Bulletin of Electrical Engineering and Informatics
Vol. 13, No. 5, October 2024, pp. 3644~3652
ISSN: 2302-9285, DOI: 10.11591/eei.v13i5.7938

Enhancing classification in high-dimensional data with robust rMI-SVM feature selection

Fung Yuen Chin[1], **Yong Kheng Goh**[2]
[1]Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Perak, Malaysia
[2]Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Selangor, Malaysia

Article Info

Article history:

Received Nov 25, 2023
Revised Apr 20, 2024
Accepted May 17, 2024

Keywords:

Classification
Feature selection
Machine learning
Mutual information
Support vector machine

ABSTRACT

Dealing with high-dimensional datasets presents notable challenges for classification modelling, primarily due to complexity and susceptibility to overfitting. Traditional feature selection methods frequently struggle to guarantee improved classification performance by including more features. Instead, they often rely on utilising the entire feature set. To address these challenges, a robust feature selection algorithm known as ranked mutual information for support vector machines (rMI-SVM) has been introduced. This approach mitigates the risk of overfitting by selecting features that augment the classification model with additional information, thereby ensuring enhanced performance as more features are selected. rMI-SVM can accommodate datasets with missing values regardless of data linearity as it does not require additional parameters or preset the number of features needed. The proposed method offers a solution to the challenges posed by high-dimensional data, and explicitly identifies the optimal number of features required for a classification model, thus circumventing the necessity of using the full feature set. These findings are supported by receiver operating characteristic (ROC) curves, which highlight the effectiveness of rMI-SVM in outperforming existing baselines and delivering a superior classification model performance.

Corresponding Author:

Fung Yuen Chin
Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman
31900 Perak, Malaysia
Email: chinfy@utar.edu.my

1. INTRODUCTION

Machine learning involves the process of data processing and building the training model [1]. When building the training model, a series of processes is involved in the training strategy, considering factors such as correlation, dependency, and relevancy [2]. However, high-dimensional data sets frequently harbour noise, irrelevant features, and redundant information, thereby diminishing their capacity to effectively discern between different labels. High dimensionality coupled with a limited sample size represents the primary challenge in data analysis. Numerous statistical techniques and machine learning tools are utilised in classification tasks [3]. One common approach employed to mitigate high-dimensional data and extract relevant information is feature selection [4]. Feature selection is categorised into three groups, the filter method, wrapper method and embedded method [5]. The primary concept behind the filter method is to rank the features based on their importance of the features, utilising measurements such as the Pearson correlation coefficient, rank correlation coefficient, chi-square test, and mutual information (MI) [6]-[8]. Over the past two decades, MI has been widely utilised in the feature selection process [9]. MI, conditional mutual information and joint mutual information (JMI) are information measures utilised to gauge the relevance and redundancy between the features and the label class [10]-[12].

Lewis [13] utilised mutual information maximisation (MIM) for feature selection in text classification. Subsequently, the features are then ranked based on expected MI and the selection of features of varying sizes is explored. Similarly, Battiti [14] proposed a greedy algorithm that considers the MI of the output class and selected features named mutual information-based feature selection (MIFS). Robust estimation can be achieved independently of coordinates, but high input dimensions require approximations. In addition to Battiti's work, Kwak and Choi [15] proposed MIFS-U, which comprises two feature selection algorithms utilising the MI method and the Taguchi method. MIFS-U represents an evolution of the MIFS. Yang and Moody [16] also adopted the use of MI and proposed an input selection method that JMI, outperforming existing methods by effectively eliminating input redundancy. Applied to real-world tasks such as finding 2D observation coordinates for data visualisation and selecting inputs for neural network classifiers, it has demonstrated superior capabilities in discovering valuable 2D projections. Peng *et al.* [17] introduced a feature selection method called minimum redundancy maximum relevance (mRMR). This method operates by maximising the correlation between features and their corresponding categories while minimising the redundancy between features [17]. Estevez defines normalised mutual information feature selection (NMIFS) as an improved version of MIFS, MIFS-U, and mRMR [18]. Unlike MIFS, mRMR, and MIFS-U, NMIFS operates independently of any parameters. The selection of parameter values for these methods lacks clear guidance. Later, the genetic algorithm is combined with NMIFS to create a genetic algorithm mutual information feature selection (GAMIFS). The difference between NMIFS and JMI is that NMIFS only considers MI between features and between features related to label classes after removing redundancy.

