Discretization and Feature Selection Based on Bias Corrected Mutual Information Considering High-Order Dependencies

Puloma Roy[1(✉)], Sadia Sharmin[2], Amin Ahsan Ali[3], and Mohammad Shoyaib[1]

[1] Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh
pulomaa92@gmail.com, shoyaib@du.ac.bd
[2] Department of CSE, Islamic University of Technology, Gazipur City, Bangladesh
sharmin@iut-dhaka.edu
[3] Department of CSE, Independent University of Bangladesh, Dhaka, Bangladesh
aminali@iub.edu.bd

Abstract. Mutual Information (MI) based feature selection methods are popular due to their ability to capture the nonlinear relationship among variables. However, existing works rarely address the error (bias) that occurs due to the use of finite samples during the estimation of MI. To the best of our knowledge, none of the existing methods address the bias issue for the high-order interaction term which is essential for better approximation of joint MI. In this paper, we first calculate the amount of bias of this term. Moreover, to select features using χ^2 based search, we also show that this term follows χ^2 distribution. Based on these two theoretical results, we propose Discretization and feature Selection based on bias corrected Mutual information (DSbM). DSbM is extended by adding simultaneous forward selection and backward elimination (DSbM_{fb}). We demonstrate the superiority of DSbM over four state-of-the-art methods in terms of accuracy and the number of selected features on twenty benchmark datasets. Experimental results also demonstrate that DSbM outperforms the existing methods in terms of accuracy, Pareto Optimality and Friedman test. We also observe that compared to DSbM, in some dataset DSbM_{fb} selects fewer features and increases accuracy.

Keywords: Feature selection · Mutual information · Interaction · Bias correction

1 Introduction

In classification tasks, the objective of feature selection (FS) process is to choose the most useful features that contribute to the prediction of class variable. Usually, all the features of a dataset do not have equal importance, rather some may

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-47426-3_64) contains supplementary material, which is available to authorized users.

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12084, pp. 830–842, 2020.
https://doi.org/10.1007/978-3-030-47426-3_64

create noise or be redundant. FS methods are used to remove such irrelevant and redundant features and can be divided into three broad categories namely Wrapper [14,18], Embedded [20], and Filter methods [13,15,16]. Among these, filter methods do not depend on a classifier to select a feature. It thus works faster, which is preferable for handling large feature sets [12].

Again, Mutual information (MI) is usually popular in filter based methods. MI can capture non-linear relationships among features and class variable, can be computed for both categorical and numerical data, and can deal with multiple classes [7]. For these reasons, in this paper, we focus on MI based filter methods.

In MI based filter methods, the main goal is to select a subset of features S from the original feature set, $F = \{f_1, f_2, f_3, ..., f_n\}$ in such a way that it will maximize joint MI ($I(S;C)$) with the class variable, C as showed in Eq. 1.

$$\begin{aligned} I(S;C) &= I(f_1, f_2, ..., f_k; C) \\ &= \sum_{f_1, f_2, ..., f_k} \sum_{C} P(f_1, f_2, ..., f_k; C) \log \frac{P(f_1, f_2, ..., f_k; C)}{P(f_1, f_2, ..., f_k)P(C)} \end{aligned} \tag{1}$$

However, the computation of $I(S;C)$ is a NP-hard problem [7]. To overcome this problem, different approximations such as MIFS [1], mRMR [10], JMI [19], RelaxMRMR [17] have been proposed over the last decades. In these methods, MI terms such as feature relevancy(R), redundancy(r), conditional redundancy(c) and interaction(i) are considered in order to achieve a better approximation. However, none of the aforementioned methods correct "bias" due to finite samples in calculating MI terms. In a recent method mDSM [16], it is shown that incorporating bias correction for R, r, and c terms improves the classification performance. However, the interaction term is not considered in mDSM which needs to be addressed for better approximation [17].

Apart from the evaluation criteria, searching is an important step in the FS methods to find out the combination of feature subset that performs well. Most popular searching techniques are forward selection, backward elimination, genetic algorithms (GA) based search [11]. Forward selection and backward elimination are greedy searching strategy that select/delete a feature one at a time. The limitation of these approaches are after selecting/deleting a feature, it cannot be deleted/re-selected later which may add redundant features [6]. On the other hand, GA based methods are computationally expensive and for a dataset with large number of features, it is not feasible to apply. Convex based Relaxation Approximation (COBRA) is proposed in [7] which provides a global solution for MI based FS. Another search strategy is introduced in mDSM where a small subset of features is selected using χ^2 based forward selection that uses dynamic discretization. However, it cannot deselect a feature once it is already selected and do not show whether it is possible to use χ^2 based search for interaction term. Considering the aforementioned issues, we propose a method called Discretization and feature Selection based on bias corrected MI (DSbM) and make the following major contribution: First, we calculate bias for the interaction terms and propose to use it for FS. Second, we show that the interaction terms follow χ^2 distribution and proposed to use it in χ^2 based search. Third, to obtain reduced

number of feature, keeping similar performances with DSbM we propose a new method for simultaneous forward selection and backward elimination (DSbM_{fb}).

