A Novel Nonparametric Feature Selection Approach Based on Mutual Information Transfer Network

Kunmei Li and Nasser Fard *

College of Engineering, Northeastern University, Boston, MA 02115, USA
* Correspondence: n.fard@northeastern.edu

Abstract: The filter feature selection algorithm is habitually used as an effective way to reduce the computational cost of data analysis by selecting and implementing only a subset of original features into the study. Mutual information (MI) is a popular measurement adopted to quantify the dependence among features. MI-based greedy forward methods (MIGFMs) have been widely applied to escape from computational complexity and exhaustion of high-dimensional data. However, most MIGFMs are parametric methods that necessitate proper preset parameters and stopping criteria. Improper parameters may lead to ignorance of better results. This paper proposes a novel nonparametric feature selection method based on mutual information and mixed-integer linear programming (MILP). By forming a mutual information network, we transform the feature selection problem into a maximum flow problem, which can be solved with the Gurobi solver in a reasonable time. The proposed method attempts to prevent negligence on obtaining a superior feature subset while keeping the computational cost in an affordable range. Analytical comparison of the proposed method with six feature selection methods reveals significantly better results compared to MIGFMs, considering classification accuracy.

Keywords: feature selection; mutual information; linear programming; maxflow; P-median problem

1. Introduction

The ‘Curse of Dimensionality’ is first mentioned by Bellman to describe various challenges that occur during data analysis due to increasing dimensionality [1]. Most of those challenges refer to computational complexity, especially the exhaustion of computational storage and time. Multiple types of dimensionality reduction (DR) approaches, such as feature selection (FRS) and feature extraction, have been proposed to overcome computational fatigue. FRS focuses on selecting a proper subset of original features from the data to maximize or minimize predetermined objectives to meet the requirements of various users. For instance, the objective can be to minimize the Akaike Information Criterion (AIC)/the Bayesian Information Criterion (BIC) or to maximize the classification accuracy [2]. FRS has also been extensively applied because of its capability to promote model interpretability. By implementing various classifiers along with machine learning approaches, FRS approaches are grouped into three types: filter, wrapper, and embedded methods. Filter methods refer to the FRS approaches without considering the classifier’s performance, except embedded methods have an intrinsic model-building metric that wrapper methods do not have during learning process. Filter methods are more computationally efficient than wrapper and embedded methods, since they do not take classifiers into account [4]. The filter methods are classified into several types based on the measures of relationship between predictors and response (or target), including distance-based, probability-based, MI-based, consistency, and correlation measures [5]. Figure 1 presents a summary of some filter measures.
Compared to other types of filter measures, MI-based measures have a few superiorities. To begin with, MI-based measures can assess the general dependence among random variables without making any assumptions in advance. Furthermore, MI-based measures are slightly affected by the monotone transformation and classifier selection [6]. These advantages allow MI-based measures for broad application in the analysis of various types of problems, including computer-aided diagnosis [7], cyber intrusion detection [8], heart failure recognition [9], and software cost estimation [10].

MI-based greedy forward methods (MIGFMs) are widely used for high-dimensional classification problems to avoid computational complexity and exhaustion. It follows the heuristic of making a greedy choice at each step of searching process based on a given optimal criterion to ensure the optimality of feature selection, which significantly reduces the searching process. For instance, to select \( m \) features out of \( n \) possible features, the total number of all the candidate subsets is \( \binom{n}{m} = \frac{n!}{m!(n-m)!} \), which could be enormous for a relatively large \( n \). In the greedy forward methods, features are selected incrementally, and only one feature is selected from the candidate subsets consisting of unselected features at each searching step based on a specified optimal criterion. Consequently, the total number of searches for determining \( m \) out of \( n \) features is:

\[
 n + (n - 1) + (n - 2) + \cdots + (n - m + 1) = \frac{(2n - m + 1)m}{2} \tag{1}
\]

This is significantly smaller compared to \( \frac{n!}{m!(n-m)!} \). More details about MIGFMs are given in Section 2.

Even though greedy forward methods deduct the total number of steps, they may simultaneously lead to negligence of a superior subset. Greedy algorithms typically lack the foresight to select inferior options during an iteration to facilitate subsequent results from the algorithm [11]. Moreover, they encounter challenges in assessing the impact of interactions among features [12]. To overcome the aforementioned limitations and to balance the efficiency and accuracy of MI-based FRS methods, we propose a new nonparametric approach based on MILP in Section 3. We will also compare the performance of the proposed method with several popular MI-based heuristic FRS methods and provide a discussion in Sections 4 and 5.