Hlaing [19] proposed cross-correlation for feature selection in the context of intrusion detection, reducing the number of attributes from 34 to 10, and employing fuzzy decision trees to detect and diagnose attacks. This approach is similar to our research, as both methods aim to enhance the performance of classification algorithms by utilising MI to select the most relevant feature subset for the current task [19]. Hoque stressed that MIFS-ND is also another evolution of the MIFS [20]. The difference between the MIFS and MIFS-ND is that the MIFS algorithm depends on the α while MIFS-ND does not rely on any parameter. Bennasar *et al.* [21] introduce two novel nonlinear feature selection methods: joint mutual information maximisation (JMIM) and normalised joint mutual information maximisation (NJMIM), which address the limitations of information theory-based methods. By utilising MI and the maximum of minimum criterion, these methods demonstrate improved performance in reducing redundant and irrelevant features, outperforming existing methods across multiple datasets. JMIM exhibited superior performance in discrete data compared to NJMIM. Lan proposed a hybrid feature selection method that combines MI with genetic algorithm technology. The objective is to optimise the feature subset, enhance classification accuracy, and reduce feature dimensionality. This approach involves two stages: a filter stage that uses MI to rank features and provide heuristic information, followed by a wrapper stage that uses a genetic algorithm to search for the best feature subset based on classifier performance and dimensionality. Experimental results show that compared to using the genetic algorithm alone, the classification accuracy improves, the feature dimensionality reduces, and the computation time decreases. The proposed method aligns with our study, emphasising the significance of MI in identifying informative features [22].

Wang *et al.* [23] introduced a novel method for feature selection based on MI, aiming to address the challenge of balancing redundant and new classification information. The proposed term, independent classification information, effectively unifies and handles both aspects, aiding in the identification of predictive features with a substantial amount of new information and minimal redundancy. Comprehensive experiments demonstrate that this method can effectively maximise the discriminative performance [23]. Gao *et al.* [24] proposed a dynamic change between the features and the label, termed dynamic change of selected feature (DCSF). Kumar *et al.* [25] proposed a two-stage hybrid intrusion detection method, that combines support vector machine (SVM) and RNN, using JMIM and correlation-based techniques for feature selection. They evaluated the performance on NSL-KDD and Kyoto 2006+ datasets, measuring detection rate, precision, FAR, accuracy, and F-score. Comparative analysis with other hybrid frameworks demonstrates its effectiveness. This study employs MI for feature selection, aiming to improve the accuracy of intrusion detection [25]. Bir-Jmel *et al.* [26] proposed the use of graph theory, Fisher scores, and a modified ant colony optimisation with a local search algorithm. This proposed method aims to reduce the high dimensionality of data. Zhou *et al.* [27] present a weighted conditional mutual information (WCFR) method that utilises standard deviation to balance relevance and redundancy. Alelyani [28] proposes an ensemble approach based on the bagging technique to enhance feature selection stability by reducing data variance. The proposed method aims to address the high dimensionality with a low sample size medical data set [28]. Macedo *et al.* [29] propose a decomposed mutual information maximisation (DMIM) method, which overcomes complementarity penalisation by separately maximising inter-feature and class-relevant redundancies.

In general, the problems encountered during the feature selection can be classified into three domains: i) selecting the most relevant features with minimum redundancy, ii) maximising the new information that is added to the predictive model, iii) determining the minimum number of features in a classification model, iv) identifying better baselines than using all the features, and v) understanding the relationship among the selected features. Several past researchers have attempted to address the first problem by maximising the MI between the feature and class such as MIM, MIFS, MIFS-ND, mRMR, and GAMIFS. However, while selecting the high-relevance features, redundant features might also be chosen simultaneously [30]. The methods mentioned earlier did not consider maximising the increment of the information content, which represents the relationship between the candidate feature with the already selected subset of features and the label class. MIFS-U, JMI, MRI, and DCSF use conditional mutual information and JMI to solve the second problem.

Previous research has encountered challenges in effectively addressing the problems posed by high-dimensional datasets with a limited number of instances. Adding additional features to a predictive model sometimes results in decreased accuracy, contrary to initial expectations, highlighting a lack of clarity on feature correlations [31]. Furthermore, there is a distinct lack of research into determining the optimal number of features required for predictive modelling, leading to uncertainty in feature selection strategies. Existing methods have failed to conduct an in-depth study of the minimum features needed in a predictive model. To date, the features used to construct a classification model remain a prominent topic in most machine-learning tasks. Additionally, the baseline of the model often includes all features and the researchers consistently always compare their findings with previous results. In light of these shortcomings, this paper proposes an improved algorithm specifically designed for classification tasks involving high-dimensional datasets with limited instances. This algorithm uses feature ranking to determine the most effective subset of features for building a predictive model. By integrating MI scores to rank features, the algorithm aims to capture the most relevant subset of features while reducing redundancy and maximising the information content of each added feature [32], [33]. In addition, this algorithm utilises a SVM classifier to reduce the dimensionality of the data set [34], [35]. This approach differs from previous methods that relied on full features and instead prioritises using the fewest features to create the most effective model.