2 Information Theoretic Feature Selection Methods

The main objective of MI based features selection methods is to determine a subset of features that have maximum dependency with the given class as shown in Eq. 1. Alternatively, this problem can be formulated for incremental feature selection that is to add one feature at a time in the selected subset to maximize $I(S;C)$. From a given set F with n number of features, a new feature f_m is added to the selected set, $S = \{f_1, f_2, ..., f_{m-1}\}$, that maximizes the score for a feature f_m:

$$J(f_m) = I(f_m \cup S; C) = I(S;C) + I(f_m; C \mid S) \tag{2}$$

Since $I(S;C)$ remains constant with respect to f_m, we choose f_m that maximizes $I(f_m; C \mid S)$. Using MI identities, this term can be expressed as

$$I(f_m; C|S) = I(f_m; C) - I(f_m; S) + I(f_m; S|C) \tag{3}$$

here, the terms $I(f_m; C)$, $I(f_m; S)$ and $I(f_m; S|C)$ represent feature relevancy, redundancy and conditional redundancy respectively [2]. Hence the score $J(f_m)$ increases if the relevancy of the feature f_m is large and redundancy with the existing features is low. However, the score also increases if the conditional redundancy is higher than the redundancy term. Hence, there is a trade-off, and the overall score is what needs to be maximized. Brown *et al.* in [2] further shows under the assumption that (a) the selected features in S are independent given the feature f_m and (b) the selected features are class-conditionally independent given the feature f_m and removing terms that have no effect on the choice of f_m one can obtain the following equivalent score function:

$$J(f_m) = I(f_m; C) - \beta \sum_{f_i \in S} I(f_m; f_i) + \gamma \sum_{f_i \in S} I(f_m; f_i|C) \tag{4}$$

with $\beta = 1$ and $\gamma = 1$, this is what we call the Rrc criterion. It can then be easily shown that the incremental FS criterion or score function of well known MI based method such as MIFS [1], mRMR [10], Extended mRMR [9], JMI [19], and MIM [5] can be derived from this parameterized version of the score function. For example, JMI [19] criteria can be derived setting the value of $\beta = \gamma = \frac{1}{|S|}$.

In [17], the authors propose a new criterion by relaxing the the first assumption. They show under the relaxed assumption that the selected features are conditionally independent given the f_m and another feature f_i in S, the redundancy term can be approximated as the following

$$I(f_m; S) = I(f_m; f_i) + \sum_{f_j \in S, i \neq j} I(f_m; f_j|f_i) + \Omega \tag{5}$$

where Ω is not dependent on f_m. Instead of finding a feature f_i to condition on, they propose to average the right-hand side over all $f_i \in S$, resulting in the following score function

$$J_{rMRMR}(f_m) = I(f_m;C) - \frac{1}{\mid S \mid}\sum_{f_i \in S} I(f_m;f_i) + \frac{1}{\mid S \mid}\sum_{f_i \in S} I(f_m;f_i \mid C) - \frac{1}{\mid S \mid\mid S-1 \mid}\sum_{f_i \in S}\sum_{f_j \in S; i \neq j} I(f_m;f_j \mid f_i) \quad (6)$$

here, the $I(f_m;f_j \mid f_i)$ terms are the second order interaction term between the features. It should be noted that sum of the second order terms is normalized by $\frac{1}{|S||S-1|}$ instead of $\frac{1}{|S|}$. The authors note that this is to prevent this sum to out-weight other terms. It can be seen that one can approximate the redundancy term using 3rd or higher order interaction terms by further relaxing the assumption. However, it is shown that the joint MI is more influenced by lower-order interaction terms in case of forward selection methods [4].