2. Information Theory and MI-Based Methods

This section briefly introduces the basic concepts of information theory and then presents a few MIGFMs with various optimal criteria.

2.1. Information Theory

Information theory is part of probability theory that was initially proposed to measure information transmitted across a noisy channel [13]. It has been broadly applied to diverse fields, such as computer science, social science, economics, and physics [14]. Two
parameters, entropy and MI, are used to measure the amount of information. For a random
discrete variable \( X = \{ x_1, x_2, x_3, \ldots, x_n \} \), the average information about \( X \) is represented
by Shannon entropy \( H(X) \), where \( p_i \) denotes the probability of \( x_i \).

\[
H(X) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{2}
\]

If there is another random discrete variable \( Y \), the amount of information needed to
describe \( X \) given a known \( Y \) is denoted by conditional entropy \( H(X|Y) \):

\[
H(X|Y) = H(X,Y) - H(Y) \tag{3}
\]

where \( H(X,Y) \) is the joint entropy measuring the entropy associated with these two variables:

\[
H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) \tag{4}
\]

The mutual information of \( X \) and \( Y \) measures the information about \( X \) (or \( Y \)) that can
be obtained through \( Y \) (or \( X \)) and is defined as:

\[
I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{5}
\]

These definitions can be expanded to multivariate cases, which are primarily ex-
pressed by the following two equations that involve conditional entropy and conditional
mutual information.

\[
H(X_1, X_2, \ldots, X_{n-1}, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{6}
\]

\[
I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1)
= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n|Y) \tag{7}
\]

Figure 2 presents a Venn diagram that distinctly describes the relationships between
entropy and mutual information of three variables [15]. The above concepts are also gen-
eralized for the case of continuous variables by replacing partial equations with differential
equations. Furthermore, various approximation approaches have been proposed to mea-
sure the mutual information between discrete and continuous variables, such as Jensen
Shannon divergence (JSD) [16] and kernel density approximation (KDA) [17].

Figure 2. Three-variable MI venn diagram.
2.2. MI-Based FRS Methods

As mentioned above, FRS seeks to obtain a subset of features that provides the optimum performance for a specific objective. For the MI-based methods, the objective could be to maximize the mutual information between a selected predictor and a class variable in a classification task. For example, suppose that there is a data set with \( n \) features denoted by \( \{f_1, f_2, f_3, \ldots, f_{n-1}, f_n\} \) and one class variable denoted by \( C \). A selected subset consisting of \( m \) features, \( S' \), should satisfy:

\[
S' = \arg \max_{S} \{I(C; S)\}
\]  

(8)

where \( S \) denotes all the possible combinations of \( m \) features. The total number of the \( S \) is \( \binom{n}{m} \), which can be a huge number for a large \( n \), as stated in Section 1. To avoid computational complexity and exhaustion caused by a large \( n \), various greedy forward methods focusing on the systematic selection of features have been proposed. Including MI variable selector under uniform information distribution (MIFS-U) [18], information gain redundancy and min-redundancy (MRMD) [21], max-relevance and max-independence (MIFU) [19], and spectral relaxation conditional MI (SPEC-CMI) [24]. These MIGFMFs are summarized in Algorithm 1.

**Algorithm 1 Summarized Algorithm.**

1. Initialization: \( SEL = \emptyset, SEL = \{f_1, f_2, f_3, \ldots, f_{n-1}, f_n\} \), \( m \) is the number of features to be selected, and \( 0 < m \leq n \)
2. while \( |SEL| < m \) do:
   3. \( f \leftarrow \arg \max_{f \in SEL} \{I(C; f) - \beta \sum_{f_i \in SEL} g(f, f_i, C)\} \)
   4. \( SEL \leftarrow SEL + \{f\} \)
   5. \( SEL \leftarrow SEL - \{f\} \)
3. end while
4. return \( SEL \)

Where \( \beta \) is a critical scale parameter, \( g(f, f_i, C) \) is a manually set function, \( SEL \) is the set of selected features, and the \( |SEL| \) denotes the set cardinality of \( SEL \). In each search step, an unselected feature that maximizes the locally optimal function \( I(C; f) - \beta \sum_{f_i \in SEL} g(f, f_i, C) \) is picked.

Multiple kinds of MI-based FRS methods can be transformed into this algorithm with different \( \beta \) and \( g(f, f_i, C) \) as provided in Table 1.

