mRelief: A Reward Penalty based Feature Subset Selection Considering Data Overlapping Problem

Suravi Akhter[1*], Sadia Sharmin[3], Sumon Ahmed[1], Abu Ashfaqur Sajib[2*], Mohammad Shoyaib[1]
bsse0827@iit.du.ac.bd, sharmin@iut-dhaka.edu, and
{abu.sajib,sumon,shoyaib}@du.ac.bd

[1]Institute of Information Technology, University of Dhaka, Bangladesh
[2] Department of Genetic Engineering and Biotechnology, University of Dhaka, Bangladesh
[3] Department of Computer Science and Engineering, Islamic University of Technology

Abstract. Feature selection plays a vital role in machine learning and data mining by eliminating noisy and irrelevant attributes without compromising the classification performance. To select the best subset of features, we need to consider several issues such as the relationship among the features (interaction) and their relationship with the classes. Even though the state-of-the-art, Relief based feature selection methods can handle feature interactions, they often fail to capture the relationship of features with different classes. That is, a feature that can provide a clear boundary between two classes with a small average distance may be mistakenly ranked low compared to a feature that has a higher average distance with no clear boundary (data overlapping). Moreover, most of the existing methods provide a ranking of the given features rather than selecting a proper subset of the features. To address these issues, we propose a feature subset selection method namely modified Relief (mRelief) that can handle both feature interactions and data overlapping problems. Experimental results over twenty-seven benchmark datasets taken from different application areas demonstrate the superiority of mRelief over the state-of-the-art methods in terms of accuracies, number of the selected features, and the ability to identify the features (gene) to characterize a class (disease).

Keywords: Feature selection · mRelief · Data overlapping

1 Introduction

Feature selection is the process of selecting a feature subset S from the original feature set F such that S includes the most informative and relevant features for classification. Through this reduction process, the noisy, redundant and irrelevant features are eliminated which in turn improves the classification accuracy, reduces over-fitting as well as the complexity of the model. In general, the existing feature selection methods can be divided into three main categories: wrapper, embedded, and filter methods [29]. Among them, filter based methods are most popular as they are computationally less expensive and not biased to any classification algorithm. Over the years different filter criteria have been introduced such as Correlation [8], Mutual Information (MI) [29, 28, 24], and Distance [18, 13, 10].

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

Earlier methods of feature selection use correlation metric to assess the quality of features. Pearson's correlation coefficient (PCC) is a well-known method in this regard [8]. However, these methods cannot capture the non-linear relationship between features and class variable. MI based selection methods overcome these problems and can detect the nonlinear relationship both for categorical and numerical data containing multiple classes [22]. One of the state-of-the-art methods in this regard, Maximum Relevance Minimum Redundancy (mRMR) [24] selects feature incrementally by maximizing the relevancy between a feature and class variable as well as minimizing the redundancy among the features. However, mRMR discards some features which provides additional information about class despite its redundancy. This problem is solved by Joint Mutual Information (JMI) [21] to some extent. Recently, Gao *et. al.* [7] introduces a new feature selection method called Min-Redundancy and Max-Dependency (MRMD) where a new feature redundancy term is proposed for better approximation of the dependency between the features and class variable. Another MI based hybrid method namely Information-Guided Incremental Selection (IGIS+) is proposed in [23] for gene selection where they employ interaction information for ranking the features and then select the final subset utilizing classifier performance. However, high classification accuracy does not always ensure that the selected genes are relevant to a particular disease identification. Again, it is also well-known that a single feature alone can not predict the class properly without considering its interaction with the other features. Thus, we need to identify the inter-relationship among the features properly to select a better subset of features. However, MI based methods often fail to capture the higher order interaction among the features.

Relief based methods (RBM) such as Relief [18] can capture feature interactions better than the MI based method. Instead of searching through feature combinations, it uses the concept of k nearest neighbors (NN) to derive feature statistics that indirectly account for interactions. RBMs are particularly interesting because they can perform better even if the feature dimension increases. Relief was originally designed for binary classification and extended to ReliefF [15] to handle multiple classes. Both in Relief and ReliefF, we need to choose the value of k appropriately otherwise, it might become difficult to capture the informative information. Instead of fixing the value of k, it will be more advantageous if one can determine a volume from where reliable information can be extracted. Inspired by this idea, SURF [13] define a volume with a radius considering the average distance of all the instances and use the hit and $miss$ within that radius (near hit/$miss$). It helps to capture the informative information even if the interaction is small. RelielfF shows low success rate in this case. Again, along with the $near$ instance, other instances might contain important information and should be used for selecting the features. SURF* [12] was designed to capture two way interaction in feature weighting using near and far instance weighting. Even though it improves the performance of SURF* but requires more computation. To reduce the computational complexity retaining the performance, MultiSURF* [11] is proposed. However, SURF, SURF* [12], and MultiSURF* are mainly designed to handle genomics data (usually features contain few discrete value such as 0, 1 or 2) and might fail to achieve better performance for other problems having different type of data. MultiSURF [10], one

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

of the recent methods in RBM group, solve this problem with less computational time compared to MultiSURF*.