Practically, all aforementioned MI terms that have been used for the approximation need bias correction due to the finite number of samples. To solve this issue, a recent method namely, mDSM [16] is proposed where bias corrected MI has been used for calculating relevancy, redundancy and complementary term. They show incorporating bias correction improves the accuracy of classification. Also, it is theoretically shown that these three terms follow χ^2 distributions.

$$J_{mDSM}(f_m) = I(f_m;C) - \frac{(\mathcal{M}-1)(\mathcal{K}-1)}{2N\ln 2} + \frac{1}{\mid S \mid}\sum_{f_i \in S}(I(f_m;f_i \mid C) - \frac{(\mathcal{M}-1)(\mathcal{I}-1)\mathcal{K}}{2N\ln 2} - I(f_m;f_i) + \frac{(\mathcal{M}-1)(\mathcal{I}-1)}{2N\ln 2}) \quad (7)$$

here, $\mathcal{M}$ and $\mathcal{I}$ are the number of intervals in feature f_m and f_i respectively. $\mathcal{K}$ is number of class and N is total number of samples. The limitation of mDSM is that it does not consider the interaction term while proposing bias corrected MI to calculate the feature score which is necessary for better approximation of joint MI.

3 Proposed Method

In this paper, we propose Discretization and feature Selection based on bias corrected MI (DSbM) which incorporates bias correction for MI based selection criteria. DSbM also uses dynamic discretization and greedy χ^2 based forward selection. Moreover, a simultaneous forward selection and backward elimination is also proposed. These are described in the following subsections.

3.1 Discretization and Feature Selection Based on Bias Corrected Mutual Information (DSbM)

DSbM incorporates the bias correction for all four terms mentioned in Eq. 6 as it is necessary for better approximation of joint MI. The bias for the first three

terms are given in Eq. 7. Theorem 1 shows the amount of bias for the interaction term and Theorem 2 shows that this term follows χ^2 distribution. Proof of the theorems are given as supplementary materials due to page limitation.

Theorem 1. *Bias is* $\frac{(\mathcal{M}-1)(\mathcal{J}-1)\mathcal{I}}{2N\ln 2}$ *for Interaction* $I(f_m; f_j \mid f_i)$ *among the features* f_m *and* f_j *given feature* f_i*, where* $\mathcal{I}$*,* $\mathcal{J}$ *and* $\mathcal{M}$ *are the number of intervals in feature* f_i *,* f_j *and* f_m *respectively.*

Incorporating this bias corrected Interaction term with Eq. 7, DSbM uses the following criteria for discretization and feature selection.

$$J_{DSbM}(f_m) = I(f_m; C) - \frac{(\mathcal{M}-1)(\mathcal{K}-1)}{2N\ln 2} + \frac{1}{\mid S \mid}\sum_{f_i \in S}(I(f_m; f_i \mid C) - \frac{(\mathcal{M}-1)(\mathcal{I}-1)\mathcal{K}}{2N\ln 2} - I(f_m; f_i) + \frac{(\mathcal{M}-1)(\mathcal{I}-1)}{2N\ln 2}) - \frac{1}{\mid S \mid\mid S-1 \mid}\sum_{f_i \in S}\sum_{f_j \in S; i \neq j}(I(f_m; f_j \mid f_i) - \frac{(\mathcal{M}-1)(\mathcal{J}-1)\mathcal{I}}{2N\ln 2}) \tag{8}$$

Theorem 2. $I(f_m; f_j \mid f_i)$ *follows* χ^2 *distribution with* $(\mathcal{M}-1)(\mathcal{J}-1)\mathcal{I}$ *degrees of freedom if* f_m*,* f_i *and* f_j *are statistically independent.*

Based on Theorem 2, the critical value of the *Interaction* term will be as Eq. 9

$$\chi^2_{(i)} = 2N\ln(2) * I(f_m; f_j \mid f_i) \tag{9}$$

As the other three terms of Eq. 6 also follows χ^2 distribution, we can use their critical values (shown in [16]) for selecting a new feature.

The overall process of DSbM is given in Algorithm 1. First, each feature $f_m \in F$ is discretized with minimum number of intervals (d_m) for which its relevancy with the class variable ($J_{rel}(f_m) = I(f_m; C) - \frac{(\mathcal{M}-1)(\mathcal{K}-1)}{2N\ln 2}$) is significant. If the feature is not significant even with some predefined maximum number of intervals (d_{max}), it is dropped. The selected candidate features (F_c) are then sorted according to their relevance J_c in descending order (line 2–12 in Algorithm 1). The first feature f_1 is then included to the final selected feature set S. The remaining features of F_c are evaluated incrementally maximizing the *Rrci* criteria. The score of J_{DSbM} (Eq. 8) is compared (in line 15) with its' critical value ($\chi^2_{(Rrci)}$), to select a new feature f_m if it is not significantly redundant. Otherwise, f_m is discarded considering that it does not contribute to the score significantly. While selecting a new feature, its discretization level is also shifted by a small value δ from its original value (as selected previously based on J_{rel} as shown in line 16–21). This process helps to select the discretization level of features dynamically considering its dependency with other feature. In this way, all the features are discretized and selected simultaneously.