The proposed method offers key benefits by addressing an identified research gap through the proposition of a systematic approach to feature selection in high-dimensional datasets with limited instances. By leveraging MI ranking and a SVM classifier, this approach aims to enhance the efficiency and accuracy of predictive models while minimising the required number of features. This underscores the significance of the proposed method in advancing the field of feature selection and classification in challenging dataset scenarios. Another advantage of this method is its suitability for data sets with missing values and non-linear data. Comparative analysis with existing methods, combined with evaluation indicators such as the confusion matrix and receiver operating characteristic (ROC) curve, illustrates the superiority of this method in prediction performance and feature correlation. The accuracy of the classification model is compared to the regression model, and the mRMR, JMI, and JMIM. The remainder of the paper is organised as follows: section 2 provides an in-depth study of the evolution of MI in feature selection. In section 3, the proposed algorithm is presented. Section 4 discusses the experimental results. Finally, the conclusion and future recommendations are given in section 5.

2. METHOD

2.1. The optimal baseline based on ranked features

MI serves as a measure of the interdependence between two random variables. In feature selection, MI quantifies the relationship between input features and target variables in classification. It assesses how much information the presence or absence of a specific feature provides about the label. The strength of MI in feature selection lies in its capability to capture both linear and nonlinear relationships between features and labels, rendering it applicable to a broad spectrum of problems. Unlike linear correlation, which only captures linear relationships, MI can detect more complex dependencies.

This study utilises MI to select the most relevant features for constructing the predictive model. Consequently, a prediction model constructed from these selected features is expected to exhibit better predictive power compared to using the full set of features. The MI between two discrete random variables X and Y is defined as (1):

$$I(X;Y) = \sum_{y} \sum_{x} p(x,y) \log \left[\frac{p(x,y)}{p(x)p(y)}\right] \tag{1}$$

where $p(x)$ is the probability mass function of the random variable X, $p(y)$ is the probability mass function of the random variable Y, and $p(x, y)$ is the joint probability mass function of the random variable X and Y.

An optimal baseline is established by ranking features based on MI. When features are ordered according to MI values, the most significant features are closely associated with the label class, providing sufficient representation. Evaluating the accuracy of these ranked features in the model demonstrates that the model's performance improves as more ranked features are included. This perspective suggests that there exists a specific number of ranked features, or a few points, indicating optimal performance for classification.

Ranking relevant features based on MI corresponds to prioritising them according to their relevance to the label class. Incorporating these relevant features into a classification is consistently superior to utilising all features, thereby establishing an enhanced baseline. In (1) is used to compute the MI scores of all features concerning the label class. The dataset is represented as a $M \times N$ matrix, where N represents the number of instances and M represents the number of features. Real experimental data were normalised to the range (-1,1) before calculating the MI score. For a feature set $X = \{x_1, x_2, \dots, x_N\}$ within a dataset D with N samples and M dimensions, MI is computed for each x_i associated with the label class.

Following this, features are sorted in descending order according to their MI scores. Subsequently, a graph is plotted to depict the accuracy of the ranked features versus the cumulative count. The peak accuracy on the graph identifies the optimal baseline while indicating the corresponding feature cutoff number. Figure 1 illustrates a flowchart of the process for determining the optimal baseline and feature-specific cutoff numbers. The evaluation of the ranking feature measure involves the utilisation of a SVM classifier, and the accuracy for this study is defined as (2):

$$Accuracy = 1 - \frac{false\ negative + false\ positive}{false\ negative + false\ positive + true\ positive + true\ negative} \tag{2}$$

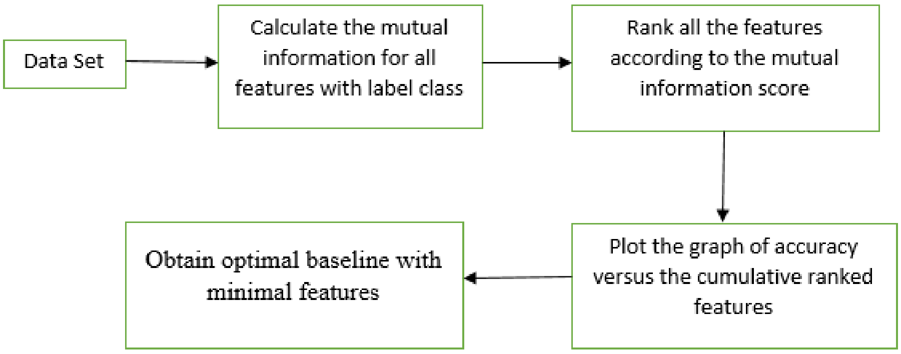

Data Set
Calculate the mutual information for all features with label class
Rank all the features according to the mutual information score
Plot the graph of accuracy versus the cumulative ranked features
Obtain optimal baseline with minimal features