Despite the importance of RBMs, most of them only rank the features and does not consider redundancy. Moreover, they may suffer if data overlapping exists and fail to capture the relationship between a feature and class. To understand this issue, let us consider an example as given in Fig.1 where F_1 and F_2 are two features having instances C_+^1 to C_+^3 for one class and C_-^1 to C_-^4 other class. C_+^T is the target instance. Relief gives more priority to feature F_1 than F_2 though F_2 have better separability. (for details, see illustrative example in section 3.3).

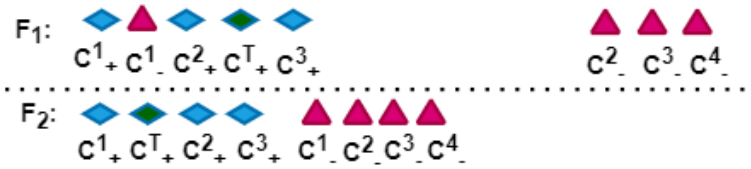


Fig. 1: Data overlapping among different classes

To solve the aforementioned problems, we propose a new feature selection method namely modified Relief (mRelief). The main contributions of this paper are as follows: First, it can capture feature interaction with the help of kNN considering all the features. Second, a $reward - penalty$ scheme is introduced here to detect the data overlapping among different classes and thus, establish a relationship between feature and class properly. Third, a redundancy criteria is proposed to remove the redundant features. In addition to superior performance in different datasets, mRelief can identify the important genes for disease identification.

The rest of the paper is organized in the following four sections. Section 2 describes the preliminaries required to understand the paper, our proposed mRelief is presented in section 3, dataset description, implementation details and experimental results are described in section 4. Finally, section 5 concludes the paper.

2 Preliminaries

Several relief based methods namely Relief, ReliefF, SURF, SURF*, MultiSURF, and MultiSURF* are described in this section.

2.1 Relief

Relief [18] is a distance based feature selection method that uses 1NN to find the nearest two instances of two different classes from a target instance. The distance is called *hit* $(\delta_h(x_n^{f_i}))$ when the target and NN belong to the same class and *miss* $(\delta_m(x_n^{f_i}))$ otherwise. The score (J_{relief}) of Relief for a feature f_i is updated using Eq.1.

$$J_{relief}(f_i) = J_{relief}(f_i) - \frac{\delta_h(x_n^{f_i})}{N} + \frac{\delta_m(x_n^{f_i})}{N} \tag{1}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

here, N is the total number of samples. For nominal feature values, both δ_h and δ_m are '0'if the values are same and '1'otherwise. For numerical values, normalized feature values are used to calculate distances.

2.2 ReliefF

ReliefF [15] is an extension of Relief to multi-class problem where instead of 1NN, kNN is used to calculate *hit* and *miss* distances. Like Relief, it also calculates kNN for the same class (hits) and for each of other classes (misses). Finally, the score ($J_{ReliefF}$) is updated using Eq.2 where $p(y_n)$ is the miss and $p(c_n)$ is the hit class probability.

$$J_{ReliefF}(f_i) = J_{ReliefF}(f_i) - \sum_{i=1}^{K} \frac{\delta_{h_i}(x_n^{f_i})}{m * k} + \sum_{c_n \neq y_n} \frac{p(y_n)}{1 - p(c_n)} \sum_{i=1}^{K} \frac{\delta_{m_i}(x_n^{f_i})}{m * k} \quad (2)$$

2.3 SURF and SURF*

SURF [13] considers the nearest instances that have small distances than a threshold (T) using Eq.3, where, 'T'is defined by taking the average distance of all instance pairs.

$$\delta(x_i, x_j) < T \quad (3)$$

here, x_i is the target instance and x_j is the *hit/miss* instance. Considering these hit and $miss$, SURF score, J_{surf} is calculated as using Eq.2.

SURF* [12] utilizes both the near and far instances (outside T) and its score, (J_{surf*}) is calculated by combining near and far instances. Near score is calculated using the traditional ReliefF scoring method (Eq.2). Far scoring is the opposite of near scoring that is this sign of the second (hit) and third term ($miss$) of the right hand side of Eq.2 become opposite.

2.4 MultiSURF* and MultiSURF

In MultiSURF* [11], threshold (T) for a target instance is identified by taking average distance of all instances from the target instance defined in Eq.4. Note that, T remains same in SURF*, but it varies in MultiSURF* for the target instance.

$$T_i = \frac{\sum_{j=1}^{N-1} \delta(x_j, x_i)}{N - 1} \quad (4)$$

In case of MultiSURF*, a dead-band zone (D_i) is defined using the standard deviation (σ_i) of the distances during the calculation of T_is. A x_n is near when $\delta(x_i, x_j) < (T_i - D_i)$ and far when $\delta(x_i, x_j) > (T_i + D_i)$. MultiSURF considers only the $near$ score instead of both $near$ and far as in MultiSURF*.