3.2 DSbM with Simultaneous Forward Selection and Backward Elimination ($\mathbf{DSbM_{fb}}$)

DSbM follows χ^2 based forward searching strategy where a feature can not be discarded once it is added to the selected subset S. When a candidate feature f_m

Algorithm 1 : DSbM

Input: Set of n features, F, Maximum discretization level d_{max}, Class C
Output: Selected set of features, $S = \{f_1, f_2, \cdots, f_k\}$ with discretization, $D = \{d_1, d_2, \cdots, d_k\}$

Subset of r candidate features, $F_c \leftarrow \emptyset$
for each $f_m \in F$ **do**
  for all $l = 2$ to d_{max} **do**
    Discretize f_m with l interval
    Calculate J_{rel} for feature f_m
    if $J_{rel}(f_m) > \chi^2_{(R)}$ **then**
      $F_c \Leftarrow F_c \cup f_m$; $D_c \Leftarrow D_c \cup l$; $J_c \Leftarrow J_c \cup J_{rel}(f_m)$;
      break
    end if
  end for
end for
Sort F_c with corresponding D_c in decreasing order based on their J_c values
select f_1 with its' corresponding d_1
$S \Leftarrow S \cup f_1$; $D \Leftarrow D \cup d_1$; $F_c \Leftarrow F_c \setminus f_1$;
for each $f_m \in F_c$ **do**
  for all $l = d_m$ - δ to $l = d_m + \delta$ **do**
    Discretize f_m with l interval
    if $J_{DSbM}(f_m) > \chi^2_{(Rrci)}$ **then**
      $d_m \Leftarrow l$; $J_m \Leftarrow J_{DSbM}(f_m)$; $T \Leftarrow \chi^2_{(Rrci)}$;
    end if
  end for
  if $J_{DSbM} > T$ **then**
    $S \Leftarrow S \cup f_m$; $D \Leftarrow D \cup d_m$;
  end if
  $F_c \Leftarrow F_c \setminus f_m$;
end for
Return S and their respective D

is found redundant with respect to the selected features from S, DSbM does not consider f_m for selection. However, it may happen that f_m is more important and contains extra information compared to the already selected features. In this case, removing the redundant features from S is more appropriate. Therefore, we modify DSbM by including backward elimination and propose DSbM_{fb} where simultaneous selection and elimination is incorporated.

The process of backward elimination is described in Algorithm 2. Here, the redundant candidate feature f_m is rechecked based on its interaction value to decide whether this feature f_m is able to replace some features from S. This checking can be done by several ways such as considering all possible combination of three way interaction of f_m with f_i and f_j and selecting the feature pair whose replacement can increase the J_{DSbM} score significantly. However, it is computationally expensive to check all possible combination pairs of features. Hence, we consider the pair for which we obtain the highest interaction value

Algorithm 2 : DSbM$_{\text{fb}}$

Input: Set of n features, F, Maximum discretization level d_{max}, Class C
Output: Selected set of features, $S = \{f_1, f_2, \cdots, f_k\}$ with discretization, $D = \{d_1, d_2, \cdots, d_k\}$

1: Line (1-14) from Algorithm 1
2: **for each** $f_m \in F_c$ **do**
3: Line (16-21) from Algorithm 1
4: **if** $J_{DSbM} > T$ **then**
5: $S \Leftarrow S \cup f_m$; $D \Leftarrow D \cup d_m$;
6: **else**
7: Set of interaction values, $E \Leftarrow \emptyset$
8: **if** $\mid S \mid >= 4$ **then**
9: **for all** $i = 1$ to $\mid S \mid$ **do**
10: **for all** $j = 1$ to $\mid S \mid$ and $i \neq j$ **do**
11: Calculate Interaction, $I(f_m; f_j \mid f_i)$ among feature f_m, f_i, f_j
12: $e_{ij} = I(f_m; f_j \mid f_i)$
13: **end for**
14: **end for**
15: Select feature f_i, f_j with highest interaction value e_{ij} from E
16: $S' \Leftarrow S \setminus \{f_i, f_j\}$
17: **if** $J_{DSbM}(f_m)$ on $S' > \chi^2_{(Rrci)}$ && $J_{DSbM}(f_m)$ on $S' > J_{DSbM}(f_m)$ on S **then**
18: $S \Leftarrow S \cup f_m$; $S \leftarrow S'$; $D \Leftarrow D \cup d_m$; $D \Leftarrow D \setminus \{d_i, d_j\}$;
19: **end if**
20: **end if**
21: **end if**
22: $F_c \Leftarrow F_c \setminus f_m$;
23: **end for**
24: **Return** S and their respective D