Figure 1. Flowchart to determine optimal baseline and feature-specific cutoff values

The proposed method aims to establish the optimal baseline for a given dataset. This baseline serves as a reference point to evaluate the performance of the predictive model in comparison to the traditional approach, which relies on the accuracy derived from all features as a baseline. By identifying this optimal benchmark, which depends on features ranked by their relevance to labels, it becomes possible to better measure the effectiveness of the predictive models. Through the use of MI, irrelevant features and noise are filtered out. This underscores the importance of prioritising the most relevant features in accurately assessing and enhancing the performance of predictive models.

2.2. Ranked mutual information with support vector machine algorithm

The methods discussed previously outline strategies for determining an optimal baseline to use as a new benchmark for a predictive model. While MI effectively identifies features that are highly relevant to the label class, it also tends to select highly correlated features. These correlated features are redundant and do not add extra information to the predictive model. Furthermore, it will also increase the model's complexity while reducing its predictive power. Therefore, filtering out these redundant features is crucial for improving the performance of the model.

In this study, the feature relevancy and feature redundancy are defined as follows:

Definition 1: (feature relevancy). Feature x_i is more relevant to the label class C than the feature x_j if the MI between the feature x_i and label class C is greater than the MI between feature x_j and label class C.

Definition 2: (feature redundancy). Feature x_i is a redundant feature to feature x_j with respect to the label class C if the MI between the feature x_i and label class C is equal to the MI between feature x_j and label class C.

MI proves to be a powerful tool for identifying the relevant features from vast and high-dimensional datasets. However, as per definition 2, relying solely on the MI score does not allow for differentiation between relevant and redundant features. While the MI score in the previous section effectively ranks the relevance among features concerning the label class, it fails to provide information about the features themselves. Therefore, removing the redundant features becomes an essential step in reducing the dimensionality of the high-dimensional datasets.

The rMI-SVM algorithm is proposed to enhance the performance of predictive models by identifying and filtering out redundant features. One of the advantages of the rMI-SVM algorithm is its ability to select features that significantly improve model performance. This improvement is achieved by dynamically evaluating the impact of newly added features versus existing features and the label class. Notably, the algorithm achieves this without resorting to overly complex calculations, speeding up the feature selection process.

Steps:

1) Let X be the initial set of N features.
2) For each feature x_i in X, normalise its value to the range [-1, 1].
3) Divide each feature x_i in X into three equal bins.
4) Compute the joint probability density function (pdf) for each feature x_i and label C.
5) For each feature x_i in X, calculate MI between x_i and the class label C using the joint pdf.
6) Sort the MI scores of each feature x_i in X in descending order in the set S.
7) According to the dynamic change criterion, select the features with the highest MI score and remove redundant features.
8) Evaluate the model performance using features from set S.
9) If the model performance improves or remains the same, continue using the selected features. Otherwise, revert to the previous set of features.
10) Repeat until all ranked features are evaluated.
11) Identify the highest accuracy achieved and the corresponding number of features.

Compared to using the entire set of features as a baseline, the proposed algorithm yields a reduced number of features, significantly enhancing the performance of the predictive model. This improved performance is achieved with a smaller feature set and is comparable to or better than the performance of the new proposed baseline derived from the ranked features in section 2.1. At the same time, the algorithm ensures that each newly added feature provides unique information to the predictive model, resulting in a graph depicting increasing accuracy relative to the number of features. Features selected by the algorithm guarantee continued enhancement of predictive model performance as more features are added.

Figure 2 illustrates the flowchart for searching a smaller selected feature. The accuracy for the ranked cumulative features is calculated, after which the feature will be deleted based on the comparison of the accuracy. A graph can be plotted using the accuracy versus the ranked cumulative features. The performance metrics used are; i) test accuracy, ii) area under curve (AUC), iii) recall, and iv) precision.

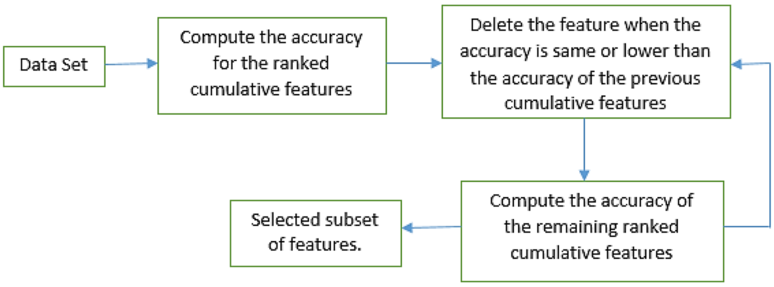

Data Set
Compute the accuracy for the ranked cumulative features
Delete the feature when the accuracy is same or lower than the accuracy of the previous cumulative features
Compute the accuracy of the remaining ranked cumulative features
Selected subset of features.