3 Modified Relief (mRelief)

We propose modified Relief (mRelief) that includes a $reward - penalty$ scheme to select a subset of features. The process of selecting features using mRelief can be divided into two parts: Candidate Feature Selection and Final Feature Subset Selection, which are described as follows:

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

3.1 Candidate Feature Selection

In this step, we filter the feature set F by removing the irrelevant and noisy features that do not contribute to the classification. A feature is irrelevant if it does not possess the ability to discriminate among the different classes properly. To measure this ability, we use paired t-test considering two sets of distances: one for *hit* and another for *miss* for a particular feature f_i. This test helps us to determine if the means of *hit* (μ_h) and *miss* (μ_m) distances are significantly different from each other under the null hypothesis, $H_0 : \mu_h = \mu_m$. When f_i accept (H_0) on α confidence interval, it is considered as irrelevant feature and removed from the feature set. Based on this hypothesis testing, we define feature irrelevance as given in Definition 1. Following this process, we remove the irrelevant and noisy features and obtain a candidate feature set S^c.

Definition 1. ***Feature Irrelevance:*** *A feature f_i is called irrelevant if there is no significant difference between the distances of hit and miss.*

3.2 Final Feature Subset Selection

Here, we first rank the candidate features S^c based on their individual relevance and then, select a subset of features S that jointly maximizes the relevancy as well as minimizes redundancy among them which is described as follows:

Ranking the Candidate Features: To rank the features in S^c, we assign a value for each feature f_i based on its score J_1 as defined in Eq.5.

$$J_1(f_i) = \frac{1}{N}\sum_{n=1}^{N} \exp\left(-\frac{\delta_m(x_n^{f_i})}{\delta_h(x_n^{f_i})}\right) \tag{5}$$

here, $x_n^{f_i}$ is the n^{th} instance of f_i, N is the total number of instances, $\delta_h(x_n^{f_i})$ and $\delta_m(x_n^{f_i})$ are the average distances of k nearest hit and miss respectively from a particular target instance (x_n). J_1 represents the class discrimination ability of f_i and it decreases if the target instance differs less from the nearby instances of the same class than the nearby instances of the other classes, and increases otherwise. It is well-known that the most discriminatory features have the least uncertainties. By minimizing the score J_1, we expect to identify the features that have better discrimination capability among the classes which in turn reduces the uncertainty. With the help of derivation given in [31], the relation between $score$ and uncertainty is specified in Theorem 1.

Theorem 1. *Score minimization is equivalent to entropy minimization.*

Proof. It is straightforward to show that the distance (δ_h) from x_n to one of its NN ($h(x_n^{f_i})$) within the same class (*hit*) of a feature f_i is proportional to the volume (V) of the hyperspheres having the radius ($x_n^{f_i}$-$h(x_n^{f_i})$). At the same time, the posterior probability, $p(c_n|x_n^{f_i})$ of x_n to be class c_n is equivalent to $\frac{1}{NV}$. Therefore, it is evident that $\delta_h \propto p(c_n|x_n^{f_i})$. From the set of probabilities calculated for all the instances of f_i, one can calculate the uncertainties of the instances to be the member of that class by using entropy $H(x^{f_i}) = \sum_{n=1}^{N} p(x_n^{f_i}) \log p(x_n^{f_i})$. Similarly, for the other classes

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

(*miss*), it can be shown that $\delta_m \propto p(y_n|x_n^{f_i})$ where $c_n \neq y_n$. Combining both *hit* and *miss* distances as given in Eq.5. and their associated probabilities, it can easily be shown that minimizing J_1 also minimizes the entropy. This proves the theorem.

Note that, entropy minimization does not always ensure $score$ minimization, but $score$ minimization always ensure entropy minimization as shown in Theorem 1. However, only minimization of J_1 in Eq.5 is not reliable enough due to data overlapping among different classes as shown in Figure 1. To solve this problem, we introduce $reward$-$penalty$ scheme. The definition of $reward$ is given in Definition 2.

Definition 2. ***Reward:*** *Reward is a distance based measure that indicate the clear class separability of hit and miss instances and can be calculated using Eq.6.*

$$dR = \frac{max(\delta_{h_k})}{min(\delta_{MoH})} * \frac{k - n_{MoH}}{k} \tag{6}$$

here, δ_{h_k} are the distances of the k nearest hit, δ_{MoH} is the distances of $miss$ outside hit, and n_{MoH} is the number of $miss$ outside hit. On the other hand, $penalty$ term is defined in Definition 3.

Definition 3. ***Penalty:*** *Penalty is a distance based measure that indicates the amount of mixing of hit and miss instances and can be calculated using Eq.7*

$$dP = \frac{min(\delta_{MiH})}{max(\delta_{h_k})} * \frac{k - n_{MiH}}{k} \tag{7}$$

here, n_{MiH} is the number of $miss$ inside of hit. The values of dR and dP ranges from 0 to 1. The lower the value of dR and dP, the more the reward and penalty. Incorporating dR and dP to Eq.5, we propose a new score J_2 defined in Eq.8 to calculate the relevance of each feature $f_i (\in S^c)$ and sort them in ascending order to get the ranking of the features.

$$J_2(f_i) = \frac{1}{N} \sum_{n=1}^{N} \exp\big(- \frac{\delta_m(x_n^{f_i})}{\delta_h(x_n^{f_i})} - \frac{dP(x_n^{f_i})}{dR(x_n^{f_i})}\big) \tag{8}$$

Generating a Criteria for Subset Selection The ranking based on J_2 does not confirm the best combination of features set for classification. Moreover, redundant features may exist in the ranking that need to be eliminated despite its relevancy. For this, let us define redundancy (δ_{red}) in term of distance which is given in Definition 4.

