Feature Selection via Synergistic Integration of Feature Correlation and Neighborhood Decision Mutual Information

WANTING WANG[1], ERIC C.C. TSANG[1]*, WEIHUA XU[2]

[1]Faculty of Innovation Engineering, Macau University of Science and Technology, Taipa, Macau
[2]College of Artificial Intelligence, Southwest University, Chongqing, 400715, PR China
E-MAIL: 957329054@qq.com, cctsang@must.edu.mo, chxuwh@gmail.com

Abstract:

The explosive growth of high-dimensional data presents significant challenges in addressing the curse of dimensionality, where feature redundancy and non-monotonic evaluation in feature selection remain critical issues. This paper proposes a novel feature selection method that integrates feature correlation and neighborhood decision mutual information. We design a dual-criterion evaluation mechanism: leveraging mutual information to measure feature relevance and constructing a non-monotonic framework based on neighborhood decision information. Guided by this theory, the UDI-MI algorithm is developed to dynamically identify discriminative features and filter redundant ones through adaptive thresholds. By integrating decision information measures with exponential decay of mutual information, the algorithm balances classification performance and feature independence. Experimental results demonstrate that the method effectively reduces computational complexity, maintains high classification accuracy, preserves the non-monotonicity of feature subsets, and performs excellently in comprehensive evaluations.

Keywords:

Feature selection; Neighborhood rough set; Feature Correlation; Entropy measures

1. Introduction

The information age has witnessed an explosive growth of high-dimensional data, which provides unprecedented support for scientific and technological advancement while simultaneously introducing the critical challenge of dimensionality curse. This issue has profoundly impacted various cutting-edge research fields including machine learning, data mining, and database systems. With the advancement of sensing technologies and data acquisition methods, contemporary datasets are characterized by exponentially increasing feature dimensions, complex nonlinear correlations among features, as well as prevalent feature redundancy and noise interference. In this context, feature selection techniques, as a crucial preprocessing step, have demonstrated their effectiveness in identifying and eliminating irrelevant and redundant features through systematic evaluation frameworks, achieving optimal dimensionality reduction while preserving the discriminative power of original data. Recent years have seen significant theoretical and methodological breakthroughs in this field, with scholars worldwide making substantial progress in addressing diverse feature selection challenges[3, 10, 13].

Rough set theory has proven to be an effective mathematical framework for handling uncertainty and consistency analysis[9]. Nevertheless, the inherent limitation of classical rough sets in processing only discrete data has motivated extensive research on its extensions. Scholars have developed various enhanced rough set models, including neighborhood rough sets (NRS), fuzzy rough sets, probabilistic rough sets, and variable precision rough sets[5, 6, 11, 12]. Particularly, the neighborhood rough set model overcomes the discrete data constraint by incorporating neighborhood operators, thereby substantially expanding the applicability of rough set theory to complex real-world problems involving continuous data.

Neighborhood rough set, with its adaptive characteristic of neighborhood radius, demonstrates strong adaptability and broad application potential in numerous fields such as data mining, pattern recognition, and artificial intelligence, emerging as an important tool for handling complex data and uncertainty problems[1, 2, 8]. Li et al.[7] estab-

[*] Corresponding author. Email: cctsang@must.edu.mo

lished a novel attribute significance measurement system based on single-attribute subsets and proposed an unsupervised attribute reduction framework using neighborhood dependency, creating a new quantitative benchmark for feature selection in unsupervised learning scenarios. Hu et al.[4] introduced an object overlap degree metric through the k-nearest-neighbor rough set model, which precisely characterizes the spatial coverage and distance relationships between different classes of data. This approach maintains decision approximation capability while significantly improving high-dimensional data processing efficiency . A low-complexity heuristic feature selection algorithm incorporating Fisher Score method was ultimately proposed by Sun et al.[14], who systematically investigated neighborhood entropy uncertainty measures by defining neighborhood credibility and coverage indices and integrating decision neighborhood entropy with mutual information theory. Starting from theoretical interpretation of multi-neighborhood entropy, Zhang et al.[15] constructed a three-dimensional evaluation model combining feature relevance, redundancy, and interaction, thereby developing a dynamic feature selection algorithm adapted to streaming data characteristics. This effectively addresses the core limitations of traditional methods that ignore feature interactions and are incompatible with the temporal nature of streaming data.