Figure 2. Flowchart of searching a smaller selected feature

3. EXPERIMENTAL RESULT AND DISCUSSION

3.1. The dataset

The dataset was randomly divided into two subsets in a 7:3 ratio: a training set and a test set. The training sample, denoted as D, is characterised by the complete feature set $X = \{x_1, x_2, \dots, x_N\}$ and the N-dimensional category C, which serves as the input for the algorithm. Subsequently, the training set is normalised to the range [-1, 1]. The ranking process is repeated 50 times to calculate the average MI score for each feature. In each iteration, new training and testing datasets are obtained. Before implementing the rMI-SVM algorithm, we will first determine the optimal baseline. Subsequently, the rMI-SVM algorithm will be utilised to reduce the complexity of the dataset and select a concise set of features to construct a predictive model. The rMI-SVM algorithm will be executed five times, and the average accuracy will be determined through these repetitions. The details of the six data are shown in Table 1.

Table 1. Summary of the data set

Data set	No. of feature	No. of sample	Number of classes
Colon	1988	62	Binary
Leukaemia	7218	72	Binary
Parkinson	22	195	Binary
Breast	30	569	Binary
Lymphoma	4026	96	9 classes
Handwriting	649	2000	10 classes

Table 2 presents two baselines: the full feature set and the optimal baseline, along with the number of features needed to achieve this optimal baseline. The optimal baseline surpasses the full feature set and necessitates fewer features. The selected features are ranked based on their MI score, signifying their relevance from high to low. According to the experimental results, the optimal baseline utilising fewer features exhibits superior accuracy compared to the baseline using the full feature set. The features obtained through this method significantly impact feature selection. The feature selection method should not exceed the number of features identified to maintain the same level of accuracy. This establishes new guidelines for determining the number of features permitted during feature selection. Unlike previous studies that relied on a full set of features to set a baseline, this guidance specifies the number of features researchers can utilise when constructing predictive models.

Table 2. The baseline uses full features and the optimal baseline with the number of features

Data set	Baseline (%)	Full features	Optimal baseline (%)	No. of features
Colon	80	1988	87.78	202
Leukaemia	96.19	7128	99.05	38
Parkinson	88.62	22	88.62	22
Breast	91.76	30	95.06	13
Lymphoma	98.33	4026	99.17	460
Handwriting	98.2	649	98.3	434

After identifying the optimal baseline and the pertinent features, the subsequent step involves employing the rMI-SVM algorithm to select a compact set of features for constructing a predictive model. This algorithm effectively filters out redundant features and notably diminishes the dimensionality of the dataset. Table 3 presents the number of features selected by the rMI-SVM algorithm alongside the corresponding percentage of dimensionality reduction. The rMI-SVM algorithm demonstrates remarkable dimensionality reduction, particularly in microarray data, and it also performs effectively in low-dimensional datasets.

Table 3. The number of features selected by the rMI-SVM algorithm

Data set	No of the features selected	Dimension to reduce (%)
Colon	7	99.65
Leukaemia	5	99.93
Parkinson	7	68.18
Breast	11	63.33
Lymphoma	22	99.45
Handwriting	50	92.3

Table 4 shows the average accuracy along with the number of features acquired from various feature selection methods for the six datasets. The results indicate that the rMI-SVM algorithm attains higher accuracy across various datasets compared to other feature selection methods. This illustrates that the rMI-SVM algorithm offers superior performance in selecting relevant features for predictive modelling in contrast to existing methods. The rMI-SVM model demonstrates superior performance compared to the optimal baselines. The predictive model constructed using features selected by rMI-SVM has proven to be relevant and minimally redundant. Consequently, the performance of this prediction model exhibits the highest accuracy using fewer features.