Definition 4. ***Redundancy:*** *A feature f_i is fully redundant in term of distance with the selected feature subset S when it has same discrimination ability as f_j ($f_j \in S$) and can be calculated using Eq.9.*

$$\delta_{red}(f_i, S) = \sum_{f_j \in S} \frac{1}{N} \sum_{n=1}^{N} \exp\big(-\frac{\delta_m(x_n^{\{f_j, f_i\}})}{\delta_h(x_n^{\{f_j, f_i\}})}\big) - J_1(f_i) \tag{9}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

here, $\delta_h(x_n^{\{f_j,f_i\}}$ represents distance considering the target instance from f_i and calculating $J_1(f_i)$ for hit and $miss$ instance for f_j. Lower value of δ_{red} indicates high redundancy with the selected feature.

Eq.8 calculates J_2 for a single feature f_i. This equation can also be used for a set of features S using kNN for that S. By incorporating this redundancy term with Eq.8 that calculates distances considering all the features in S and f_i, we propose our final criteria (J_3) for feature subset selection as given in Eq.10.

$$J_3(f_i, S) = \frac{1}{N}\sum_{n=1}^{N} \exp\left(-\frac{\delta_m(x_n^{S,f_i})}{\delta_h(x_n^{S,f_i})} - \frac{dP(x_n^{S,f_i})}{dR(x_n^{S,f_i})} - \delta_{red}\right) \tag{10}$$

Search Strategy: We adopt a greedy forward search strategy to select the best subset of features without having low redundancy among them. At first, we include the top ranked (obtained using J_2) feature to our final subset S as the first selected feature. After that, we consider the subsequent features ($S^c \setminus f_1$) in the ranked list one by one and evaluate their goodness in combination with the selected subset S using the score J_3. The overall process of mRelief is shown in Algorithm-1. If the feature f_i minimizes the J_3, it shows that f_i gives additional information with the selected feature and is added f_i to S otherwise discarded. Following this process, we select the final subset of feature.

Algorithm 1 mRelief

Input: Dataset (D): all instances, X ={ x_1,x_2,x_3, ...x_n} and features, F={f_1,f_2,f_3, ...f_p}
Parameter: Number of neighbour (k)
Output:Subset of features, S $\subseteq$ F

Select the candidate feature set S^c performing t-test
Calculate relevance score J_2 of each feature $f_i \in S^c$ using Eq-(8)
Sort S^c based on their score J_2 in ascending order
select f_1 with minimum score J_2 value
$S \leftarrow f_1; S^c \leftarrow S^c \setminus f_1; T \leftarrow J_2(f_1)$
for all i = 2: $|S^c|$ **do**
    Calculate score $J_3(f_i, S)$ using Eq-(10)
    if $J_3(f_i, S) < T$ **then**
        $S \leftarrow S \cup f_i; T \leftarrow J_3(f_i, S)$
    end if
    $S^c \leftarrow S^c \setminus f_i$
end for
return S

3.3 An Illustrative Example

To understand the impact of $reward$-$penalty$ scheme, let us consider an example where F_1 and F_2 are two features having the instance values shown in Table 1. Here C_x^y represents instance y (1 to m) belong to class x (+ and -) and T represents the target

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

Table 1: Sample dataset for the illustrative example

Instance	C_+^1	C_-^1	C_+^2	C_+^T	C_+^3	C_-^2	C_-^3	C_-^4
F_1	0.13	0.15	0.17	0.18	0.19	0.6	0.62	0.65
F_2	0.11	0.3	0.14	0.12	0.15	0.31	0.35	0.40
class	+	-	+	+	+	-	-	-

instance. Considering Eq.2, the score of ReliefF for F_1 and F_2 are 0.0342 and 0.0225 respectively. According to ReliefF, F_1 gets higher value compared to F_2. However, it is evident that the instances of F_2 are more clearly separable than F_1 and thus, F_2 should get higher priority. This is because, for F_1, an instance of '-ve'class resides within the '+ve'class (also shown in Fig.1. In case of mRelief (Eq.8), the score of F_1 and F_2 are $0.1368e^{-38}$ and $\simeq 0$ respectively. As mRelief gives priority to the minimized score, it ranks F_2 better than F_1 which is desired.

4 Result Analysis and Discussions

To demonstrate the experimental result, we first describe the datasets and then, present the experimental setup of different methods along with the proposed one and their evaluation process. Finally, mRelief is compared with other state-of-the-art methods from different aspects.

4.1 Dataset Description

To compare mRelief with other state-of-the-art methods, we use twenty benchmark datasets collected from UCI machine learning repository[1]. We also use seven cancer related datasets namely (Lung [2], CNS [26], SRBCT [17], Colon [1], Leukemia [9], MLL [25]) and GDS3341 [5] collected from different sources. The characteristics of the datasets are presented in column 2 to 4 of Table 2.