**Table 1. MI-based FRS methods in summarized algorithm format.**

| MI-Based FRS Methods | \( \beta \) | \( g(f, f_i, C) \) |
|----------------------|--------|-----------------|
| MIFS-U               | \( \beta \) | \( I(f; C) - I(f; f_i) \) |
| IGFS                 | \( -\frac{1}{|SEL|} \) | \( I(f; C, f_i) \) |
| CIFE                 | \( 1 \) | \( I(f; C, f) \) |
| MRMD                 | \( \frac{1}{|SEL|} \) | \( I(f; f_i) - I(f; C|f_i) \) |
| MRI                  | \( 1 \) | \( 2I(f; f_i; C) - I(f; C) - I(f_i; C) \) |
| mRMR                 | \( \frac{1}{|SEL|} \) | \( I(f; f_i) \) |
| SPEC-CMI             | \( -1 \) | \( I(f; C|f_i) \) |

In addition to the methods indicated in Table 1, the methods with nonlinear combinations of mutual information, such as CMIM [25], ICAP [26], NMIFS [27], and DISR [28] are also adopted in various aspects [29].

When employing greedy forward methods, some concerns need to be specified, such as selecting appropriate parameters such as \( \beta \) and the stopping criterion of the greedy
procedure [30]. The improper specification of parameters and stopping criteria may lead
to a negligence of obtaining a better subset [31]. To circumvent the disadvantages of
the parametric methods, we present a novel method which converts the FRS problem
into a mixed-linear optimization model. P-median method is used for the selection of
facility location according to an objective function, similarly MILP approaches could be
introduced for feature selection. The following section introduces some basic definitions of
the P-median model, as it relates to our proposed method.

3. P-Median Problem and Proposed Algorithm

In this section, we first present a brief description of the P-median problem before
presenting the proposed algorithm.

3.1. P-Median Problem

The P-median problem is an imperative type of discrete location problem that focuses
on selecting optimal locations for \( P \) facilities from a set of potential locations to minimize
the total delivery cost or distance [32]. The selection of suitable locations for facilities
is one of the most critical tasks in almost every sector. For instance, the transportation
authority must determine the locations of bus and subway stations; the government needs
to locate public services, including hospitals and schools; the delivery services company is
required to locate the mail drop boxes. Such strategical problems are known as discrete
location problems or facility location problems. P-median problem is an NP-hard problem
that takes polynomial time to solve [33]. As the number of variables and constraints
involved in this problem grows, the complexity of the P-median problem increases, and
when the number of arcs in graphs built from the P-median problem exceeds 360,000,
commercial MILP coding becomes difficult [34]. Therefore, instead of seeking a globally
optimal result, we set termination criteria such as elapsed time, relative gap tolerance, and
absolute gap tolerance to identify a relatively optimum solution that can be recognized
within a reasonable amount of time while maintaining acceptable accuracy. Typically, the
P-median problem is formulated as following MILP:

\[
\begin{align*}
\text{minimize} & \quad \sum_{i \in I} \sum_{j \in J} b_i d_{ij} y_{ij} \\
\text{subject to} & \quad \sum_{j \in J} x_j = P \\
& \quad \sum_{j \in J} y_{ij} = 1, \quad \forall i \\
& \quad y_{ij} - x_j \leq 0, \quad \forall i, j \\
& \quad y_{ij} \in \{0, 1\}, \quad \forall i, j \\
& \quad x_j \in \{0, 1\}, \quad \forall j
\end{align*}
\]  

(9)

where:

- \( b_i \): demand of \( i \)th customer
- \( d_{ij} \): distance between \( i \)th customer and \( j \)th candidate facility
- \( P \): number of facilities to be located
- \( x_j \): binary variable, =1 if locate facility at \( j \)th candidate, =0 if not
- \( y_{ij} \): binary variable, =1 if \( i \)th customer is served by \( j \)th facility, =0 if not

The optimal solution for \( x_j \) and \( y_{ij} \) indicates the locations of \( P \) facilities and the clients’
allocation that provides the minimum total distance between the clients and the facilities.
As mentioned above, the goal of the P-median problem is to choose \( P \) locations from all
candidates that optimize an objective function. It is analogous to the FRS problem, in which
the goal is to locate a subset of features that maximize the mutual information between
the selected features and the response variable. They both attempt to choose entities from
a pool of all potential entities to achieve a specific goal. Because of this similarity, we
apply MILP to the FRS problem and propose a new approach to select features by solving
MI-transfer-network-based integer linear programming (MITN-ILP). Section 3.2 provides
more details about this method.
3.2. The Proposed Algorithm (MITN-ILP)

The P-median problem and its similarities to the FRS problem were presented in the previous section. It has clearly demonstrated the feasibility of handling the feature selection problem by reframing MILP models for the P-median problem. Next, we give more details about the proposed MITN-ILP method, which includes two main steps listed below:

- First, create the MI transfer network from original data.
- Then, build the MILP model and solve it.