(line 9–15) and replace that feature pair with f_m if their removal from S passes the χ^2 value and increases the total score (line 17–18). As a result, DSbM$_{\text{fb}}$ obtains a smaller subset of features compared to DSbM.

4 Experimental Result

In this section, the experimental setup and evaluation process of different methods along with the proposed ones is presented. Furthermore, a number of experiments are performed to highlight the effectiveness of the proposed contributions.

4.1 Dataset Description and Implementation Details

In this experiment, twenty benchmark datasets collected from UCI Machine Learning Repository [3] are used as they are also employed in [16] and [19]. The description of these datasets are given in Table 1. For classification, we use SVM and KNN, and conduct 10-fold cross-validation on each dataset.

We compare DSbM with four state-of-the-art methods namely mDSM, JMI, JMI with COBRA search (JC) and RelaxMRMR. Here, DSbM, mDSM and JC are feature selection method, however, JMI and RelaxMRMR are feature ranking method. Hence, the number of selected feature obtained in DSbM are used to generate the results for these two methods. For JMI and RelaxMRMR, we use forward selection whereas, JC performs COBRA search and mDSM uses χ^2 based search. For comparing the methods we use three metrics namely accuracy, Score (defined in Eq. 10) and Pareto Optimality(PO). PO returns a set of non-dominant candidate solutions.

$$Score = \frac{\sum_{i=1}^{n} w_i * \alpha_i}{\sum_{i=1}^{n} w_i} \tag{10}$$

here, α_i and w_i indicates the performance evaluation criteria and weights respectively. For our method α_1 and α_2 indicates the percentage accuracy, and $\alpha_2 = (N_t - N_s)/N_t$ is the percentage of reduction features. Here, N_t is the total number of features in a dataset and N_s is the number of selected features. We use equal weights. To calculate PO, we use α_1 and α_2 and to perform Friedman test we use Score to incorporate the joint impact of number of selected features and the corresponding accuracy. We also calculate Win/Tie/Loss which indicates the number of datasets for which comparing method performs better/equally-well/worse than other methods unless otherwise stated. To determine whether the wins are statistically significant we perform t-test at 0.05 significance level.

4.2 Results and Discussion

Here, we first discuss how DSbM performs compared to other methods and then we compare the performance of DSbM with DSbM_{fb}.

Comparison of DSbM with Other Methods. To investigate the impact of high-order term for approximating joint MI in DSbM, let us first consider Table 2. For this table, win/tie/loss is calculated using the accuracies given in Table 3. RelaxMRMR performs slightly better than JMI due to the incorporation of interaction term. Whereas, mDSM outperforms RelaxMRMR even though mDSM does not consider high-order term. It is due to the bias correction, dynamic discretization and χ^2 based search. This indicates that mDSM with high-order term might perform well which is the proposed DSbM. mDSM also performs better than JC. Table 3 compares DSbM with mDSM, JC, JMI and RelaxMRMR. The number inside the parenthesis represents the number of selected feature. For example, DSbM achieves 96% accuracy using SVM with 2 selected features for Iris dataset.

It is evident from Table 3 that DSbM outperforms all the four state-of-the-art methods. The second last and the last row of Table 3 represent the pair wise win/tie/loss and significant win/loss of DSbM with the existing methods respectively. Even though DSbM wins in thirteen datasets among the twenty compared to mDSM for SVM classifier, the differences in accuracies are not