4.2 Implementation Detail

We conduct 10 fold cross-validation (10-CV) for the datasets with a large number of samples (> 250), ten runs of 5-CV for the datasets having their samples between 100 and 250; and Leave-One-Out (LOO) otherwise. For fair comparison, the same strategy is followed for all other methods used in this paper. The results of mRelief are compared with other RBMs such as ReliefF, SURF, SURF*, MultiSURF*, and MultiSURF. We also compare mRelief with two MI based methods namely MRMD and IGIS+. Features are normalized to the interval [0,1] for all RBMs along with mRelief as suggested in [15]. For MRMD and IGIS+, we follow the suggested discretization as given in [7] and [23]. There are various type of classifier such as SVM, KNN, XGBoost [4] and for imbalance data PEkNN [16], kENN [19] can be used to measure the accuracy. The average accuracy of n-fold cross-validation is calculated using SVM (linear kernel)

[1] https://archive.ics.uci.edu/ml

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

Table 2: Dataset overview and performance comparison among different algorithms in terms of Accuracy. Significant win is marked as * and loss as ° means mRelief significantly perform better/worse in the comparison of existing methods.

Dataset	Features	Class	Instance	MRMD	MultiSURF	MultiSURF*	SURF*	SURF	ReliefF	mRelief
yeast	8	10	1484	0.517*	0.532	0.522*	0.562	0.497*	0.560	**0.563(7)**
wine	13	3	178	0.963*	0.982	0.983	0.984	0.987	0.984	**0.989(12)**
heart	13	2	270	0.819	0.833	0.804	0.800*	0.833	0.833	**0.837(11)**
segment	19	7	2310	0.894*	0.919	0.800*	0.809*	0.914	0.906	**0.926(10)**
steel	27	7	1941	**0.684**	0.649	0.676	0.677	0.637	0.662	0.668(18)
ionosphere	33	2	351	0.833	0.822	0.831	0.819	0.836	0.831	**0.856(17)**
dermatology	34	6	366	0.948	0.815*	0.838*	0.868*	0.800*	0.840*	**0.950(13)**
appendicitis	7	2	106	0.850	0.858	0.833	0.842	0.875	0.842	**0.875(6)**
german	20	2	1000	**0.763**	0.747	0.742	0.740	0.749	0.753	0.752(15)
sonar	60	2	208	0.755	0.764	0.790	0.750	0.764	0.759	**0.791(17)**
libras	91	15	360	**0.744**°	0.571*	0.469*	0.527*	0.567*	0.584*	0.702(29)
page-blocks	10	5	5472	0.912*	0.923	0.901*	0.901*	0.921	0.921	**0.921(5)**
saheart	9	2	462	0.710	0.681	0.668*	0.685	0.672*	0.685	**0.717(6)**
southgerman	21	2	1000	**0.771**	0.753	0.752	0.753	0.750	0.754	0.767(15)
page-blocks0	10	2	5472	0.926*	0.918*	0.901*	0.901*	0.928	0.906*	**0.933(7)**
vehicle0	18	2	846	0.819*	0.822*	0.921	0.921	0.809*	0.833*	**0.926(6)**
ecoli3	7	2	336	**0.926**°	0.886	0.886	0.886	0.886	0.886	0.886(6)
new-thyroid1	5	2	215	0.941	0.945	0.818*	0.818*	0.945	0.945	**0.950(2)**
musk	166	2	476	0.783	0.790	0.775	0.775	**0.810**	0.763*	0.796(45)
semeion	256	10	1593	0.874*	0.820*	0.750*	0.759*	0.867*	0.855*	**0.895(99)**
Win/ Tie/ Loss	-	-	-	15/0/5	19/1/0	19/1/0	18/2/0	17/2/1	18/1/1	-
Sig. Win/ Loss	-	-	-	7/2	5/0	9/0	8/0	6/0	6/0	-

to measure the performance of the compared methods. We use the same number of features that mRelief selects for all other compared methods (except IGIS+) as they are all ranking methods.

4.3 Comparison of mRelief with the State-of-the-art Methods

Table 2 and Table 3 present the average accuracies of mRelief in comparison with other state-of-the-art methods along with the number of selected features given in the parenthesis. For these tables, win/tie/loss is calculated using the accuracies and shown in the second last row of the tables. To evaluate the significance of the improvements in accuracies among different methods, paired t-tests (at 95% significance level) are performed and shown in the last row of the tables. Here, Win/Tie/Loss indicates the number of datasets for which mRelief performs better/equally-well/worse than other aforementioned methods. The best performing method is presented in boldfaced. Analyzing these tables, we find that for most of the datasets, mRelief outperforms the other methods with less number of features. The reason behind such performance is that mReleif is able to detect feature interaction, data overlapping and redundant features. The detailed discussion of experimental results is presented as follows.

Impact of feature interaction and data overlapping To understand the impact of feature interaction and data overlapping, let us consider the data visualization for $Semeion$ dataset shown in Fig.2. We observe a major data overlapping between class 3 and 10 in this figure. As an MI based method, MRMD can handle data overlapping, its overall accuracy for this dataset is 87.4% which is close to mRelief (89.5%) (mRelief significantly wins in this case). Note that, MRMD can classify class 3 and 10 more accurately

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

Table 3: Performance comparison among different algorithms on high Dimensional Datasets in terms of Accuracy. Significant win is marked as * and loss as ° means mRelief significantly perform better/worse in the comparison of existing methods.