The solution of the MILP model designates the selection of features. In subsequent subsections, we present each step explicitly.

3.2.1. Create the MI Transfer Network

Before developing the MILP model for the FRS problem, we need to transform the information from the FRS problem into sets of variables and constraints for the MILP model, by constructing an MI transfer network. Figure 3 depicts the ubiquitous structure of MI transfer network.

![Figure 3: Structure of MI-transfer-network.](image)

The network within the square box is a fully connected network consisting of all candidate features, where the $i$th feature is represented by a node $f_i$ with a supply $b_i = I(f_i; C)$, and the capacity of the arc between $f_i$ and $f_j$ is the MI between the $i$th feature and the $j$th feature. Furthermore, there is also a designated target node $T$ for gathering all the transferred information, as well as a dummy node $D$ for ensuring the network’s flow balance. The arc capacity from any feature node to node $T$ or node $D$ is infinite. After calculating the paired MI of the given data set, we can create the flow network as illustrated in Figure 3 and convert the FRS problem into an MILP problem that maximizes the flow into node $T$. In the MI-transfer-network, the communicated information between each paired predictors is quantitated by the MI between them, and the information of the response variable contained in each predictor is quantitated by the MI between it and the response variable. Since we assume that the communicated information between two predictors includes the joint information of the response variable contained in those two predictors, only one node of the fully connected network is permitted to transmit the MI into node $T$ to restrict the redundancy of response information. Section 3.2.2 presents detailed MILP formulations for this problem.

3.2.2. Develop the MILP Model

For the FRS problem, our aim is to select a set of features that maximize the MI between the selected features and the target variable. To clarify the formulation, we add some definitions before constructing the MILP model. First, an MI-Matrix called $U$ is defined to represent the mutual information between each pair of nodes, as shown in Equation (10). $U_{f_i, r+1}$ and $U_{f_i, r+2}$ are equal to zero, while $U_{f_i, r+1}$ and $U_{f_i, r+2}$ are equal to $E$ (an immense positive value that times zero is still zero to ensure the calculability) since nodes $D$ and $T$ only have incoming flows.
where:

\[ f_1 \quad f_2 \quad \ldots \quad f_n \quad D \quad T \]
\[ 0 \quad I(f_1;f_2) \quad \ldots \quad I(f_1;f_n) \quad E \quad E \quad f_1 \]
\[ I(f_2;f_1) \quad 0 \quad \ldots \quad I(f_2;f_n) \quad E \quad E \quad f_2 \]
\[ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \]
\[ I(f_n;f_1) \quad I(f_n;f_2) \quad \ldots \quad 0 \quad E \quad E \quad f_n \]
\[ 0 \quad 0 \quad \ldots \quad 0 \quad 0 \quad 0 \quad D \]
\[ 0 \quad 0 \quad \ldots \quad 0 \quad 0 \quad 0 \quad T \]

Meanwhile, all mutual information between each feature and the target variable \( C \) is denoted by the supply vector \( b \):

\[
\begin{bmatrix}
  b_1 = I(f_1; C) \\
  b_2 = I(f_2; C) \\
  \vdots \\
  b_n = I(f_n; C)
\end{bmatrix}
\]  

(11)

The target variables for MILP are given by \( x, y, \) and \( v \), while both \( y \) and \( v \) are binary vectors.

\[
\begin{bmatrix}
  x_{1,1} & x_{1,2} & \ldots & x_{1,n} & x_{1,t} \\
  x_{2,1} & x_{2,2} & \ldots & x_{2,n} & x_{2,t} \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  x_{n,1} & x_{n,2} & \ldots & x_{n,n} & x_{n,t} \\
  x_{d,1} & x_{d,2} & \ldots & x_{d,n} & x_{d,t} \\
  x_{l,1} & x_{l,2} & \ldots & x_{l,n} & x_{l,t}
\end{bmatrix}, \quad
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n \\
  y_{n+1} \\
  \vdots \\
  y_{n+t}
\end{bmatrix}, \quad
\begin{bmatrix}
  v_1 \\
  v_2 \\
  \vdots \\
  v_n \\
  v_{n+1} \\
  \vdots \\
  v_{n+t}
\end{bmatrix}
\]  

