Turk J Elec Eng & Comp Sci, Vol.18, No.6, 2010, © TÜBİTAK
doi:10.3906/elk-0906-75

Using covariates for improving the minimum Redundancy Maximum Relevance feature selection method

Olcay KURŞUN[1], C. Okan ŞAKAR[2], Oleg FAVOROV[3], Nizamettin AYDIN[4], Fikret GÜRGEN[5]
[1]*Department of Computer Engineering, İstanbul University, İstanbul-TURKEY*
[2]*Department of Computer Engineering, Bahcesehir University, İstanbul-TURKEY*
[3]*Department of Biomedical Engineering, University of North Carolina, Chapel Hill, USA*
[4]*Department of Computer Engineering, Yıldız Technical University, İstanbul-TURKEY*
[5]*Department of Computer Engineering, Bogazici University, İstanbul-TURKEY*
e-mail: olcay.kursun@istanbul.edu.tr; okan.sakar@bahcesehir.edu.tr; favorov@bme.unc.edu; naydin@yildiz.edu.tr; gurgen@boun.edu.tr

Abstract

Maximizing the joint dependency with a minimum size of variables is generally the main task of feature selection. For obtaining a minimal subset, while trying to maximize the joint dependency with the target variable, the redundancy among selected variables must be reduced to a minimum. In this paper, we propose a method based on recently popular minimum Redundancy-Maximum Relevance (mRMR) criterion. The experimental results show that instead of feeding the features themselves into mRMR, feeding the covariates improves the feature selection capability and provides more expressive variable subsets.

Key Words: *Mutual information; mRMR; unsupervised learning; support vector machines; SINBAD covariates.*

1. Introduction

Feature selection is one of the most crucial steps of many pattern recognition and artificial intelligence problems [1]. In this paper, we build on mutual information, a measure of relevance/dependence, which is used in filter methods with the aim of measuring the relevance levels of the features with the target variable. Mutual Information [2] is a classical measure of dependence which has recently been used for feature selection and ranking as a filter (sorting the variables from most relevant to the least) in several studies in many fields—medicine, neuroscience, genomics and related fields, ecology, economics, etc [3, 4, 5]. Peng et al.'s study [5] is based on an approach called mRMR (*minimum Redundancy-Maximum Relevance*), which aims at obtaining

maximum classification/prediction performance with a minimal subset of variables by reducing the redundancies among the selected variables to a minimum.

In this paper, we will show that using mRMR can lead to inaccurate orderings of the variables since it does not deal with the *type* of the dependency, but only with the quantity of dependency. Instead, we propose a method called CmRMR (*Covariate minimum Redundancy-Maximum Relevance*) which explores and uses the correlated functions (covariates) between variables to compute the unique information about the target variable that the variables possess.

The remainder of this paper is organized as follows: Section 2 reviews the mRMR approach. Section 3 presents the method we propose, which explores the correlated functions between the variables and uses them to compute the redundancy term instead of using themselves (their exact values) directly. Section 4 presents the experimental results on two bioinformatics datasets, Parkinson's and Arrhythmia, selected from UCI machine-learning archive [6] and on a real environmental engineering dataset (next-day ozone-level prediction) [7]. We present discussions and conclusions; and report future work in Section 5.

2. The minimum Redundancy-Maximum Relevance approach

The *minimum Redundancy-Maximum Relevance* (mRMR) approach [5] is based on recognizing that the combinations of individually good variables do not necessarily lead to good classification/prediction performance. In other words, to maximize the joint dependency of top ranking variables on the target variable, the redundancy among them must be reduced, which suggests incrementally selecting the maximally relevant variables while avoiding the redundant ones. Firstly, the mutual information (MI) between the candidate variable and the target variable is calculated (the relevance term). Then the average MI between the candidate variable and the variables that are already selected is computed (the redundancy term). The entropy-based mRMR score (higher it is for a feature, more that feature is needed) is obtained by subtracting the redundancy from relevance. According to mRMR approach, m^{th} feature chosen for inclusion in the set of selected variables S must satisfy the condition

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; T) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right], \tag{1}$$

where X is the whole set of features; T is the target variable; x_i is the i^{th} feature; and I is the mutual information.

Although it has been showed that mRMR algorithm works well for some experimental studies, it is known that it causes inaccurate orderings in some cases since it only measures the quantity of redundancy between the candidate variables and the selected variables but does not deal with the type of this redundancy [8]; also Peng, et al. [5] themselves apply the backward elimination wrapper technique after feature selection step by mRMR to get rid of these ineffective variables. Specifically, it chooses some irrelevant variables too early and some useful variables too late. Because, candidate variable that seems highly redundant with the already selected variables might carry unique information about target variable.