Table 1. Dataset description

Index	Dataset	Dimension	Instance	Class	Index	Dataset	Dimension	Instance	Class
1	**Iris**	4	150	3	**11**	**Parkinson**	22	197	2
2	**Pima**	8	768	2	**12**	**Steel**	27	1941	7
3	**Yeast**	8	1484	10	**13**	**Breast**	30	569	2
4	**Glass**	9	214	6	**14**	**Dermatology**	34	366	6
5	**Wine**	13	178	3	**15**	**Spambase**	57	4601	2
6	**Heart**	13	270	2	**16**	**Sonar**	60	208	2
7	**Australian**	14	690	2	**17**	**Liver**	6	345	2
8	**Segment**	17	2310	7	**18**	**Breast Tissue**	9	106	6
9	**Cardio**	21	2126	10	**19**	**Arrhythmia**	279	452	16
10	**Waveform**	21	5000	3	**20**	**Semeion**	256	1593	10

Table 2. Comparison of different methods (Win/Tie/Loss)

	RelaxMRMR vs. JMI	mDSM vs. RelaxMRMR	**mDSM vs. JC**
SVM	7/7/6	13/1/6	14/0/6
KNN	14/1/5	14/0/6	16/0/4

significant in most of the cases. DSbM wins significantly only for three datasets and losses for one. However, DSbM selects less number of features than other feature selection methods as it considers the bias corrected interaction term for which some redundant features are discarded. For example, in Wine dataset the accuracy of SVM is 96.84% for both DSbM and mDSM. However, DSbM selects only 9 features whereas mDSM selects 12. But in some cases mDSM selects less feature than DSbM. This is due to the greedy nature of forward selection and difference in the score functions. DSbM and mDSM may select different features in any iteration due to the inclusion of interaction term in DSbM, This may results in DSMb selecting a larger number of features compared to mDSM (for example, in case of Spambase and Sonar).

To understand the joint impact of accuracy and number of selected features, let us consider Table 4, where the ranking of the above mentioned methods is shown according to their frequency in the PO set and Friedman test. In both cases, DSbM achieves the highest rank. In Friedman test, after rejecting the null hypothesis that all the methods perform equivalently, a post-hoc test called Nemenyi test [8] is used to determine the which method performs significantly better than the others. The test indicates that DSbM significantly (at 95% confidence level) outperforms the four other methods both for SVM and KNN.

Impact of DSbM$_{\mathbf{fb}}$ over DSbM. To understand the impact of simultaneous forward selection and backward elimination using DSbM$_{\text{fb}}$, let us consider Fig. 1a and Fig. 1b. We observe, in most of the cases DSbM$_{\text{fb}}$ selects less features than DSbM (number of selected features is given on the top of each bar and on the x-axis the index of datasets are given according to their order in Table 1). These figures also illustrate that when the total number of features for a dataset

Table 3. Comparison among different methods based on its accuracy. (∗) and (○) represents that DSbM wins and loses significantly from that method respectively and bold values represent the overall win among all methods.