(12)

Let \( x_{i,j} \) denote the \( i \)th row of a matrix \( x \), and \( x_{s,j} \) denote the \( j \)th column of a matrix \( x \), where \( i, j \in \{1, 2, 3, \ldots, n, t, d\} \). Hence, the MILP model identifying \( P \) features among all potential candidates that transfer maximum MI in the network is given in Section 3.2.1 is:

\[
\begin{align*}
\text{maximize} & \quad \sum x_{s,n+2} \\
\text{subject to} & \quad e^T y = P \\
& \quad e^T v = 1 \\
& \quad \sum x_{i,s} = \sum x_{s,j} = b_i \quad \forall i \\
& \quad x_{s,n+2} \leq U_{s,n+2} \circ v \\
& \quad x_{i,n+2} \leq k \quad \forall i \\
& \quad x_{s,n+2} \leq U_{s,n+2}
\end{align*}
\]  

(13)

where:

- \( e \): a column vector = \((1, 1, \ldots, 1)^T\), and \( e^T \) is its transpose
- \( U \): MI-Matrix, where \( U \) the arc capacity between node \( i \) and node \( j \)
- \( b \): supply vector, where \( b_i \) = the amount of information at node \( i \)
- \( x \): flow matrix, where \( x_{i,j} \) = the amount of information shifted from node \( i \) to node \( j \)
- \( y \): binary vector, where \( y_i = 1 \) if \( f_i \) is selected and \( y_i = 0 \) if not
- \( v \): binary vector, where \( v_i = 1 \) if the arc \((f_i, T)\) is activated and \( v_i = 0 \) if not
- \( P \): total number of features are selected
- \( k \): the precalculated parameter for a better bound for the solution
- \( \circ \): the Hadamard product operator

It should be noted that in the proposed method, only one node can transmit the MI into node \( T \), as indicated by \( e^T v \) in the above formulation. Without this constraint, the selected nodes will be the top \( P \) nodes by their supply, which does not meet our expectations. Furthermore, an upper bound \( k \) is also applied to reduce the calculation time by shrinking the feasible area. To obtain an appropriate \( k \), we separate all the nodes into two types based
on their value of \( v \). All nodes with \( v_i = 0 \) form the first layer, while the nodes with \( v_i = 1 \) constitute the second layer as displayed in Figure 4.

![End part of MI-transfer-network.](image)

Figure 4. End part of MI-transfer-network.

If the \( i \)th node is selected from the second layer, then only the \((P-1)\) nodes can be selected from \( \mathcal{UNS} \), where \( \mathcal{UNS} \) is a set composed of all the unselected features, to form the first layer. The maximum transferable MI from the first layer into the second layer should be less than the maximum total supplies in the first layer, which is the sum of the top \((P-1)\) values in \( \{b_j, \forall j \in \mathcal{UNS}\} \). Moreover, the maximum MI should also be limited by the maximum total capacities between the first layer and the second layer, which is the sum of the top \((P-1)\) values in \( \{u_{ij}, \forall j \in \mathcal{UNS}\} \). Consequently, we can obtain one upper limit \( k \) (the right side of the inequality given below) for the total maximum flow:

\[
x_{i,n+2} \leq \min \{b_i + \sum_{j=1}^{P-1} \max_{j \in \mathcal{UNS}} \{u_{ij}, \forall j \in \mathcal{UNS}\}, b_i + \sum_{j=1}^{P-1} \max_{j \in \mathcal{UNS}} \{b_j, \forall j \in \mathcal{UNS}\}\}
\]

To compute optimal solutions, various solvers may be utilized. We used the Gurobi solver in this study. As stated in Section 3.1, it is not easy to find the global optimum efficiently, especially for the high-dimensional data. Therefore, the relative MIP optimality gap (MIPGap) is adopted as the termination criterion for the Gurobi solver [35]. MIPGap refers to at least the gap value that Gurobi has to reach before declaring optimality, and a larger MIPGap usually reduces the computational time. Optimal values of \( x, y, \) and \( v \) under this criterion are obtained from the solver. The optimal value of \( y \) is the most essential result, since it indicates the selection of the features.