3. The proposed method: CmRMR

To overcome the problem with the mRMR approach, the functions that represent the existing relations between the variables and the target T must be discovered and the mutual information scores must be calculated regarding these functions. For this task, an unsupervised machine learning tool, SINBAD, suggested by Ryder and Favorov [9–12] is used as a basic computational strategy for finding the functions of dependencies between variables.

3.1. SINBAD method for finding correlated functions

To illustrate the goal of SINBAD [9–12], a continuous XOR problem is given below, in which our task is to identify the variables that have statistical dependence with T. Suppose that in reality T is determined by the relation

$$T = X_1 cXORX_2 = X_1 + X_2 - 2 \cdot X_1 \cdot X_2, \tag{2}$$

and assume the presence of variables X_1, X_2, X_3, X_4, X_5, and so on, to search for dependent relation with T. Suppose that the variables are real valued and uniformly distributed in the range from zero to one. The plot of X_1 versus T is shown in Figure 1, which seems to (and should) have some order. However, when we use X_1 to predict T, using it as input to a support vector machine (SVM) [12], or any sort of artificial neural network (ANN) [9, 11, 13], it will result in zero correlation of the learnt output with the desired (target) output T (as a matter of fact, that would successfully minimize the mean squared error by learning to produce T=0.5 at all times as the solution). The correlations shown in Figure 1 even indicate, by chance, that the magnitude of the correlation can be more between the independent variables than the dependent ones. So why categorize X_1 (but not X_3, X_4, or X_5) as to have some dependency with T? In fact, the prediction accuracy of T in terms of mean squared error or correlation coefficient is not really better when we use X_1 as input in comparison to when we use, say, X_3.

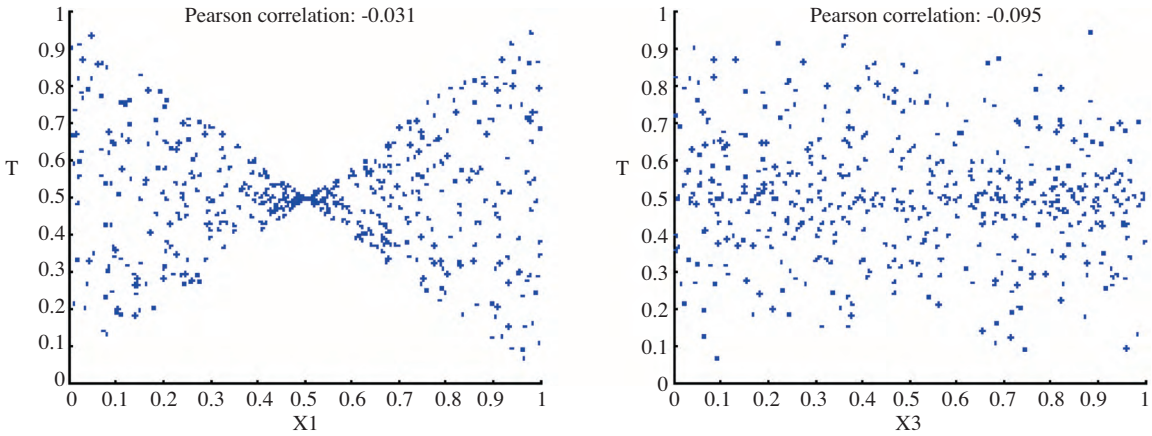

Figure 1. Left Panel: The plot of X_1 versus T (not correlated but actually dependent case). Right Panel: The plot of X_3 versus T (independent case).

As seen in Figure 1, when X_1 is around 0.5, then T is also around 0.5. It is true that, when X_1 is close to zero or close to one, the value of T is unpredictable (left and right margins of the plot indicates for X_1 is

zero, T can take any values from zero to one, of course depending on the value of X_2 as implied by equation 2). Nevertheless, X_1 has some predictive power of T; more precisely, a function of X_1 has high correlation with a function of T. In fact, combined with X_2, it can predict T perfectly but without X_2 it can only do some prediction of a function of T, but not the actual value of it.