The rest of this paper is structured as follows. Section 2 reviews the fundamental theories of neighborhood rough sets, information entropy, and feature evaluation metrics. Section 3 proposes a novel feature selection method integrating feature correlation and neighborhood decision mutual information, including adaptive threshold design and comprehensive evaluation criteria. Section 4 details the UDI-MI algorithm and presents experimental results on UCI datasets, verifying its efficiency and classification performance. Finally, we have summarized this article in Section 5.

2. Preliminaries

In this section, we systematically review the fundamental theories and core concepts of neighborhood rough sets, information entropy, and information gain.

Definition 2.1 In a neighborhood decision system $\mathcal{NDS} = (U, AT \cup D, V, f)$, for any feature subset $B \subseteq AT$, define the joint neighborhood granule of an object $x \in U$ with respect to B and D as:

$$N^{\delta}_{B,D}(x) = NG^{\delta}_{B}(x) \cup [x]_D$$

where $NG^{\delta}_{B}(x)$ denotes the neighborhood granule of x under B, and $[x]_D$ represents the decision equivalence class of x.

Definition 2.2 For an object $x \in U$ and feature subset $B \subseteq AT$, the neighborhood entropy is formulated by the expression.

$$NE^{\delta}_{x}(B) = -\log\left(\frac{|NG^{\delta}_{B}(x)|}{|N^{\delta}_{B,D}(x)|}\right)$$

This measure innovatively employs the joint neighborhood granule to quantify data uncertainty.

Definition 2.3 The average neighborhood entropy of the universe U with respect to B is given by:

$$NE^{\delta}(B) = \frac{1}{|U|}\sum_{x \in U} NE^{\delta}_{x}(B)$$

Moreover, $NE^{\delta}(B)$ satisfies the inequality

$$0 \leq NE^{\delta}(B) \leq -\frac{1}{|U|}\sum_{x \in U}\log\left(\frac{1}{|N^{\delta}_{B,D}(x)|}\right).$$

Definition 2.4 In a neighborhood decision system $\mathcal{NDS} = (U, AT \cup D, V, f)$, for any feature subset $B \subseteq AT$, the neighborhood joint entropy of D and B, neighborhood conditional entropy of D given B, and decision information between D and B are defined by the following formulas, which satisfy corresponding properties:

$$NE^{\delta}(D,B) = -\frac{1}{|U|}\sum_{x \in U}\log\left(\frac{|NG^{\delta}_{B}(x)|^2 \cdot |[x]_D|}{|NG^{\delta}_{B}(x) \cap [x]_D|^2 \cdot |N^{\delta}_{B,D}(x)|}\right)$$

$$NE^{\delta}(D|B) = -\frac{1}{|U|}\sum_{x \in U}\log\left(\frac{|NG^{\delta}_{B}(x)| \cdot |[x]_D|}{|NG^{\delta}_{B}(x) \cap [x]_D|^2}\right)$$

$$DI^{\delta}(D,B) = -\frac{1}{|U|}\sum_{x \in U}\log\left(\frac{|NG^{\delta}_{B}(x) \cap [x]_D|^2}{|N^{\delta}_{B,D}(x)| \cdot |NG^{\delta}_{B}(x)|}\right)$$

As a fundamental criterion in feature selection, information gain measures feature relevance by evaluating the reduction in sample set uncertainty. Focusing on continuous features within high-dimensional datasets, Qu et al.[10] develop an optimized information gain calculation approach incorporating adaptive threshold determination.

For any feature subset $B \subseteq AT$, the information entropy is formally defined as $E(B) = -\sum_{i=1}^{r} p(X_i) \log_2 p(X_i)$ a fundamental measure that serves as the theoretical foundation for the subsequent definition.