4. Experiments and Results

To verify the performance of our proposed algorithm, we applied the algorithm in conjunction with six other FRS methods: mRMR, CMIM, CIFE, Relief [36], CFS [37], and Fisher’s Score (FS), to a set of benchmark data sets, then compared their classification accuracy. Unlike the other six methods, MITN-ILP requires a mathematical optimization solver. We applied AMPL [38] as the algebraic modeling language cooperating with Gurobi, and increased MIPGap from its default value 0.01% to 5% which reduced the computational time by 88.3%. The summary of the data sets is given in Table 2, which contains the number of observations, features, and classes. Data sets comprise of both binary data sets and multiclass data sets from various aspects. The data sets are from the UCI machine learning repository [39], except Mfeat and Volkert that are from OpenML [40].

To build classification models, all seven FRS methods, including our proposed method, have been used to select the features with feature numbers \( P = \{5, 10, 15, \ldots, 50\} \). Afterward, there are four kinds of model: K-Nearest-Neighbor (KNN) [41], Naïve Bayes (NB) [42], Support Vector Machine (SVM) [43], and Linear Discriminant Analysis (LDA) [44], are implemented to exclude the influences of classifiers. Finally, the overall performance of
each FRS method is measured by the average classification accuracy obtained from these
four models.

Table 2. Summary of data sets.

| Data Sets    | Number of Observations | Number of Features | Number of Classes |
|--------------|------------------------|-------------------|------------------|
| APS          | 60,000                 | 170               | 2                |
| Arrhythmia   | 279                    | 179               | 16               |
| Cortex       | 1080                   | 81                | 8                |
| Gas          | 13,910                 | 128               | 6                |
| LSVT         | 126                    | 310               | 2                |
| Madelon      | 4400                   | 500               | 2                |
| Mfeat        | 2000                   | 217               | 10               |
| Musk         | 6598                   | 168               | 2                |
| PdSpeech     | 756                    | 754               | 2                |
| Spectrometer | 531                    | 102               | 48               |
| Urban        | 675                    | 148               | 9                |
| Volkert      | 58,310                 | 181               | 10               |

Experimental results are presented in Figures 5 and 6. The classification accuracy of
FRS methods with each classifier has been ranked in descending order. The rankings of the
proposed approaches for each data and classifier are shown in Figure 5. Moreover, Figure 6
provides 12 subfigures reporting the average classification accuracy from all classifiers
of 12 data sets. Different colors denote different FRS methods, and the solid black line
symbolizes our proposed method. The proposed method outperforms all other methods in
terms of mean accuracy across almost all other data sets, and it ranks among the top three
overall when the number of selected features is greater than 20.

Figure 5. Classification accuracy rank of MITN-ILP for each data and classifier.
Figure 6. Cont.
Figure 6. Mean classification accuracy for all our data.

Despite the fact that our proposed method shows its superiority, there is some slight variation across data sets. For instance, MITN-ILP’s superiority is significant when $p \leq 25$ for the Musk and LVST, but it is not for the Gas and Mfeat. To figure out the factors that might account for the discrepancy of the proposed method, analysis of variance (ANOVA) is used to detect significant factors from: the number of selected features (FeatureLevel), the number of classes (NumClass), and $OF-Ratio = \frac{|\text{observations}|}{|\text{features}|}$. $OF-Ratio$ is arisen here instead of directly using the number of features and the number of observations, because it standardized the number of observations and the number of features ranging from hundreds to billions in real-world problems. Before further analysis, we separate $OF-Ratio$ into three levels: $(0, 1]$ for ‘small’, $(1, 10]$ for ‘medium’, and $(10, \infty)$ for ‘large’. The number of classes is separated into ‘binary’ and ‘multiclass’, while the number of selected features is categorized into $[5, 25]$ (‘low’) and $[30, 50]$ (‘high’). The interaction plots from ANOVA are given in Figure 7.

![Interaction plots for mean accuracy](image)

Figure 7. Interaction plots for mean accuracy.

The $p$-value of interaction effect between FeatureLevel and NumClass is 0.029, while that between FeatureLevel and $OF-Ratio$ is 0.046. Therefore, we can conclude the significance of these interaction effects. From the interaction plots, the MITN-ILP performs better when FeatureLevel = ‘low’ for binary classification and FeatureLevel = ‘high’ for multiclass. Furthermore, MITN-ILP provides higher accuracy for the data with an $OF-Ratio > 10$ when selecting more than 25 features. For the data set with an $OF-Ratio \leq 1$, MITN-ILP performs better when choosing a smaller size of features.