SINBAD can conclude that X_1 has more relation to T than, for example, X_3. We could ask the neural network or SVM to learn not necessarily T but a nonlinear function of T. In fact, the correlation between $(X_1-0.5)^2$ and $(T-0.5)^2$ is around 0.6. This is just a simple ad-hoc function we came up with based on the nature of the relation between X_1 and T; that is, when X_1 is around 0.5 then T is around 0.5. The correlation is striking. We can really learn to predict a nonlinear function of T from X_1. Therefore, unlike in the case of supervised setting, in which a function of X_1 is forced to approximate T, the prediction accuracy in the unsupervised setting (when the training signal T is left free) will be more. The next question to be answered is how to decide what this nonlinear function is. Two neural nets (or SVMs or any other machine learning tools) can be set up to teach each other what is common to their different inputs (Figure 2). One neural net (ANN) can take variable X_1 as input and the other ANN can take variable T as input and they can teach each other until they converge to a common function, h, that both can compute well enough, such as

$$f_1(X_1) = (X_1 - 0.5)^2 \cong h \cong (T - 0.5)^2 = g_1(T). \tag{3}$$

$f_1(X_1)$ and $g_1(T)$ are called correlated functions (Figure 3), also known as covariates in statistics [14]. Our proposal boils down to using unsupervised learning for finding dependencies based on SINBAD method. Note that mutual information [2] can, too, detect that X_1 and T has more in common but mutual information calculation does not give us learnt functions f and g as in Equation 3, it only gives an entropy-based score, which is hard to interpret (unlike easily interpretable Pearson correlation coefficient that is between -1 and 1). The SINBAD algorithm is outlined in Algorithm, where two SVMs train each other iteratively until convergence. In our experiments we used 5 iterations of the repeat-until loop shown.

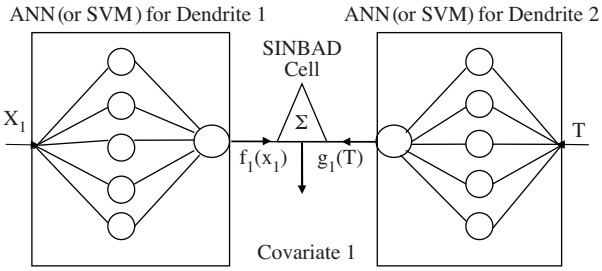

Figure 2. The SINBAD model with two dendrites modeled using artificial neural networks (ANN) or alternatively support vector machines (SVMs).

In its implementation of SINBAD in this work, the learnt functions f_i and g_i are kernel functions as we use kernel-SVMs for SINBAD dendrites. To acquire nonlinear dependences, RBF-kernel is the most recommended. The correlated functions (covariates) learned from a feature X_i and the target variable T are in the following form that resembles kernel canonical correlation analysis (KCCA) [9]:

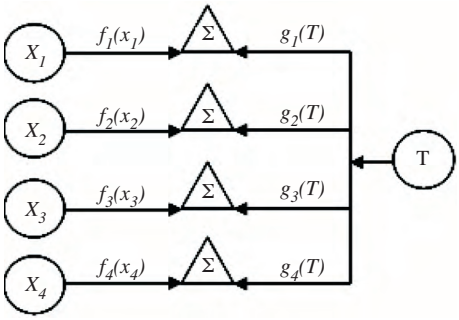

Figure 3. The functional relations among the variables.

$$f_i(X_i = x) = a_i + \sum_{j=1}^{n} b_{ij} K_1(x_i^j, x)$$

$$g_i(T = t) = c_i + \sum_{j=1}^{n} d_{ij} K_2(t^j, t)$$

where n is the number of training samples, K_1 and K_2 are kernel functions, b_{ij} and d_{ij} are Lagrange multipliers (j=1,2,...,n), and a_i and c_i are bias terms learnt by SVM in the mutual prediction of f_i and g_i.

SINBAD Algorithm for Learning Correlated Functions

Input: Observations of *X* and *T* variables
Output: *f*(*X*) and *g*(*T*) that are maximally correlated

Randomly initialize ***g*(*T*)** with μ = 0, σ = 1
Repeat
Train SVM$_1$ on *X* to approximate *g*(*T*)
f(*X*) ← Outputs learnt by *SVM*$_1$
f(*X*) ← Normalize *f*(*X*) to μ = 0, σ = 1
Train SVM$_2$ on *T* to approximate *f*(*X*)
g(*T*) ← Outputs learnt by *SVM*$_2$
g(*T*) ← Normalize *g*(*T*) to μ = 0, σ = 1
Until convergence or #iterations reached the limit

Algorithm. Training algorithm of a SINBAD cell.