Definition 2.5 Given a neighborhood decision system $\mathcal{NDS} = (U, AT \cup D, V, f)$, for any continuous feature $a \in AT$, the information gain $\mathrm{IG}(a)$ is defined as:

$$\mathrm{IG}(a) = \max_{d \in \mathcal{T}_a} \left[H_D(U) - \sum_{s \in \{L,R\}} \frac{|U_d^s|}{|U|} H_D(U_d^s) \right]$$

where $\mathcal{T}_a = \{\frac{v_i + v_{i+1}}{2} \mid v_i \in \mathrm{Sort}(V_a)\}$ denotes the candidate threshold set obtained by computing midpoints between consecutive values of the sorted feature values, $H_D(\cdot)$ represents the conditional entropy with respect to decision attribute D, and U_d^L and U_d^R indicate respectively the left and right subsets partitioned by threshold d, containing samples with feature values less than or equal to and greater than d.

The feature selection process begins with systematically evaluating all features using the Information Gain (IG) metric. After ranking the features in descending order of their IG values, the top-K most discriminative features are selected to form the candidate subset Q. For high-dimensional data, this method significantly improves computational efficiency while maintaining selection accuracy by incorporating specific dimensionality reduction algorithms. The entire process achieves $O(n)$ time complexity, demonstrating excellent performance in handling high-dimensional data.

3. Feature Selection Based on Feature Correlation and Neighborhood Decision Mutual Information

This section proposes a feature selection method fusing feature correlation and neighborhood decision mutual information, aiming to address a series of issues caused by redundant features in high-dimensional data, such as weakened classification performance, insufficient algorithm stability, and surging feature space complexity. First, addressing the limitations of traditional methods in handling feature redundancy, we design a dual-criterion evaluation mechanism: measuring the correlation between features and decision attributes through mutual information, and constructing a non-monotonic evaluation framework by integrating neighborhood decision information. Second, based on this evaluation mechanism, we develop the Unsupervised Decision Information with Mutual Information (UDI-MI) algorithm with correlation-based screening, which dynamically identifies discriminative features and filters out redundant features through adaptive thresholds. Finally, this method preserves the non-monotonicity of feature subsets while achieving collaborative optimization of correlation analysis and classification performance in the feature selection process.

Definition 3.1 In a neighborhood decision system $\mathcal{NDS} = (U, AT \cup D, V, f)$, for any candidate feature $a_i \in AT$ and selected subset $B \subseteq AT$, their mutual information is defined as:

$$\mathrm{MI}(a_i, B) = \sum_{a_i \in V_{a_i}, b \in V_B} p(a_i, b) \log \frac{p(a_i, b)}{p(a_i)p(b)}$$

where V_{a_i} and V_B denote the value spaces of a_i and B respectively, $p(a_i, b)$ is the joint probability distribution, with $p(a_i)$ and $p(b)$ being marginal distributions. This measure quantifies nonlinear feature redundancy via statistical dependency, where higher $\mathrm{MI}(a_i, B)$ values indicate stronger redundancy between a_i and B.

The measure captures nonlinear feature redundancy through statistical dependency, surpassing conventional linear correlation techniques that fail to detect nonlinear associations, which allows for reliable detection of redundant features in diverse data environments.

Definition 3.2 Given the mutual information matrix $\mathbf{MI}_B$ of feature subset B, the adaptive threshold τ is:

$$\tau = \mu_{\mathbf{MI}_B} + k \cdot \sigma_{\mathbf{MI}_B}$$

where $\mu_{\mathbf{MI}_B}$ and $\sigma_{\mathbf{MI}_B}$ are the mean and standard deviation of matrix elements, and k is a tuning factor. The threshold adapts to the mutual information distribution by increasing with growing correlation variability as $\sigma_{\mathbf{MI}_B}$ rises to prevent over-pruning, while decreasing to improve redundancy removal when variability reduces.

Through dynamically adjusting to data-specific correlation distributions, this approach resolves the over-pruning and redundant retention problems of fixed thresholds while providing reliable redundancy control in high-dimensional feature spaces with diverse dependency patterns.