Table 4. Average accuracy (Ave Acc) with the number of features obtained from several feature selection methods for six data sets

Colon	No. of features	Ave acc (%)	Leukaemia	No. of features	Ave acc (%)
Optimal	202	87.78	Optimal	38	99.05
rMI-SVM	7	87.78	rMI-SVM	5	100
Regression	7	82.22	Regression	5	98.1
mRMR	7	81.11	mRMR	5	94.28
JMI	11	85.4	JMI	5	99
JMIM	12	85.4	JMIM	4	97.25
Parkinson	No. of features	Ave Acc (%)	Breast	No. of features	Ave Acc (%)
Optimal	22	88.97	Optimal	13	95.06
rMI-SVM	7	90.69	rMI-SVM	11	96.59
Regression	7	84.83	Regression	11	93.82
mRMR	7	86.21	mRMR	11	93.29
JMI	3	89.5	JMI	20	95.8
JMIM	8	91	JMIM	5	96.2
Lymphoma	No. of features	Ave Acc (%)	Handwriting	No. of features	Ave Acc (%)
Optimal	460	99.17	Optimal	434	98.3
rMI-SVM	22	99.17	rMI-SVM	50	98.8
Regression	22	55	Regression	50	96.5
JMI	55	91	JMI	33	97
JMIM	59	91	JMIM	39	97.5

Table 5 presents the area under the curve, recall and precision for the six datasets utilising rMI-SVM. The sensitivity of the classifier was evaluated using the ROC curve, and all six datasets yielded a recall value of more than 0.5. Additionally, the classifier employing the selected features demonstrated a high value in the area under the curve. A high AUC value indicates that the predictive model excels in distinguishing between different classes. It underscores the model's reliability in accurately classifying instances, rendering it valuable for its intended application. A high recall value suggests that the predictive model adeptly identifies relevant instances among all actual positive instances. This demonstrates that rMI-SVM can effectively capture the majority of positive cases, exhibiting higher sensitivity, and mitigating the risk of missing critical cases. A high precision value implies that the predictive model is highly accurate in its predictions. This indicates that when the model predicts a positive outcome, it is likely to be correct. Consequently, there is a reduced likelihood of false positives, thereby enhancing the model's reliability in making accurate predictions.

Table 5. The AUC, recall, and precision for six data sets using rMI-SVM

Data set	AUC	Recall	Precision
Colon	0.94	0.69	0.92
Leukaemia	1	1	0.94
Parkinson	0.92	1	0.85
Breast	0.99	0.98	0.97
Lymphoma	0.97	0.68	1
Handwriting	1	0.98	0.99

Research results indicate that the rMI-SVM algorithm achieves superior performance in predictive modelling by selecting a compact feature subset, thereby surpassing traditional methods. This aligns with the research objective of assessing the efficacy of the rMI-SVM algorithm in enhancing the accuracy of predictive models through feature selection. This suggests that employing fewer relevant features can result in improved predictive model performance, challenging the notion of incorporating all features into the model. The high recall, precision, and AUC values obtained using the selected features demonstrate the reliability and effectiveness of the prediction model established using the rMI-SVM algorithm.

4. CONCLUSION

The rMI-SVM algorithm in feature selection represents a significant advancement in predictive modelling. This algorithm adeptly handles high-dimensional data by eliminating redundant features, thereby mitigating overfitting and enhancing the overall performance of the model. Moreover, exploring feature relationships using MI networks yields valuable insights. This breakthrough addresses the issue where previous predictive models often exhibited fluctuating performance when more features were added to the model. The rMI-SVM algorithm ensures that the performance of the predictive model improves as more features are incorporated.

Additionally, the cumulative ranking of features establishes an optimal baseline through MI scores. This ranking process suggests that higher scores indicate more information about the feature. The results show that the optimal baseline outperforms existing baselines that include all features. This improvement is attributed to the inclusion of all features introducing irrelevant redundancy and noise into the model. This optimal baseline will serve as a benchmark for the predictive model, as past studies have not provided clear guidelines on setting benchmarks.

The rMI-SVM algorithm offers a promising method for processing high-dimensional data and improving the performance of predictive models. The ranked mutual information scores of features establish the optimal baseline, providing a benchmark for model evaluation and comparison. This optimal baseline outperforms existing baselines by eliminating irrelevant redundancy and noise, thereby offering clearer guidance on benchmark performance in future studies. Overall, the findings demonstrate that the rMI-SVM algorithm and its related methods have the potential to enhance predictive modelling practices in various fields where high-dimensional data prevail.

Further investigation into the relationship between the selected features can be conducted by constructing a network model using MI. The ordering of features has consistently posed a challenge in feature selection. Therefore, additional research on how to select features when encountering a tie condition is crucial, as the contribution of these features to the label class may vary. Additionally, the current algorithm can be enhanced by exploring more relevant features through the integration of other selection methods.

ACKNOWLEDGEMENTS

We gratefully acknowledge the partial financial support provided by Universiti Tunku Abdul Rahman. Their support has been invaluable in enabling us to conduct this research.