3.2. CmRMR formulation

In our method that we called CmRMR (*Covariate minimum Redundancy-Maximum Relevance*), while computing the redundancy the mutual information among the correlated functions, $f_i(X_i)$, of features explored by SINBAD will be used instead of using the features, X_i, themselves as in mRMR algorithm. The functions $f_i(X_i)$ correspond to the maximal information of X_i about T(or a function of T). Thus, the irrelevant information of X_i 's about T is filtered out. The mutual information between $f_i(X_i)$ and $f_j(X_j)$ now gives the

relevant redundancy about T (i.e. the commonality in what can be inferred about T *by* X_i and X_j). Thus, the variable that will be selected next among the candidate variables must satisfy the condition

$$\max_{x_j \in X - S_{m-1}} \left[I(f_j(x_j); T) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I\left(f_j(x_j); f_i(x_i)\right) \right], \tag{4}$$

where $f_i(x_i)$ is the correlated function of variable x_i. In equation 4, rather than $I(x_j; T)$, we used $I(f_j(x_j); T)$ as the relevance in order to preserve the commonality of the $f_j(x_j)$ term with redundancy.

4. Experimental results

In the experiments given below, all the linear-valued features are normalized to zero-mean and unit-variance, and for computing the mutual information scores we simply discretized them to 9 discrete levels, similar to [5]. That is, the mean μ and the standard deviation σ of each variable is used such that the feature values between μ-$\sigma/2$ and $\mu+\sigma/2$ are converted to 0. The 4 intervals of size σ to the right of $\mu+\sigma/2$ are converted to discrete levels from 1 to 4 and the 4 intervals of size σ to the left of μ-$\sigma/2$ are mapped to discrete levels from -1 to -4. Very large positive or negative feature values are truncated and discretized to ± 4 appropriately. We used LIBSVM package [15] for the implementation of SVMs of SINBAD, and also for comparing the accuracy of selected variables of mRMR and CmRMR.

4.1. Parkinson dataset

4.1.1. Dataset description

Parkinson Dataset (PD) [16], which is available online at UCI machine learning archive [6], consists of 22 real-valued speech-features and a binary PD-score of a total of 195 speech recordings of 32 individuals (24 of which are with PD and 8 are healthy with six or seven recordings per subject). The features of PD are linear valued; a PD-score of 1 indicates that the feature vector belongs to a person with PD and a score of 0 indicates that it belongs to a healthy subject. The features in the dataset are diverse. Some of them are traditional measures based on the application of the short-time autocorrelation to successive segments of the signal, some are non-standard measures based on nonlinear dynamical systems theory. The labels and short explanations for the measurements along with some basic statistics of the original dataset are given in Table 1.

4.1.2. Feature ranking on parkinson dataset

Top 10 features of PD dataset according to their mRMR and CmRMR scores are shown in Table 2. In order to compare the selected 10 features of mRMR and CmRMR, we fed the top 10 features of both methods to SVMs. Firstly, we split the PD dataset into training and test set by randomly bipartitioning the 195 samples. For statistical significance, we repeated this process 100 times. As it is seen in Table 3, the average classification accuracies of mRMR and CmRMR selected features are 87.46 and 86.89, respectively.

Table 1. Description of the features of the parkinson dataset.

Description	Feature Label	Min.	Max.	Mean	Std. Dev.
Average vocal fundamental freq.	MDVP:F0(Hz)	88.33	260.11	154.23	41.39
Max. vocal fundamental freq.	MDVP:Fhi(Hz)	102.15	592.03	197.11	91.50
Min. vocal fundamental freq. frequency	MDVP:Flo(Hz)	65.48	239.17	116.33	43.52
Several measures of variation in fundamental frequency	MDVP:Jitter(%)	0.002	0.033	0.006	0.005
	MDVP:Jitter(Abs)	7E-06	26E-05	4.4E-05	3.48E-05
	MDVP:RAP	0.001	0.021	0.003	0.003
	MDVP:PPQ	0.001	0.020	0.003	0.003
	Jitter:DDP	0.002	0.064	0.010	0.009
Several measures of variation in amplitude	MDVP:Shimmer	0.01	0.119	0.03	0.019
	MDVP:Shimmer(dB)	0.085	1.302	0.282	0.195
	Shimmer:APQ3	0.005	0.056	0.016	0.010
	Shimmer:APQ5	0.006	0.079	0.018	0.012
	MDVP:APQ	0.007	0.138	0.024	0.017
	Shimmer:DDA	0.014	0.169	0.047	0.030
Two measures of ratio of noise to tonal components in the voice	NHR	0.001	0.315	0.025	0.040
	HNR	8.441	33.047	21.886	4.426
Two nonlinear dynamical complexity measures	RPDE	0.257	0.685	0.499	0.104
	D2	1.423	3.671	2.382	0.383
Signal fractal scaling exponent	DFA	0.574	0.825	0.718	0.055
Three nonlinear measures of fundamental frequency variation (Last one, PPE, is the proposed measurement of dysphonia)	Spread1	-7.965	-2.434	-5.684	1.090
	Spread2	0.006	0.450	0.227	0.083
	PPE	0.045	0.527	0.207	0.090

Table 2. Top 10 features of mRMR and CmRMR rankings.