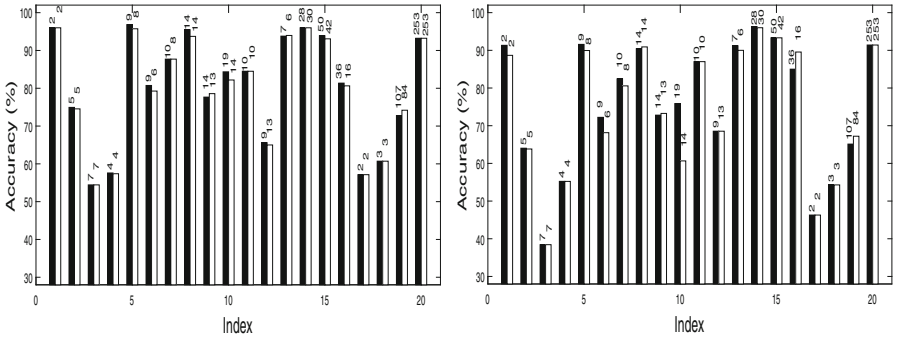
	SVM(accuracy in %)					KNN(accuracy in %)				
	DSbM	mDSM	JC	JMI	Relax MRMR	DSbM	mDSM	JC	JMI	Relax MRMR
Iris	**96.00**(2)	94.67(2)	91.30(2)*	93.33	93.33	**91.33**	86.00*	83.30*	87.32	87.33
Pima	**74.94**(5)	74.29(7)	73.60(8)	73.12	73.12	**64.03**	61.69	58.70*	48.83*	48.84*
Yeast	**54.44**(7)	53.46(7)	50.00(7)*	51.83	51.50*	**38.43**	37.45	30.30*	33.33*	32.55*
Glass	**57.59**(4)	50.00(5)*	51.70(7)	54.35	54.35	**55.23**	54.35	51.30	50.00	51.00
Wine	**96.84**(9)	**96.84**(12)	91.60(9)*	94.21	95.79	91.58	91.58	84.20*	**96.32**	94.74
Heart	80.74(9)	80.00(10)	81.10(10)	**83.33**	82.22	72.22	72.59	71.90	**81.11**$^\circ$	75.93
Australian	**87.71**(10)	**87.71**(10)	68.30(11)*	87.14	87.57	**82.43**	81.14	59.30*	77.43*	75.23*
Segment	**95.51**(14)	94.06(16)	88.80(12)*	89.26*	89.26*	90.43	**91.60**	86.90*	87.45*	87.47*
Cardio	**77.66**(14)	76.88(16)	63.30(13)*	68.99*	68.95*	**72.80**	71.15	61.90*	68.21*	67.02*
Waveform	84.35(19)	**85.29**(19)	80.80(13)*	83.97	83.97	75.87	**76.71**	70.30*	74.81	74.83
Parkinson	**84.50**(10)	83.00(17)	**84.50**(14)	**84.50**	84.00	87.00	87.00	**92.00**	91.50	84.50
Steel	65.61(9)	**72.02**(26)$^\circ$	69.60(20)$^\circ$	51.56*	63.79	68.54	**69.39**	69.30	21.01*	62.42*
Breast	93.79(7)	**96.38**(26)	95.70(20)	93.62	95.17	91.21	**94.66**$^\circ$	92.20	72.93*	89.66
Dermatology	96.00(28)	95.50(33)	95.30(23)	**96.75**	96.50	**96.25**	96.00	93.50*	92.75*	92.75*
Spambase	**93.90**(50)	93.06(47)*	73.30(41)*	74.51*	74.71*	**93.32**	92.62	67.30*	68.31*	68.31*
Sonar	**81.36**(36)	75.91(21)	72.70(60)*	70.45*	73.64*	85.00	83.64	**88.60**	87.27	84.55
Liver	57.14(2)	57.14(2)	**59.43**(6)	57.14	57.14	46.29	46.27	**52.00**$^\circ$	43.14	43.14
Breast Tissue	60.71(3)	**61.43**(4)	55.00(6)	57.81	56.43	**54.29**	53.57	51.43	52.14	49.29
Arrhythmia	**72.79**(107)	66.28(118)*	70.70(253)	72.56	74.19	65.12	65.35	58.60*	64.65	**69.30**
Semeion	**93.23**(253)	92.93(255)	93.20(254)	93.17	93.17	**91.41**	90.73	91.20	91.40	**91.41**
Win/Tie/Loss		**13/3/4**	**15/1/4**	**16/2/2**	**15/1/4**		**12/2/6**	**15/0/5**	**16/0/4**	**16/1/3**
Sig. Win/Loss		**3/1**	**9/1**	**5/0**	**5/0**		**1/1**	**11/1**	**9/1**	**8/0**

Table 4. Ranking of existing feature selection criteria.

	Frequency in Pareto optimal set				
SVM	DSbM(12)	JC(8)	mDSM(6)	JMI(2)	RelaxMRMR(2)
KNN	DSbM(13)	JC(10)	mDSM(6)	JMI(4)	RelaxMRMR(1)
	Average rank from Friedman test				
SVM	DSbM(1.70)	RelaxMRMR(2.73)	JMI(2.85)	mDSM(3.78)	JC(3.95)
KNN	DSbM(1.68)	RelaxMRMR(2.75)	JMI(2.83)	mDSM(3.80)	JC(3.95)

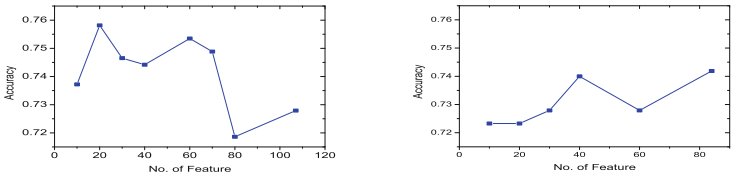
is comparatively small then the performance of both DSbM and DSbM_{fb} are similar in terms of number of selected features and accuracy (e.g., Iris, Yeast, Glass etc.). Note that in some cases such as in Cardio, Arrhythmia etc., DSbM_{fb} selects fewer features with higher accuracy.

Furthermore, a limitation of mDSM is that, the set of selected features might contain a subset for which better accuracy can be found. DSbM also has similar problem which can be observed in Fig. 2a. This issue is resolved to some extent in DSbM_{fb}. Here, we get 74.19% accuracy with 84 selected features (see Fig. 2b) while DSbM obtains an accuracy of 72.79% with 107 features.



