Maximum Relevancy Maximum Complementary feature selection for multi-sensor activity recognition

Saisakul Chernbumroong[a], Shuang Cang[b], Hongnian Yu[a]

[a] *Faculty of Science and Technology, Bournemouth University, Poole, Dorset, BH12 5BB, UK*
[b] *School of Tourism, Bournemouth University, Poole, Dorset, BH12 5BB, UK*

Abstract

In the multi-sensor activity recognition domain, the input space is often large and contains irrelevant and overlapped features. It is important to perform feature selection in order to select the smallest number of features which can describe the outputs. This paper proposes a new feature selection algorithms using the maximal relevance and maximal complementary criteria (MRMC) based on neural networks. Unlike other feature selection algorithms that are based on relevance and redundancy measurements, the idea of how a feature complements to the already selected features is utilized. The proposed algorithm is evaluated on two well-defined problems and five real world data sets. The data sets cover different types of data i.e. real, integer and category and sizes i.e. small to large set of features. The experimental results show that the MRMC can select a smaller number of features while achieving good results. The proposed algorithm can be applied to any type of data, and demonstrate great potential for the data set with a large number of features.

Keywords: Feature selection, Neural networks, Mutual information, Activity recognition.

1. Introduction

The aim of feature selection is to identify the smallest subset of input features which explains the output classes. This process is important especially to the classification problems with a large number of input features. For example, a multi-sensor activity classification system normally contains a large number of input features generated from different sensors. Feature selection can help reduce the size of feature space which leads to reduction in computational cost and complexity in the classification system. In real world problems where input features contain irrelevant and redundant features, feature selection can help identify a relevance feature set which leads to improvement in classification performances.

There are three main approaches in feature selection found in wearable sensor-based activity recognition applications: intuition, filter, and wrapper. Intuition based feature selection requires a domain knowledge or understanding which is required in the classification of the interested activities. This approach is often used in conjunction with visual inspection, statistical analysis of the features e.g. histogram, distribution graph, or observation made during activity occurrence [1] and [2]. Filter based-feature selection measures the relevance between features and the outputs by using techniques such as information theory, distance, correlation, receive operating curve (ROC), etc. Each feature is evaluated for its relevance then given a ranking score. For example, features which have the best performance in discriminating the interested activities were selected using ROC [3, 4]. Many of the statistical tests are used with this approach e.g. chi-square, T-test, etc. The study

in [4] found that features selected from the ranking quality group technique based on discrimination and robustness, ROC, T-test or the wilcoxon with support vector machine produced remarkable results. Mutual information (MI) is another popular measurement used for measuring the relationship between two variables. Feature selection techniques which use MI are such as maximum relevance minimum redundancy [5], normalized mutual information feature selection-feature space [6], feature selection based on cumulate conditional mutual information [7], etc. Some techniques are based on neural networks to rank the features e.g. neural network feature selection [8], clamping technique [9], constructive approach for feature selection [10], etc. The main advantages of the filter approach are due to its simplicity, speed and independence of the classification algorithm [11]. However, most of the techniques in this approach usually consider two variables i.e. a feature and class output, thus ignoring dependencies among a set of features. This may lead to a selection of redundant features resulting in low classification accuracy. In some techniques such as MRMR [5] and NMIFS [12], another criteria i.e. redundancy is used to reduce the chance of selecting redundant features.

Wrapper based-feature selection is the most popular technique in wearable sensor-based activity recognition. In this technique, various set of feature subsets are generated and evaluated using a classification algorithm. The most optimum feature subset is selected using search techniques. Examples of this approach are forward selection [13, 14, 15], backward selection, forward-backward selection [16], exhaustive search [17], etc. In forward selection, one feature is added into a feature subset each time and the subset is evaluated for

its performance. On the other hand, backward selection removes one feature from the feature subset each time and evaluates the subset performance. Forward-backward selection employs both directions where forward selection is carried out first then the subset is refined using backward selection. This approach is computationally more extensive than the filter method, however it can provide a better result as it takes into account the features dependency and interaction with the classification algorithm. Some studies combine both filter and wrapper methods. For example, the study in [18] combined the features filtered by information gain and F-score, then used the wrapper method to improve classification accuracy.

The most feature selection methods in the current literature are based on two criteria i.e. relevancy - how the feature is relevant to outputs, and redundancy - to reduce the chance of selecting redundant features. However, feature selection using these two criteria does not consider how a feature will complement the already selected features. This may result in selecting a larger number of feature than actually required. Also, in some feature selection techniques which only consider the relevancy criteria, redundant features may be selected. For techniques which use the wrapper approach, considering all possible feature subsets suffers a high computational cost.

Considering the above limitations, we propose a new feature selection algorithm with a new criterion i.e. complementary - how a feature complements the already selected features. In addition, based on our knowledge, this criterion has not yet been considered in any other feature selection algorithm. The clamping technique is employed to measure the feature relevance. We introduce a new measurement to calculate the complementary value of

the feature to the already selected feature set. The feature is selected based on the criteria of maximum relevance and maximum complementary. The main difference between the proposed technique and the other algorithms are that the complementary measurement is used instead of the redundancy measurement. Feature redundancy can be detected through the complementary measurement such that the redundant feature should give a low complementary score.