Ranking	mRMR	CmRMR
1	spread1	spread1
2	MDVP:Fo(Hz)	spread2
3	Shimmer:APQ3	MDVP:Fhi(Hz)
4	D2	MDVP:Flo(Hz)
5	DFA	D2
6	MDVP:Flo(Hz)	MDVP:Fo(Hz)
7	spread2	MDVP:Jitter(Abs)
8	MDVP:RAP	MDVP:RAP
9	RPDE	Jitter:DDP
10	MDVP:Fhi(Hz)	PPE

Table 3. Classification rates using top 10 features of mRMR and CmRMR rankings with SVMs.

	100 random splits of samples	100 random splits of individuals
mRMR-10	87.46 ± 3.42	76.76 ± 6.67
CmRMR-10	86.89 ± 3.08	80.61 ± 6.61

As the PD dataset has 6 (but 7 for three of the individuals) recordings per individual, bipartitioning the samples results in including the same individual's different samples both in the training and test sets. It

is the most likely case that the training set contains samples from all the individuals; and therefore, the test set greatly overlaps with the training set, which creates artificially high, biased prediction accuracy [17]. In order to remove this bias, we split the PD dataset into training and test set by randomly bipartitioning the 32 individuals (not the 195 samples). Thus, all the samples of an individual exist only in either the training or the test set. We repeated this 100 times as well (Table 3). It is seen that the average classification accuracy of the SVMs with mRMR and CmRMR top 10 features is 76.76 and 80.61, respectively. CmRMR selected features increases the accuracy nearly 4% in contrast to mRMR selected features when the dataset is split into the training and test sets without the aforementioned bias. This shows that CmRMR selects a more compact and robust feature set than mRMR. For example, pitch period entropy (PPE) is selected in top 10 of CmRMR but not selected of mRMR. In [16], PPE has been proposed as one of the most important features of PD telediagnosis which removes the natural variations in human voice in order to better capture pitch period variation due to PD-related dysphonia.

We also evaluate the "speaker-identification" memorization (over-fitting) risk of the selected feature sets (top 10 selected features) of mRMR and CmRMR by applying the nearest neighbor rule. For this purpose, we calculated the percentage of times, out of 195 samples, when the nearest neighbor is another sample of the same individual (Table 4). The probability of the nearest neighbor being of the same speaker using the mRMR top 10 features is 0.7487 whereas it is 0.60 using the CmRMR top 10 features. Probability of either or both of the 1^{st} and the 2^{nd} Nearest Neighbor being of the same speaker is 0.7846 and 0.6667 with mRMR and CmRMR feature sets, respectively. These results also show that CmRMR feature set has lower memorization risk than mRMR feature set; therefore, it is expectable that mRMR is more likely to be affected by the aforementioned bias of training/test split.

Table 4. Memorization risk estimation of top 10 features of mRMR and CmRMR with nearest neighbor.

	Percentage of 1^{st} Nearest Neighbor being of the same speaker	Percentage of 1^{st} or 2^{nd} Nearest Neighbor being of the same speaker
mRMR	74.87%	78.46%
CmRMR	60.00%	66.67%

4.2. Air pollutants dataset

Ozone (O_3) is an issue of increasing public concern due to its recognized adverse effects on human health. Therefore, accurate O_3 prediction models are very important tools in developing public warning strategies.

4.2.1. Dataset description

The air pollutant parameter measurements used in this study were procured from Istanbul Metropolitan Municipality Environment Protection and Control Office which has 10 automatic air quality measuring stations in Istanbul, Turkey, to observe the air pollution in the atmosphere of Istanbul continuously [7]. These measurements have been observed at 15-minute intervals. The dataset contains the measurements of two of these stations, Kadikoy and Sarachane, from July 2003 to June 2004. The meteorological variables were chosen from

Florya and Goztepe meteorological stations of Government Meteorology Works Office which are the nearest stations to Kadikoy and Sarachane, respectively. The dataset contains 328 samples (one year's worth of data) with 13 linear valued variables (features). The variables, their abbreviations and some statistical properties are given in Table 5.