(a) SVM (b) KNN

Fig. 1. DSbM(black bar) vs. DSbM_{fb}(white bar)



(a) DSbM (b) DSbM_{fb}

Fig. 2. Accuracy (SVM) vs. Number of features for Arrhythmia dataset

5 Conclusion

In this paper, we propose a method DSbM which includes bias correction for high-order dependencies among features and use χ^2 based search that also consider high-order dependencies. Results over a large amount of dataset demonstrate that DSbM outperforms current state-of-the-art methods. Beside this, a χ^2 based simultaneous forward and backward search is also proposed here that shows similar performances with DSbM with less number of features. This method can be applied for different applications such as activity recognition and cancer classification for gene expression data. Incorporation of further high-order terms might improve the overall performance which require further theoretical analysis and experimentation with global feature selection which will be addressed in future work.

Acknowledgement. This research is supported by ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh. 56.00.0000.028.33.093. 19-427, 20-11-2019.

References

1. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw **5**(4), 537–550 (1994)
2. Brown, G., Pocock, A., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J. Mach. Learn. Res. **13**(1), 27–66 (2012)
3. Dua, D., Graff, C.: UCI machine learning repository (2017)
4. Lee, J., Kim, D.W.: Mutual information-based multi-label feature selection using interaction information. Exp. Syst. Appl. **42**(4), 2013–2025 (2015)
5. Lewis, D.D.: Feature selection and feature extraction for text categorization. In: Proceedings of the Workshop on Speech and Natural Language, pp. 212–217 (1992)
6. Mao, K.Z.: Orthogonal forward selection and backward elimination algorithms for feature subset selection. IEEE Trans. Syst. Man Cybern. Part B **34**(1), 629–634 (2004)
7. Naghibi, T., Hoffmann, S., Pfister, B.: A semidefinite programming based search strategy for feature selection with mutual information measure. IEEE Trans. Pattern Anal. Mach. Intell. **37**(8), 1529–1541 (2014)
8. Nemenyi, P.: Distribution-free multiple comparisons. Ph.D. thesis, Princeton University (1963)
9. Nguyen, X.V., Chan, J., Romano, S., Bailey, J.: Effective global approaches for mutual information based feature selection. In: ACM SIGKDD, pp. 512–521 (2014)
10. 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. **8**, 1226–1238 (2005)
11. Saidi, R., Bouaguel, W., Essoussi, N.: Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. In: Hassanien, A.E. (ed.) Machine Learning Paradigms: Theory and Application. SCI, vol. 801, pp. 3–24. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-02357-7_1
12. Senawi, A., Wei, H.L., Billings, S.A.: A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. Pattern Recogn. **67**, 47–61 (2017)
13. Sharmin, S., Aktar, F., Ali, A.A., Khan, M.A.H., Shoyaib, M.: BFSp: a feature selection method for bug severity classification. In: R10-HTC, pp. 750–754 (2017)
14. Sharmin, S., Arefin, M.R., Wadud, M.A., Nower, N., Shoyaib, M.: SAL: an effective method for software defect prediction. In: 18th ICCIT, pp. 184–189 (2015)
15. Sharmin, S., Ali, A.A., Khan, M.A.H., Shoyaib, M.: Feature selection and discretization based on mutual information. In: icIVPR, pp. 1–6. IEEE (2017)
16. Sharmin, S., Shoyaib, M., Ali, A.A., Khan, M.A.H., Chae, O.: Simultaneous feature selection and discretization based on mutual information. Pattern Recogn. **91**, 162–174 (2019)
17. Vinh, N.X., Zhou, S., Chan, J., Bailey, J.: Can high-order dependencies improve mutual information based feature selection? Pattern Recogn. **53**, 46–58 (2016)
18. Wanderley, M.F.B., Gardeux, V., Natowicz, R., de Pádua Braga, A.: GA-KDE-Bayes: an evolutionary wrapper method based on non-parametric density estimation applied to bioinformatics problems. In: 21st ESANN, pp. 155–160 (2013)

19. Yang, H., Moody, J.: Feature selection based on joint mutual information. In: Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis, pp. 22–25. Citeseer (1999)
20. Yuan, G.X., Chang, K.W., Hsieh, C.J., Lin, C.J.: A comparison of optimization methods and software for large-scale L1-regularized linear classification. J. Mach. Learn. Res. **11**, 3183–3234 (2010)