Definition 3.3 For candidate feature $a_j \in AT \setminus B$, we define the comprehensive score $\mathrm{Score}(a_j)$ as follows.

$$\mathrm{Score}(a_j) = \mathrm{DI}^{\delta}(D, B \cup \{a_j\}) \cdot \exp\left(-\lambda \cdot \mathrm{MI}(a_j, B)\right)$$

where $\mathrm{DI}^{\delta}(D, B \cup \{a_j\})$ denotes the decision information measure from UDI, and the balance parameter λ has a default value of 0.8. The exponential decay term suppresses contributions from highly redundant features, ensuring selected features exhibit both high discriminability and low correlation.

Integrating both discriminative power (DI^{δ}) and redundancy (MI), this unified metric achieves adaptive balance between classification performance and feature independence, where the exponential term effectively suppresses highly redundant features to maintain high discriminability and low correlation in the selected features.

Definition 3.4 In a neighborhood decision system $\mathcal{NDS} = (U, AT \cup D, V, f)$, for any candidate feature $B \subseteq AT$, We determine the importance of feature $a_i \in (AT - B)$ through the following expression.

$$\mathrm{Sg}_{out}(a_i, B, D) = \mathrm{DI}^{\delta}(D, B \cup \{a_i\}) - \mathrm{DI}^{\delta}(D, B)$$

When $B = \emptyset$, the feature importance measure simplifies to $\mathrm{Sg}_{out}(a_i, \emptyset, D) = \mathrm{DI}^{\delta}(D, \{a_i\})$. This metric evaluates the discriminative power of individual features when no prior features are selected. More generally, $\mathrm{Sg}_{out}(a_i, B, D)$ measures the marginal contribution of feature a_i to the discriminative information of subset B. The sign and magnitude of this measure capture the complex, non-monotonic relationships between features in high-dimensional spaces. These characteristics enable the Unsupervised Decision Information (UDI) algorithm to implement an adaptive feature selection strategy that goes beyond conventional monotonic approaches.

This definition quantifies the incremental impact of a feature on decision information, capitalizing on the non-monotonicity of DI^{δ} to model intricate feature interactions. It provides a crucial link between feature discriminability and the non-monotonic evaluation framework, enabling the mutual-information-enhanced feature selection approach to balance feature relevance and redundancy during the selection process.

4. Algorithm and experimental analysis

This section presents the UDI-MI feature selection algorithm, which integrates neighborhood decision information and mutual information to balance feature relevance and redundancy. The algorithm first performs high-dimensional preprocessing using Information Gain when the feature dimension exceeds 500, then iteratively selects features by maximizing decision information while pruning redundant features via adaptive mutual information thresholds. Experimental evaluations are conducted to validate the algorithm's effectiveness in feature selection and computational efficiency.

The experimental setup employs a neighborhood radius $\delta = 0.2$ to quantify feature similarity, with the mutual information threshold parameter k optimized within $[0, 2]$ to regulate redundancy pruning through the adaptive threshold $\tau = \mu_{\mathbf{MI}_R} + k \cdot \sigma_{\mathbf{MI}_R}$. For high-dimensional datasets, we preset $K = 200$ features during preprocessing to maintain computational efficiency. To comprehensively evaluate the proposed method, we analyze both the efficiency of feature selection and the preservation of classification performance. Comparative experiments with UDI-IG confirm the superior performance of UDI-MI in computational efficiency and feature selection quality. Experimental results consistently demonstrate that UDI-MI outperforms UDI-IG across all three evaluation dimensions, maintaining robust performance across diverse data characteristics.

The performance of the designed algorithm is tested and evaluated, followed by a detailed analysis of the experimental results. The experiments are conducted using Python on a Windows 10 PC equipped with 16 GB RAM and a 3.10 GHz i5-11300H CPU. Five datasets are selected from the UCI repository (http://archive.ics.uci.edu/ml/datasets.html) for the experiments, and their detailed information is presented in Table 1.