Table 5. Abbreviations and statistical parameters of the air pollutant dataset.

Variable	Abb.	Minimum	Maximum	Average	Std. Deviation
Ozone (μg/m^3)	O_3	0	86	14.45	10.535
Sulfur Dioxide (μg/m^3)	SO_2	0	82	16.102	13.068
Nitric Oxide (μg/m^3)	NO	3	587	46.551	63.169
Nitrogen Dioxide(μg/m^3)	NO_2	13	158	53.898	24.866
Dust (μg/m^3)	PM	9	191	55.662	29.21
Total Hydrocarbon (μg/m^3)	THC	162	4091	1588.755	419.33
Outdoor Temperature (^{o}C)	OT	-5.3	28.8	13.509	7.686
Wind Speed (m/s)	WS	0.4	7.1	2.537	1.137
Solar Irradiance (Hour)	SI	0	13.2	6.086	4.284
Cloudiness (0 – 10)	C	0	10	4.855	3.431
Pressure (mbar)	P	988.2	1032.1	1012.452	6.514
Relative Humidity (%)	RH	45.7	95.7	73.032	11.058
Rain (mm)	R	0	48.6	1.91	4.994

4.2.2. Feature ranking on air pollutants dataset

Both mRMR and CmRMR are carried out for the sensitivity analysis of the variables according to their importance in ozone prediction (see Table 6).

Start Here Next Ozone concentration has the highest MI with ozone concentration at time $t+1$. Thus, it is chosen in the first order by both mRMR and CmRMR. However, while relative humidity (RH) is selected in the 4^{th} order by mRMR, it is the 10^{th} in the ranking produced by our method, CmRMR. It is well known that RH is not a key factor in the prediction of ozone level [18]. This inaccurate preference for this variable by the mRMR ranking seems to be due to the problem described in Section 3. In other words, as mRMR takes the irrelevant redundancies also into account, the mRMR scores of other variables must have turned out to be less than that of RH by mistake.

Table 6. Importance of input variables in the prediction of ozone concentration.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
mRMR	O_3	THC	OT	RH	NO	WS	SI	SO_2	C	PM	R	NO_2	P
CmRMR	O_3	OT	NO_2	SI	NO	THC	WS	P	SO_2	RH	C	R	PM

In the 4^{th} order, solar irradiance (SI) is ranked by our method. Previous studies in related fields indicate that SI is one of the most important meteorological variables in the prediction of ozone concentration with temperature [18–21]. mRMR ranked in the 7^{th} order, probably because it seems mostly redundant with outdoor temperature (OT). However, SI is a key factor for O_3 forming and accumulation [19], and thus carries important unique information distinct from OT about ozone level that could be successfully captured by our method.

One of the other variables ordered in the same manner with SI is NO_2, which is ranked as 12^{th} order by mRMR; and 3^{th}, i.e. one of the most significant variables, by CmRMR. Ordering of NO_2 as one of the least significant features by mRMR seems arising from the redundancy with NO which is ranked before NO_2 (in the 5^{th} order). However, this is probably irrelevant redundancy; because NO_2 has unique information about ozone concentration, distinct from what NO has. This coincides with the environmental literature stating that NO_2 has a major importance in the prediction of ozone concentration [19, 20, 22].

We also fed the mRMR and CmRMR rankings to SVMs, incrementally adding one feature at a time. We divided the dataset into training and test sets as odd/even days; that is, by taking a day for training and the next day for the test. The results are shown in Table 7. SVMs give higher accuracy with CmRMR ranked features than with mRMR ranked features in all iterations; and the highest correlation of 0.7605, is obtained with top 3 features of CmRMR, which are O_3, OT, and NO_2.

Table 7. SVM prediction correlations of ozone concentration by incrementally adding features one at a time by mRMR and CmRMR.