REFERENCES

[1] R. Toor and I. Chana, “DIDACE: Literature Mining and Exploration of Disease-Diet Associations,” *Journal of Information Science and Engineering*, vol. 38, no. 1, pp. 207–221, 2022, doi: 10.6688/JISE.202201_38(1).0011.

[2] L. Hu, W. Gao, K. Zhao, P. Zhang, and F. Wang, “Feature selection considering two types of feature relevancy and feature interdependency,” *Expert Systems with Applications*, vol. 93, pp. 423–434, Mar. 2018, doi: 10.1016/j.eswa.2017.10.016.

[3] N. X. Vinh and J. Bailey, “Comments on supervised feature selection by clustering using conditional mutual information-based distances,” *Pattern Recognition*, vol. 46, no. 4, pp. 1220–1225, Apr. 2013, doi: 10.1016/j.patcog.2012.11.001.

[4] J. M. Alostad, “Improved probabilistic distance based locality preserving projections method to reduce dimensionality in large datasets,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 593–601, Feb. 2019, doi: 10.11591/ijece.v9i1.pp593-601.

[5] A. Ashraf, A. Sophian, A. A. Shafie, T. S. Gunawan, and N. N. Ismail, “Machine learning-based pavement crack detection, classification, and characterization: a review,” *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3601–3619, Dec. 2023, doi: 10.11591/eei.v12i6.5345.

[6] F. Li, D. Miao, and W. Pedrycz, “Granular multi-label feature selection based on mutual information,” *Pattern Recognition*, vol. 67, pp. 410–423, Jul. 2017, doi: 10.1016/j.patcog.2017.02.025.

[7] G. Herman, B. Zhang, Y. Wang, G. Ye, and F. Chen, “Mutual information-based method for selecting informative feature sets,” *Pattern Recognition*, vol. 46, no. 12, pp. 3315–3327, Dec. 2013, doi: 10.1016/j.patcog.2013.04.021.

[8] F. Fleuret, “Fast binary feature selection with conditional mutual information,” *Journal of Machine Learning Research*, vol. 5, no. 12, pp. 1531–1555, 2004.

[9] X. V. Nguyen, J. Chan, S. Romano, and J. Bailey, “Effective global approaches for mutual information based feature selection,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA: ACM, Aug. 2014, pp. 512–521, doi: 10.1145/2623330.2623611.

[10] J. Novovičová, P. Somol, M. Haindl, and P. Pudil, “Conditional Mutual Information Based Feature Selection for Classification Task,” *Progress in Pattern Recognition, Image Analysis and Applications: 12th Iberoamericann Congress on Pattern Recognition*, vol. 4756, pp. pp. 417–426, 2007, doi: 10.1007/978-3-540-76725-1_44.

[11] H. Cheng, Z. Qin, C. Feng, Y. Wang, and F. Li, “Conditional Mutual Information-Based Feature Selection Analyzing for Synergy and Redundancy,” *ETRI Journal*, vol. 33, no. 2, pp. 210–218, Apr. 2011, doi: 10.4218/etrij.11.0110.0237.

[12] H. Peng and Y. Fan, “Feature selection by optimizing a lower bound of conditional mutual information,” *Information Sciences*, vol. 418–419, pp. 652–667, Dec. 2017, doi: 10.1016/j.ins.2017.08.036.

[13] D. D. Lewis, “Feature Selection and Feature Extraction for Text Categorization,” in *Proceedings of a Workshop Held at Harriman*, New York, Feb. 1992.

[14] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, Jul. 1994, doi: 10.1109/72.298224.

[15] N. Kwak and Chong-Ho Choi, “Input feature selection for classification problems,” *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002, doi: 10.1109/72.977291.

[16] H. H. Yang and J. Moody, “Data visualization and feature selection: new algorithms for nongaussian data,” in *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS’99)*, MIT Press, 1999, pp. 687–693.

[17] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[18] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized Mutual Information Feature Selection,” *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, Feb. 2009, doi: 10.1109/TNN.2008.2005601.

[19] T. Hlaing, “Feature Selection and Fuzzy Decision Tree for Network Intrusion Detection,” *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 1, no. 2, pp. 109–118, Sep. 2012, doi: 10.11591/ij-ict.v1i2.591.

[20] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, “MIFS-ND: A mutual information-based feature selection method,” *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371–6385, Oct. 2014, doi: 10.1016/j.eswa.2014.04.019.

[21] M. Bennasar, Y. Hicks, and R. Setchi, “Feature selection using Joint Mutual Information Maximisation,” *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, Dec. 2015, doi: 10.1016/j.eswa.2015.07.007.

[22] Y.-D. Lan, “A Hybrid Feature Selection Based on Mutual Information and Genetic Algorithm,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 214–225, Jul. 2017, doi: 10.11591/ijeecs.v7.i1.pp214-225.