Dataset	Features	Class	Instance	IGIS+	MultiSURF	MultiSURF*	SURF*	SURF	ReliefF	mRelief
Lung	12600	5	203	0.902(9)	0.852*	0.892*	0.899*	0.871*	0.918*	**0.931(77)**
CNS	7129	2	60	0.618(3)	0.582*	0.603*	0.629	0.605	0.628	**0.631(7)**
SRBCT	2308	4	83	0.923(9)*	0.987*	0.942*	0.928*	0.982*	0.994*	**0.997(51)**
Colon	2000	2	62	0.727(3)*	0.829	0.695*	0.649*	**0.845**	0.835	0.843(21)
Leukemia	7129	2	72	0.903(3)	0.932	0.865*	0.861*	0.937	0.937	**0.939(37)**
MLL	12582	3	72	0.843(5)*	0.944*	0.873*	0.844*	0.953	0.951*	**0.965(64)**
GDS-3341	30802	2	41	0.862(4)	**1.000**	0.780	0.927	0.976	**1.000**	**1.000(26)**
Win/ Tie/ Loss	-	-	-	7/0/0	6/1/0	7/0/0	7/0/0	6/0/1	6/1/0	-
Sig. Win/ Loss	-	-	-	3/0	4/0	5/0	5/0	2/0	3/0	-

than RBMs (with 80% and 87% accuracy respectively), mRelief achieve 88.7% and 89.07% accuracy in this case. It shows although MRMD performs better than other RBMs, it can not exceed the performance of mRelief due to its lower capability of approximating the higher-order feature interaction. Instead of MRMD, we use IGIS+ (an MI based method) for the genomics datasets which is mainly designed to detect genes (features) for such datasets. This reason helps mRelief to win for all genomics dataset compared to IGIS+ shown in Table 3. On the other hand, RBMs (Table 3) can capture feature interaction but they confuse among the classes if data overlapping exists. Therefore, even the best performing RBM (SURF) does not perform well and the accuracies of class 3 and 10 are 85.2%, and 82.9% respectively. These results demonstrate the superiority of mRelief over the other methods in terms of capturing interactions and the ability to handle data overlapping.

Impact of feature selection mRelief uses Eq.10 to select the best set of features. Note that, among all the compared methods, mRelief is the only feature selection method. To demonstrate the capability of mRelief in this regard, we plot the accuracies of different methods for different number of selected features using $Lung$, $Semeion$ and MLL dataset as shown in Fig.3 (arrow mark indicates the performance of total se-

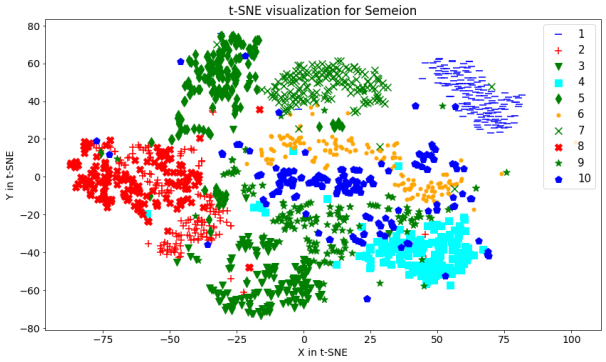
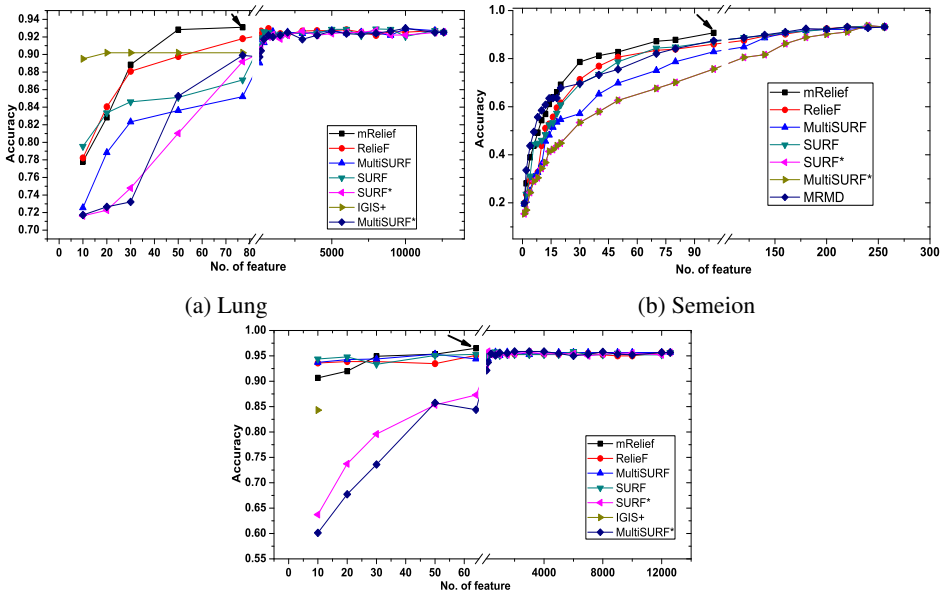


Fig. 2: Data Visualization using tSNE[20]. Here, (1-10) represents different classes.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24



(a) Lung (b) Semeion

(c) MLL

Fig. 3: Comparison of accuracies for different number of selected features

lected features using mRelief). From these figures, we observe that mRelief can identify the relevant features properly and reach the highest point of accuracy for $Lung$, and MLL dataset and very close to the highest accuracies for $Semeion$ dataset. Note that, for $Lung$ dataset, mRelief selects 7089 candidate features from 12600 features and achieves 93.2% accuracy using 77 selected features. On the other hand, other methods require considerably more features to attain such performance. This justifies the purpose of feature selection methods that the selected features are highly relevant to the class and less redundant among themselves.