Feature Ranking	mRMR feature	mRMR accuracy	CmRMR accuracy	CmRMR feature
1	O_3	.7349	.7349	O_3
2	THC	.6981	**.7523**	OT
3	OT	.7386	**.7605**	NO_2
4	RH	.7330	**.7482**	SI
5	NO	.7266	**.7344**	NO
6	WS	.7127	**.7326**	THC
7	SI	.7016	**.7095**	WS
8	SO_2	.6858	**.7214**	P
9	C	.6718	**.7059**	SO_2
10	PM	.6697	**.7008**	RH
11	R	.6753	**.6859**	C
12	NO_2	.6783	**.6869**	R
13	P	.6837	.6837	PM

4.3. Arrhythmia dataset

4.3.1. Dataset description

Arrhythmia dataset is available on the UCI machine learning archive [6, 23]. Arrhythmias are disorders of the regular rhythmic beating of the heart. The aim is to classify the sample in one of the 16 groups of arrhythmia of which class 1 means 'normal', classes 2 to 15 refer to different classes of arrhythmia, and class 16 refers to one of the unclassified arrhythmia types [23]. The dataset contains 452 samples with 279 attributes, 206 of which are linear valued (the other 73 attributes/features are binary).

4.3.2. Feature ranking on arrhythmia dataset

We ranked the features of arrhythmia with mRMR and CmRMR, and fed the rankings to SVMs. For each variable, 10 iterations are done while exploring the correlated functions. We used 10-fold cross validation with

default parameters of LIBSVM [15]. The classification accuracies with mRMR and CmRMR ranked features are shown in Table 8.

Table 8. SVMs classification accuracies on arrhythmia dataset with various number of features.

Feature #	mRMR accuracy	CmRMR accuracy
5	69.03	**70.58**
10	69.47	**70.35**
15	69.25	**70.35**
20	69.25	**70.80**
25	69.91	**72.12**
30	69.91	**72.12**
35	70.80	**71.68**
40	**71.90**	71.24
45	**71.24**	71.02
50	**71.68**	71.24

The results in Table 8 show that CmRMR starts selecting better features than mRMR. The highest accuracy is obtained using 25 top features of CmRMR with 72.12%. It seems that mRMR adds some irrelevant features at the beginning, but as we add more and more features, eventually the good features are added and gives comparable results with CmRMR.

Table 9. Run times for exploring the correlated functions with SINBAD.

	Feature #	Samples #	Average time for each feature (sec)	Total Time (sec)
Parkinson Dataset	22	195	0.01 ± 0.01	0.40
Air Pollutants Dataset	13	328	0.54 ± 0.07	7.01
Arrhythmia Dataset	279	452	0.38 ± 0.29	106.02

5. Discussions, conclusions, and future work

There is a vast collection of research on improving the use of mutual information (MI) as a filter in feature selection methods. One of the most successful studies is by Peng et al. [5] called mRMR (*minimum Redundancy – Maximum Relevance*) approach and is based on choosing a subset that aims at minimizing the pairwise redundancies in the set among the selected variables while maximizing the overall relevance with the target variable. It is true that such redundancies (variables with nearly the same information content about the target variable) must be avoided in order to obtain a minimal subset that maximizes the joint inferential dependency with the target variable.

In its iterative implementation, candidate variables are ranked based on the difference of their relevance (mutual information) with the target and their redundancy with the already selected variables. As for the redundancy term of a candidate variable, the mRMR approach computes plainly its mutual information with the already selected variables. It does not consider whether that redundancy is related to the target variable or not. However, as a more effective redundancy term, we propose to deal with the part of the redundancy between

the correlated functions (with the target variable) of the candidate and the selected variables. Computing the redundancies between the correlated functions quantifies the unique information that a candidate variables possesses about the target (i.e., unique in the sense that different contribution from what is already learnable from the selected variables). We propose to first find the covariates (correlated functions) between each one of the candidate variables and the target variable and then we use these functions while computing the relevance and redundancies among the variables. Experimental results on three real datasets (Parkinson's disease telediagnosis, Arrhythmia classification, and ozone level prediction) also compare our method favorably to mRMR.

Exploring the correlated functions (covariates) among the features and the target variable in order to calculate the CmRMR scores brings an extra time complexity for feature selection. It involves training and testing SVMs for a number of iterations. This preprocessing stage is needed by CmRMR, which simply passes the correlated functions to the plain mRMR (instead of passing the features directly). However, the added time complexity of this processing is negligible as shown in Table 9 for the experimental studies of Section 5.

As a future direction, more than just a single pair of correlated functions can be extracted possibly using higher-speed neural implementations (e.g. the optoelectronic implementation [13]) for the SINBAD method, in order to better represent the relations between features and the target variable.

Acknowledgements

The work of C. O. Sakar was supported by the Ph.D. scholarship (2211) from Turkish Scientific Technical Research Council (TÜBİTAK). He is a Ph.D. student in the Computer Engineering Department at Bogazici University, Istanbul, Turkey.