[23] J. Wang, J.-M. Wei, Z. Yang, and S.-Q. Wang, “Feature Selection by Maximizing Independent Classification Information,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 4, pp. 828–841, Apr. 2017, doi: 10.1109/TKDE.2017.2650906.

[24] W. Gao, L. Hu, and P. Zhang, “Class-specific mutual information variation for feature selection,” *Pattern Recognition*, vol. 79, pp. 328–339, Jul. 2018, doi: 10.1016/j.patcog.2018.02.020.

[25] B. N. Kumar, M. S. V S. B. Raju, and B. V. Vardhan, “A novel approach for selective feature mechanism for two-phase intrusion detection system,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 101–112, Apr. 2019, doi: 10.11591/ijeecs.v14.i1.pp101-112.

[26] A. Bir-Jmel, S. M. Douiri, and S. Elbernoussi, “Gene Selection via a New Hybrid Ant Colony Optimization Algorithm for Cancer Classification in High-Dimensional Data,” *Computational and Mathematical Methods in Medicine*, vol. 2019, pp. 1–20, Oct. 2019, doi: 10.1155/2019/7828590.

[27] H. Zhou, X. Wang, and Y. Zhang, “Feature selection based on weighted conditional mutual information,” *Applied Computing and Informatics*, vol. 20, no. 1/2, pp. 55–68, Jan. 2024, doi: 10.1016/j.aci.2019.12.003.

[28] S. Alelyani, “Stable bagging feature selection on medical data,” *Journal of Big Data*, vol. 8, no. 1, p. 11, Dec. 2021, doi: 10.1186/s40537-020-00385-8.

[29] F. Macedo, R. Valadas, E. Carrasquinha, M. R. Oliveira, and A. Pacheco, “Feature selection using Decomposed Mutual Information Maximization,” *Neurocomputing*, vol. 513, pp. 215–232, Nov. 2022, doi: 10.1016/j.neucom.2022.09.101.

[30] C. Pascoal, M. R. Oliveira, A. Pacheco, and R. Valadas, “Theoretical evaluation of feature selection methods based on mutual information,” *Neurocomputing*, vol. 226, pp. 168–181, Feb. 2017, doi: 10.1016/j.neucom.2016.11.047.

[31] I. Chlioui, A. Idri, I. Abnane, and M. Ezzat, “Ensemble Case based Reasoning Imputation in Breast Cancer Classification,” *Journal of Information Science and Engineering*, vol. 37, no. 5, pp. 1039–1051, 2021, doi: 10.6688/JISE.202109_37(5).0004.

[32] R. Ahuja and S. C. Sharma, “Exploiting Machine Learning and Feature Selection Algorithms to Predict Instructor Performance in Higher Education,” *Journal of Information Science and Engineering*, vol. 37, no. 5, pp. 993–1009, 2021, doi: 10.6688/JISE.202109_37(5).0001.

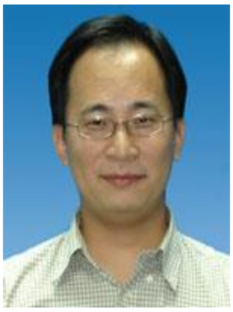
[33] T. A. Assegie, V. Elanangai, J. S. Paulraj, M. Velmurugan, and D. F. Devesan, “Evaluation of feature scaling for improving the performance of supervised learning methods,” *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1833–1838, Jun. 2023, doi: 10.11591/eei.v12i3.5170.

[34] F. Y. Chin, K. H. Lem, and K. M. Wong, “Improving handwritten digit recognition using hybrid feature selection algorithm,” *Applied Computing and Informatics*, Jul. 2022, doi: 10.1108/ACI-02-2022-0054.

[35] N. Sureja, B. Chawda, and A. Vasant, “A novel salp swarm clustering algorithm for prediction of the heart diseases,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 265–272, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp265-272.

BIOGRAPHIES OF AUTHORS

Fung Yuen Chin received the B.S. (Honours) and M.S. degrees in Mathematics from Universiti Kebangsaan Malaysia, Malaysia. She is currently pursuing a Ph.D. degree in the Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman, Malaysia. Her research interests include bioinformatics, data analysis, and machine learning. She can be contacted at email: chinfy@utar.edu.my.



Yong Kheng Goh received his Ph.D. degree in Mathematics from the University of London, England. He is currently an Associate Professor at Universiti Tunku Abdul Rahman, Malaysia. His research interests include bioinformatics, computational finance, computational physics, data analysis and visualisation, and statistical mechanics. He can be contacted at email: gohyk@utar.edu.my.