5. Discussion and Conclusions

In this study, we proposed a new approach that attempts to overcome the shortcomings of MIGFMs, such as the ignorance of relatively optimum feature subsets, the sensitivity to parameters, and the stopping criterion. By converting it into a maximum flow problem in the MI transfer network, we are able to solve it without considering the necessary parameters for MIGFM. As mentioned in Section 3, MILP can be time-consuming when it involves too many variables and constraints. Therefore, we introduce an upper limit to MI-transfer-network-based linear programming model for computational time reduction.
To analyze the performance of the proposed method, we designed experiments to select features for classification problems with 12 data sets in our proposed method and the other 6 FRS methods. The results achieved from these comparative experiments demonstrated MITN-ILP’s superiority in feature selection concerning classification precision, especially for high-dimensional data that have fewer observations. The insights from this study can assist the healthcare or clinical investigation that are more sensitive to precision and suffers from a lack of patient records, such as electronic health records (EHRs) [45] for newly established healthcare or rural hospitals. Furthermore, unlike greedy forward FRS methods, MITN-ILP helps avoid potential negligence of a better predictor set, since it is unaffected by the preset parameters and the current greedy state.

Despite these encouraging results, questions remain since the complexity of the converted MILP problem increases with the size of the data sets, particularly in terms of the number of features. While our proposed method achieves superior performance efficiently with the aforementioned upper limit of maximum flow and the optimality gap, future work should consider whether a better upper limit can be generalized to make our approach feasible for extremely large data. Furthermore, an acceptable optimality gap that balances efficiency and selection performance is also essential, which allows a broader range of applications of our proposed method.

Author Contributions: Conceptualization, K.L.; Methodology, K.L. and N.F.; Supervision, N.F.; Writing—review & editing, K.L. and N.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Bellman, R. Dynamic programming. *Science* **1966**, *153*, 34–37. [CrossRef]
2. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4.
3. Remeseiro, B.; Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **2019**, *112*, 103375. [CrossRef]
4. Hsu, H.H.; Hsieh, C.W.; Lu, M.D. Hybrid feature selection by combining filters and wrappers. *Expert Syst. Appl.* **2011**, *38*, 8144–8150. [CrossRef]
5. Freeman, C.; Kulić, D.; Basir, O. An evaluation of classifier-specific filter measure performance for feature selection. *Pattern Recognit.* **2015**, *48*, 1812–1826. [CrossRef]
6. Cover, T.M.; Thomas, J.A. Elements of information theory second edition solutions to problems. *Internet Access* **2006**, 19–20.
7. Tourassi, G.D.; Frederick, E.D.; Markey, M.K.; Floyd, C.E., Jr. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Med. Phys.* **2001**, *28*, 2394–2402. [CrossRef]
8. Mohammadi, S.; Desai, V.; Karimipour, H. Multivariate mutual information-based feature selection for cyber intrusion detection. In Proceedings of the 2018 IEEE Electrical Power and Energy Conference (EPEC), Toronto, ON, Canada, 10–11 October 2018; pp. 1–6.
9. Yu, S.N.; Lee, M.Y. Conditional mutual information-based feature selection for congestive heart failure recognition using heart rate variability. *Comput. Methods Programs Biomed.* **2012**, *108*, 299–309. [CrossRef]
10. Liu, Q.; Shi, S.; Zhu, H.; Xiao, J. A mutual information-based hybrid feature selection method for software cost estimation using feature clustering. In Proceedings of the 2014 IEEE 38th Annual Computer Software and Applications Conference, Västerås, Sweden, 21–25 July 2014; pp. 27–32.
11. Simmons, B.I.; Hoeppke, C.; Sutherland, W.J. Beware greedy algorithms. *J. Anim. Ecol.* **2019**, *88*, 804–807. [CrossRef] [PubMed]
12. Cilia, N.D.; De Stefano, C.; Fontanella, F.; di Freca, A.S. A ranking-based feature selection approach for handwritten character recognition. *Pattern Recognit. Lett.* **2019**, *121*, 77–86. [CrossRef]
13. Reza, F.M. *An Introduction to Information Theory*; Courier Corporation: North Chelmsford, MA, USA, 1994.
14. Brillouin, L. *Science and Information Theory*; Courier Corporation: North Chelmsford, MA, USA, 2013.
15. Wang, B.; Shen, Y. A method on calculating high-dimensional mutual information and its application to registration of multiple ultrasound images. *Ultrasonics* 2006, 44, e79–e83. [CrossRef] [PubMed]

16. Ross, B.C. Mutual information between discrete and continuous data sets. *PLoS ONE* 2014, 9, e87357. [CrossRef]

17. Moon, Y.I.; Rajagopalan, B.; Lall, U. Estimation of mutual information using kernel density estimators. *Phys. Rev. E* 1995, 52, 2318. [CrossRef] [PubMed]

18. Kwak, N.; Choi, C.H. Input feature selection for classification problems. *IEEE Trans. Neural Netw.* 2002, 13, 143–159. [CrossRef] [PubMed]

19. El Akadi, A.; El Ouardighi, A.; Aboutajdine, D. A powerful feature selection approach based on mutual information. *Int. J. Comput. Sci. Netw. Secur.* 2008, 8, 116.