Impact of redundancy mRelief achieves a better ranking with a small set of features as it can identify non-redundant features for most of the datasets. For this, let us consider $Lung$ and MLL datasets shown in Fig.3. In MLL dataset, mRelief performs better than the existing methods with less number of selected features, because mRelief has the mechanism of removing the redundant features and selects the features having better class separability. However, RBMs achieve similar accuracies with a higher number of features as they can not eliminate the redundant ones. In MLL dataset, IGIS+ obtains satisfactory performance (though these accuracies are lower compared to mRelief) with a fewer number of selected features in comparison with mRelief. We use IGIS+ only for gene datasets as the proposal is only for these types of datasets. However, the selected features often fail to identify the relevant genes for a disease whereas the selected features of mRelief are truly relevant for classification as well as disease identification which is described in the following subsections.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

Table 4: Ranking of the selected gene

Method	Pathway	Rank	FDR
Using both score and frequency			
mRelief	Proteasome	1	4.41E-42
	Epstein-Barr virus infection	2	1.62E-32
	Cell cycle	3	1.40E-23
	Viral carcinogenesis	4	7.13E-18
	Hepatitis B	5	1.09E-15
	Kaposi's sarcoma-associated herpesvirus infection	6	3.19E-14
	Pathways in cancer	8	4.95E-13
Using the respective method's score			
mRelief	Viral carcinogenesis	2	6.94E-10
	Hepatitis B	3	8.11E-10
	Kaposi's sarcoma-associated herpesvirus infection	4	1.17E-08
	Epstein-Barr virus infection	5	2.58E-07
ReliefF	Pathways in cancer	4	3.06E-11
	Kaposi's sarcoma-associated herpesvirus infection	5	3.29E-11
SURF	N/A		
MultiSURF	Pathways in cancer	1	0.000716
SURF*	Hepatitis B	5	6.07E-12
	Viral carcinogenesis	7	7.88E-12
MultiSURF*	Viral carcinogenesis	2	1.31E-10
	Cell cycle	6	2.06E-09
IGIS+	Pathways in cancer	4	0.387

Impact of Important Gene Identification To demonstrate that mRelief can select relatively a small subset of best performing features (gene) to identify a particular class (disease related gene), we use $GDS3341$ [5], gene expression dataset. Both qualitative and quantitative changes in gene expression contribute to the development of different diseases. Therefore, we investigate the biological significance of the top fifty genes selected by different methods, except IGIS+. IGIS+ can select only ten different genes, which are used as input in further analysis. To explore the biological importance of the top selected genes, we use NetworkAnalyst 3.0 [33] for identifying the cellular pathways.

The $GDS3341$ dataset includes nasopharyngeal carcinoma tissue samples. Epstein-Barr virus (EBV) is well known to cause nasopharyngeal carcinoma (NPC), which is an epithelial cancer prevalent in Southeast Asia [3, 32]. Hepatitis B virus (HBV) infection plays a role in the development of NPC [30]. Epstein-Barr virus (EBV) and human herpesvirus, which is also known as Kaposi sarcoma-associated herpesvirus (KSHV), belong to the human gammaherpes virus family [6]. EBV manipulates the ubiquitin-proteasome system in EBV-associated malignancies [14]. Table 4 shows the pathways that have been identified based on protein-protein interaction networks. The list of pathways are ranked based on false discovery rates (FDR), which are adjusted p-values used in the analysis of large datasets generated in high-throughput experiments in order to correct for random events that falsely appear significant [27]. Only the top ten most significant pathways are considered. Table 4 shows only those pathways relevant to NPC. Based on the ranks of the NPC associated pathways, mRelief unequivocally perform better than the other methods. Although the selected genes of IGIS+ can identify "Pathways in cancer" with only ten selected genes, it is not significant statistically ($FDR > 0.05$).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

5 Conclusion

In this paper, we propose mRelief that selects a better feature subset with higher accuracies compared to the state-of-the-art methods over a large set of benchmark datasets. Moreover, it identifies a set of features that is highly representative of a particular class and thus can be used in different applications including gene selection for disease identification. However, mRelief can be adapted using SURF or MultiSURF instead of fixed k, which we will address in the future.

Acknowledgement

This research is supported by a grant (19IF12116, 2019-2020) from the Innovation fund of the ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh. The authors are thankful for the support.