In terms of runtime efficiency, UDI-MI demonstrates significant superiority over UDI-IG on large-scale datasets. For instance, on the Madelon dataset, UDI-MI reduces the

Algorithm 1: UDI-MI Feature Selection Algorithm

```
Input  : Neighborhood decision system
         $\mathcal{NDS} = (U, AT \cup D, V, f)$, neighborhood
         radius $\delta$, mutual information threshold
         parameter $k$, preprocessing feature number
         $K$
Output: Selected feature subset $R$
begin
    Notation: $\mathcal{C}$ denotes candidate feature set, $\mathcal{S}$
      denotes survived feature set after redundancy
      pruning, $f^*$ denotes the optimal feature;
    $R \leftarrow \emptyset$;
    if $|AT| > 500$ then
        $Q \leftarrow$ Top $K$ features selected by
          Information Gain;
        $AT \leftarrow Q$;
    end
    $\mathcal{C} \leftarrow AT$;
    while $\mathcal{C} \neq \emptyset$ do
        for Each $f \in \mathcal{C}$ do
            Compute $\mathrm{DI}^{\delta}(D, R \cup \{f\})$;
        end
        if $R \neq \emptyset$ then
            $\tau \leftarrow \mu_{\mathbf{MI}_R} + k \cdot \sigma_{\mathbf{MI}_R}$;
            $\mathcal{S} \leftarrow \{f \in \mathcal{C} \mid \mathrm{MI}(f, R) \leq \tau\}$;
        end
        else
            $\mathcal{S} \leftarrow \mathcal{C}$;
        end
        if $\mathcal{S} = \emptyset$ then
            $\mathcal{C} \leftarrow \emptyset$;
        end
        else
            $f^* \leftarrow$
              $\left\{f \in \mathcal{S} \mid \mathrm{DI}^{\delta}(D, R \cup \{f\}) = \mathrm{DI}^{\delta}_{max}\right\}$;
            $R \leftarrow R \cup \{f^*\}$;
            $\mathcal{C} \leftarrow \mathcal{C} \setminus \{f^*\}$;
        end
    end
    return $R$;
end
```

TABLE 1. Description of the experimental datasets

No.s	Datasets	Objects	Attributes	Classes
1	dermatology	358	54	6
2	Madelon	2200	500	10
3	waveform	5000	21	3
4	au2-10000	10000	250	2

computation time from 69291.98s for UDI-IG to 36184.70s, achieving a 47.8% runtime reduction. This optimization is attributed to the adaptive redundancy pruning mechanism, which filters features exceeding the dynamic threshold $\tau = \mu_{\mathbf{MI}_R} + 1.5\sigma_{\mathbf{MI}_R}$. As a result, UDI-MI not only accelerates the feature selection process but also ensures the selection of informative and non-redundant features. As shown in Table 2, UDI-MI consistently outperforms UDI-IG in runtime efficiency, particularly in high-dimensional scenarios, effectively mitigating the curse of dimensionality through its preprocessing and dynamic thresholding strategy.

TABLE 2. Attribute reduction time consumption

No.s	Datasets	UDI-IG	UDI-MI
1	dermatology	948.3376	547.7914
2	Madelon	69291.9843	36184.6977
3	waveform	1435.4663	7098.2438
4	au2-10000	23181.2409	10534.2915

Furthermore, the runtime reduction does not compromise classification performance. UDI-MI consistently outperforms UDI-IG in classification accuracy across all datasets, with significant improvements observed in both low-dimensional and high-dimensional scenarios. The adaptive threshold mechanism ensures that critical discriminative features are retained, leading to enhanced classification performance. This demonstrates that UDI-MI effectively balances computational efficiency and classification accuracy, making it a reliable approach for feature selection in high-dimensional data analysis.

TABLE 3. Classification accuracy of KNN classifier

No.s	Datasets	UDI-IG	UDI-MI
1	dermatology	0.7633 ± 0.0391	0.8437 ± 0.0244
2	Madelon	0.4065 ± 0.0219	0.4976 ± 0.0172
3	waveform	0.7152 ± 0.0086	0.7884 ± 0.0048
4	au2-10000	0.6197 ± 0.0094	0.6538 ± 0.0051

Regarding the analysis of remaining attribute quantities after reduction, UDI-MI demonstrates remarkable effectiveness in dimensionality reduction while preserving key features. Compared with UDI-IG, UDI-MI can achieve a higher reduction rate. This significant difference highlights the algorithm's capability to identify redundant features through mutual information constraints. The adaptive thresholding mechanism ensures that only features with low redundancy and high decision information are retained, leading to more compact and discriminative feature subsets. This not only simplifies model complexity but also enhances interpretability, as the reduced feature sets highlight the most informative attributes for classification tasks.