20. Lin, D.; Tang, X. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *Proceedings of the European Conference on Computer Vision*, Graz, Austria, 7–13 May 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 68–82.

21. Zou, Q.; Zeng, J.; Cao, L.; Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 2016, 173, 346–354. [CrossRef]

22. Wang, J.; Wei, J.M.; Yang, Z.; Wang, S.Q. Feature selection by maximizing independent classification information. *IEEE Trans. Knowl. Data Eng.* 2017, 29, 828–841. [CrossRef]

23. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 2005, 27, 1226–1238. [CrossRef] [PubMed]

24. Nguyen, X.V.; Chan, J.; Romano, S.; Bailey, J. 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, 24–27 August 2014; pp. 512–521.

25. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* 2004, 5, 1531–1555.

26. Jakulin, A. Machine Learning Based on Attribute Interactions. Ph.D. Thesis, Univerza v Ljubljani, Ljubljana, Slovenia, 2005.

27. Estévez, P.A.; Tesmer, M.; Perez, C.A.; Zurada, J.M. Normalized mutual information feature selection. *IEEE Trans. Neural Netw.* 2009, 20, 189–201. [CrossRef]

28. Meyer, P.E.; Bontempi, G. On the use of variable complementarity for feature selection in cancer classification. In *Proceedings of the Workshops on Applications of Evolutionary Computation*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 91–102.

29. Brown, G.; Pocock, A.; Zhao, M.J.; Luján, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* 2012, 13, 27–66.

30. Verleysen, M.; Rossi, F.; François, D. Advances in feature selection with mutual information. In *Similarity-Based Clustering*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 52–69.

31. Mielniczuk, J.; Teisseire, P. Stopping rules for mutual information-based feature selection. *Neurocomputing* 2019, 358, 255–274. [CrossRef]

32. Daskin, M.S.; Maass, K.L. The p-median problem. In *Location Science*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 21–45.

33. Reese, J. Solution methods for the p-median problem: An annotated bibliography. *Netw. Int. J.* 2006, 48, 125–142. [CrossRef]

34. Avella, P.; Sassano, A.; Vasil’ev, I. Computational study of large-scale p-median problems. *Math. Program.* 2007, 109, 89–114. [CrossRef]

35. Gao, Q.; Yang, Z.; Yin, W.; Li, W.; Yu, J. Internally Induced Branch-and-Cut Acceleration for Unit Commitment Based on Improvement of Upper Bound. *IEEE Trans. Power Syst.* 2022, 37, 2455–2458. [CrossRef]

36. Robnik-Sikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* 2003, 53, 23–69. [CrossRef]

37. Zhao, Z.; Morstatter, F.; Sharma, S.; Aleyani, S.; Anand, A.; Liu, H. Advancing Feature Selection Research. *ASU Feature Selection Repository*. 2010. pp. 1–28.

38. Gay, D.M. The AMPL modeling language: An aid to formulating and solving optimization problems. In *Numerical Analysis and Optimization*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 95–116.

39. Dua, D.; Graff, C. *UCI Machine Learning Repository*; University of California, School of Information and Computer Science: Irvine, CA, USA, 2017.

40. Vanschoren, J.; Van Rijn, J.N.; Bischof, B.; Torgo, L. OpenML: Networked science in machine learning. *ACM SIGKDD Explor. Newsl.* 2014, 15, 49–60. [CrossRef]

41. Goldberger, J.; Hinton, G.E.; Roweis, S.; Salakhutdinov, R.R. Neighbourhood components analysis. In *Proceedings of the Advances in Neural Information Processing Systems* 17 (NIPS 2004), Vancouver, BC, Canada, 13–18 December 2004.

42. Schütze, H.; Manning, C.D.; Raghavan, P. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008; Volume 39.

43. Crammer, K.; Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.* 2001, 2, 265–292.

44. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.

45. Li, Y.; Bai, C.; Reddy, C.K. A distributed ensemble approach for mining healthcare data under privacy constraints. *Inf. Sci.* 2016, 330, 245–259. [CrossRef] [PubMed]