References

1. Alon, Uri, e.a.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences **96**(12), 6745–6750 (1999)
2. Bhattacharjee, Arindam, e.a.: Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences **98**(24), 13790–13795 (2001)
3. Cao, Y.: Ebv based cancer prevention and therapy in nasopharyngeal carcinoma. NPJ precision oncology **1**(1), 1–5 (2017)
4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd. pp. 785–794 (2016)
5. Dodd, Lori E., e.a.: Genes involved in dna repair and nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma. Cancer Epidemiology and Prevention Biomarkers **15**(11), 2216–2225 (2006)
6. Frere, Corinne, e.a.: Therapy for cancer-related thromboembolism. Seminars in oncology **41**(3) (2014)
7. Gao, W., Hu, L., Zhang, P.: Feature redundancy term variation for mutual information-based feature selection. Applied Intelligence **50**(4), 1272–1288 (2020)
8. Goh, L., Song, Q., Kasabov, N.: A novel feature selection method to improve classification of gene expression data. In: Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29. pp. 161–166. Australian Computer Society, Inc. (2004)
9. Golub, Todd R., e.a.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science **286**(5439), 531–537 (1999)
10. Granizo-Mackenzie, D., Moore, J.H.: Multiple threshold spatially uniform relieff for the genetic analysis of complex human diseases. In: EvoBIO. pp. 1–10. Springer (2013)
11. Granizo-Mackenzie, D., Moore, J.H.: Multiple threshold spatially uniform relieff for the genetic analysis of complex human diseases. EvoBIO. Springer, Heidelberg pp. 1–10 (2013)
12. Greene, C.S., Himmelstein, D.S., Kiralis, J., Moore, J.H.: The informative extremes: using both nearest and farthest individuals can improve relief algorithms in the domain of human genetics. In: EvoBIO. pp. 182–193. Springer (2010)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24

13. Greene, C.S., Penrod, N.M., Kiralis, J., Moore, J.H.: Spatially uniform relieff (surf) for computationally-efficient filtering of gene-gene interactions. BioData mining **2**(1), 1–9 (2009)
14. Hui, K.F., Tam, K.P., Chiang, A.K.S.: Therapeutic strategies against epstein-barr virus-associated cancers using proteasome inhibitors. Viruses **9**(11), 352 (2017)
15. Igor, K.: Estimating attributes: analysis and extensions of relief. European conference on machine learning. Springer, Berlin, Heidelberg (1994)
16. Kadir, M.E., Akash, P.S., Sharmin, S., Ali, A.A., Shoyaib, M.: A proximity weighted evidential k nearest neighbor classifier for imbalanced data. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 71–83. Springer (2020)
17. Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F.: Classification and diagnostic prediction of cancers using gene expression profiling 300 and artificial neural networks. Nature medicine **7**, 673–679 (2001)
18. Kira, Kenji, Rendell, L.A.: The feature selection problem: traditional method and a new algorithm. AAAI **2**, 129–134 (1992)
19. Li, Y., Zhang, X.: Improving k nearest neighbor with exemplar generalization for imbalanced classification. In: PAKDD. pp. 321–332. Springer (2011)
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
21. Moody, J., Yang, H.: Data visualization and feature selection: New algorithms for nongaussian data. Advances in neural information processing systems **12**, 687–693 (1999)
22. Naghibi, T., Hoffmann, S., Pfister, B.: A semidefinite programming based search strategy for feature selection with mutual information measure. IEEE transactions on pattern analysis and machine intelligence **37**(8), 1529–1541 (2014)
23. Nakariyakul, S.: A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification. PloS one **14**(2) (2019)
24. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence **27**(8), 1226–1238 (2005)
25. Pollack, Jonathan R., e.a.: Genome-wide analysis of dna copy-number changes using cdna microarrays. Nature genetics **23**(1), 41–46 (1999)
26. Pomeroy, S., e.a.: Gene expression-based classification and outcome prediction of central nervous system embryonal tumors. Nature **415**(24), 436–442 (2002)
27. Rouam, S.: False discovery rate (fdr)." encyclopedia of systems biology. Cancer Epidemiology and Prevention Biomarkers **36**, 731–732 (2013)
28. Roy, P., Sharmin, S., Ali, A.A., Shoyaib, M.: Discretization and feature selection based on bias corrected mutual information considering high-order dependencies. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 830–842. Springer (2020)
29. Sharmin, S., Shoyaib, M., Ali, A.A., Khan, M.A.H., Chae, O.: Simultaneous feature selection and discretization based on mutual information. Pattern Recognition **91**, 162 – 174 (2019)
30. Weng, Jing-Jin, e.a.: Effects of hepatitis b virus infection and antiviral therapy on the clinical prognosis of nasopharyngeal carcinoma. Cancer medicine **9**(2), 541–551 (2020)
31. Yang, S.H., Hu, B.G.: Discriminative feature selection by nonparametric bayes error minimization. IEEE Transactions on knowledge and data engineering **24**(8), 1422–1434 (2012)
32. Young, L.S., Dawson, C.W.: Epstein-barr virus and nasopharyngeal carcinoma. Chinese journal of cancer **33**(12), 581 (2014)
33. Zhou, Guangyan, e.a.: Networkanalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic acids res. **47**(W1), W234–W241 (2019)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_24