We would like to thank Dr. Goksel Demir, from the Environmental Engineering department at Bahcesehir University for providing the air-pollutants database.

References

[1] J. G. Zhang, & H.W. Deng, "Gene selection for classification of microarray data based on the Bayes error", BMC Bioinformatics, Vol. 8, pp. 370, 2007.

[2] C. E. Shannon, "A mathematical theory of communication", Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656, 1948.

[3] C. Ding & H. Peng, "Minimum redundancy feature selection from microarray gene expression data", Journal of Bioinformatics and Computational Biology, Vol. 3(2), pp. 185-205, 2005.

[4] N. Kwak & C. H. Choi, "Input feature selection by mutual information based on Parzen Window", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24(12), pp. 1667-1671, 2002.

[5] H. Peng, F. Long, F., C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27(8), pp. 1226-1238, 2005.

[6] UCI Learning Repository, http://archive.ics.uci.edu/ml/, June 2008.

[7] H. Ozdemir, G. Demir, G. Altay, S. Albayrak, C. Bayat, "Prediction of tropospheric ozone concentration in Istanbul by employing artificial neural network", Environmental Engineering Science, Vol. 25(9), 2008.

[8] P. A. Estévez, M. Tesmer, C. A. Perez, J. M. Zurada, "Normalized mutual information feature selection", IEEE Transactions on Neural Networks, Vol. 20(2), 2009.

[9] O. Favorov & D. Ryder, "SINBAD: a neocortical mechanism for discovering environmental variables and regularities hidden in sensory input", Biological Cybernetics, Vol. 90, pp. 191-202, 2004.

[10] O. Kursun & O. Favorov, "Feature selection and extraction using an unsupervised biologically-suggested approximation to Gebelein's maximal correlation", International Journal of Pattern Recognition and Artificial Intelligence, 2010 (to appear).

[11] O. Kursun & O. Favorov, "SINBAD automation of scientific discovery: From factor analysis to theory synthesis", Natural Computing, Vol. 3(2), pp. 207-233, 2004.

[12] O. Kursun & O. Favorov, "What can SVMS teach each other?", Artificial Neural Networks in Engineering (ANNIE 2004), St Louis, MO, USA, 2004.

[13] A. Bal, "Widrow-cellular neural network and optoelectronic implementation", Optik, Vol. 115, pp. 295-300, 2004.

[14] C. Fyfe, Hebbian Learning and Negative Feedback Networks, Springer, 2005.

[15] C. W. Hsu & C. J. Lin, "A comparison of methods for multi-class support vector machines", IEEE Trans. Neural Networks, Vol. 13, pp. 415-425, 2002 (LIBSVM software available for downloading at http://www.csie.ntu.edu.tw/∼cjlin/libsvm).

[16] M. A. Little, P. E. McSharry, E. J. Hunter, L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", IEEE Transactions on Biomedical Engineering, Vol. 56(2), pp. 1015-1022, 2009.

[17] O. Sakar & O. Kursun, "Telediagnosis of Parkinson's disease using measurements of dysphonia", Journal of Medical Systems, 2009 (Online published).

[18] A. C. Comrie, "Comparing neural networks and regression models for ozone forecasting", J. Air Waste Manage, Vol. 47(6), pp. 653-663, 1997.

[19] D. Wang, W. Z. Lu, "Interval estimation of urban ozone level and selection of influential factors by employing automatic relevance determination model", Chemosphere, Vol. 62, pp. 1600-1611, 2006.

[20] J. Gomes-Sanchis, J. D. Martin-Guerrero, E. Soria-Olivas, J. Villa-Frances, J. L. Carrasco, S. D. Valle-Tascon, "Neural networks for analyzing the relevance of input variables in the prediction of tropospheric ozone concentration", Atmospheric Environment, Vol. 40, pp. 6173-6180, 2006.

[21] O.P. Barcenas, E. Soria-Olivas, J. D. Martin-Guerrero, G. Camps-Valls, J. L. Carrasco-Rodriguez, S. del Valle-Tascon, "Unbiased sensitivity analysis and pruning techniques in neural networks for surface ozone modeling", Ecological Modelling, Vol. 182, pp. 149–158, 2005.

[22] S. A. Abdul-Wahab, S. M. Al-Alawi, "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks", Environmental Modelling & Software, Vol. 17, pp. 219-228, 2002.

[23] H. A. Guvenir, B. Acar, G. Demiroz, A. Cekin, "A supervised machine learning algorithm for Arrhythmia analysis", Proceedings of the Computers in Cardiology Conference, Lund, Sweden, 1997.