TABLE 4. The number of attributes

No.s	Datasets	UDI-IG	UDI-MI
1	dermatology	27	18
2	Madelon	245	193
3	waveform	20	16
4	au2-10000	143	117

5. Conclusion

This paper presents UDI-MI, a feature selection method integrating feature correlation and neighborhood decision mutual information. It uses mutual information to measure feature relevance and an adaptive threshold to prune redundant features. The algorithm balances classification performance and feature independence via a comprehensive score combining decision information and mutual information. Experiments on UCI datasets show UDI-MI reduces runtime, improves classification accuracy, and achieves higher feature reduction rates than UDI-IG, effectively addressing the curse of dimensionality in high-dimensional data.

Acknowledgements

This paper is supported by the Macao Science and Technology Development Fund (0036/2023/RIA1, 0037/2024/RIA1) and Jiangsu Kaite Auto Parts Co., Ltd. (MUST Code: 1330).

References

[1] T. Al-shami, D. Ciucci, Subset neighborhood rough sets, Knowledge-Based Systems, 2022, 237: 107868.

[2] A. El Atik, A. Wahba, Topological approaches of graphs and their applications by neighborhood systems and rough sets, Journal of Intelligent & Fuzzy Systems, 2020, 39(5): 6979-6992.

[3] J. Chen, P. Zhu, Feature selection of dominance-based neighborhood rough set approach for processing hybrid ordered data, International Journal of Approximate Reasoning, 2024, 167: 109134.

[4] M. Hu, E. Tsang, Y. Guo, et al., Attribute reduction based on overlap degree and k-nearest-neighbor rough sets in decision information systems, Information Sciences, 2021, 584: 301-324.

[5] X. Jia, Z. Tang, W. Liao, et al., On an optimization representation of decision-theoretic rough set model, International Journal of Approximate Reasoning, 2014, 55(1): 156-166.

[6] Q. Kong, C. Yan, W. Xu, Simplified rough sets, Information Sciences, 2025, 686: 121367.

[7] Y. Li, B. Zhang, Z. Yuan, et al., Unsupervised attribute reduction based on neighborhood dependency, Applied Intelligence, 2024, 54(21): 10653-10670.

[8] N. Li, R. Zhou, Q. Hu, et al., Mechanical fault diagnosis based on redundant second generation wavelet packet transform, neighborhood rough set and support vector machine, Mechanical Systems and Signal Processing, 2012, 28: 608-621.

[9] Z. Pawlak, Information systems theoretical foundations, Information Systems, 1981, 6: 205-218.

[10] K. Qu, J. Xu, Q. Hou, et al., Feature selection using Information Gain and decision information in neighborhood decision system, Applied Soft Computing, 2023, 136: 110100.

[11] Y. Qian, J. Liang, Rough set method based on multi-granulations, In: Proceedings of the 5th IEEE International Conference on Cognitive Informatics, 2006, 1: 297-304.

[12] Y. Qian, J. Liang, Q. Wang, et al., Local rough set: a solution to rough data analysis in big data, International Journal of Approximate Reasoning, 2018, 97: 38-63.

[13] T. Shu, Y. Lin, L. Guo, et al., Online hierarchical streaming feature selection based on adaptive neighborhood rough set, Applied Soft Computing, 2024, 152: 111276.

[14] L. Sun, X. Zhang, Y. Qian, et al., Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification, Information Sciences, 2019, 502: 18-41.

[15] G. Zhang, J. Hu, J. Yang, et al., Interactive streaming feature selection based on neighborhood rough sets, Engineering Applications of Artificial Intelligence, 2025, 139: 109479.