior performance.



(a) (b) (c)

(d) (e) (f)

Figure 5: The ROC curve for random forest classifiers with different feature selections. (a) MI^3 with AUC=0.9824, (b) MID with AUC=0.9816, (c) JMI with AUC=0.9804, (d) MIM with AUC=0.9762, (e) SVD-QR with AUC=0.9385, (f) All 12 features with AUC=0.9825.