The paper is organized as follows: Section 2 presents some popular feature selection algorithms which are used for comparison in this study. Section 3 presents the proposed feature selection technique in detail. We evaluate our algorithm using two well-defined problems and four benchmark data sets and one multi-sensor activity recognition data set collected from a real home. The experimental results are presented in Section 4. Finally, the discussion and conclusion are presented in Section 5.

2. Related works

Many techniques have been proposed for feature selections as discussed in the previous section. In this paper, we look at two different approaches used for feature ranking i.e. mutual information (MI) and neural networks (NN).

2.1. Mutual information based feature selection

MI, which is based on information theory [19], measures the dependency between two variables. The MI value is zero if and only if the variables are independent. Given continuous variables f_i and f_j, the MI is:

$$MI(f_i; f_j) = \int \int p(f_i, f_j) \log \frac{p(f_i, f_j)}{p(f_i)p(f_j)} \mathrm{d}f_i \mathrm{d}f_j$$

In practice, it is difficult to calculate MI of the continuous values and often the variables are discretized using bins. The MI of discrete variables is:

$$MI(f_i; f_j) = \sum_c \sum_j p(f_i, f_j) \log \frac{p(f_i, f_j)}{p(f_i)p(f_j)}$$

There are many feature ranking algorithms based on the MI [5, 6, 7, 12]. The maximal relevant minimal redundant (MRMR) is one of the most popular feature selection algorithms. Many algorithms have been based on MRMR. For example, the normalized mutual information feature Selection (NMIFS) which enhance MRMR by using entropy of the variables to normalize the MI values when calculating the redundancy between variables. MRMR is enhanced by using the kernel canonical correlation analysis as inputs rather than the actual features [20].

In this study we investigate the commonly used feature selection algorithms based on MI which are MRMR and NMIFS algorithms.

2.1.1. MRMR

The MRMR algorithm [5] ranks the features based on the minimal redundancy and maximal relevance criterion. It calculates the MI between two features to measure the redundancy and the MI between a feature and the outputs to measure the relevance. Using the MRMR concept and greedy selection, a set of feature rankings S can be obtained as follow:

(A) Given $S = \{\}$ where S is a set of selected features and $F = \{f_1, f_2, f_i, f_j ..., f_N\}$ where F is a set of N features, select the feature f_s in F which has

the maximum mutual information between itself and output C where $C = \{c_1, c_3, ..., c_K\}$ and $f_s = \max_{f_i \in F} MI(f_i; C)$, and update S and F.

$$S = S \cup \{f_s\} \tag{1}$$

$$F = F \setminus \{f_s\} \tag{2}$$

(B) Select feature f_s in F which satisfies the following condition:

$$f_s = \max_{f_i \in F} \{MI(f_i; C) - \frac{1}{|s|} \sum_{f_j \in S} MI(f_i; f_j)\}$$

Update S and F using (1) and (2).

Repeat Step (B) until the desired number of features is obtained.

2.1.2. NMIFS

The NMIFS algorithm [12] is an enhancement of the MRMR algorithm. A normalized MI (NMI) between two features are used instead:

$$NMI(i; j) = \frac{MI(i; j)}{\min\{H(i), H(j)\}}$$

where $H()$ is the entropy function. Similar steps as MRMR are carried out, however the condition in Step (B) is changed to:

$$f_s = \max_{f_i \in F} \{MI(f_i; C) - \frac{1}{|s|} \sum_{f_j \in S} NMI(f_i; f_j)\}$$

2.2. Neural network based feature selection

Some studies have proposed to use NN for feature selection [8, 9, 10]. For example, the neural network feature selector (NNFS) [8] selects features based on weights associated with that features. The weights associated with unimportant features would have values close to zero. NNFS adds a penalty term to the cross-entropy error function in order to distinguish redundant network connection. The technique proposed in [21] trains the network by minimizing the cross-entropy error function augmented with additional terms to constraint the derivatives. The features are selected based on the reaction of the validation classification error as a result of the removal of individual features. The algorithm is tested on three real-world problems and the results indicate that it outperforms the techniques such as NNFS, the fuzzy entropy, the discriminant analysis, the neural network output sensitivity based feature saliency measure, and the weights-based feature saliency measure. The clamping technique proposed in [9] is designed to determine the importance of the feature by observing the network performance when each feature is clamped. In this study we compare the performance of the proposed algorithm with the clamping algorithm.

The idea of the clamping algorithm is that if the feature is important by clamping the feature to a certain value i.e. mean value, the generalized classification accuracy will be affected. Firstly, train the network using all features on training data set and obtain the generalized performance ($g(F)$) by calculating the classification accuracy on test data set. The generalized performance of clamped feature $g(F|f_i = \bar{f}_i)$ is obtained by calculating the classification accuracy of test data set where feature f_i is clamped to its mean

value. The importance of feature f_i is calculated as:

$$Im(f_i) = 1 - \frac{g(F|f_i = \bar{f}_i)}{g(F)} \tag{3}$$

The features are then ranked based on their importance in descending order.

The clamping technique provides robust ranking even in noisy data. However, it only considers the relationship between one feature and the classes. It does not consider any relationship between the features. MRMR and NMIFS do consider the relationship between features. However, the relationship between only two features are measured. None of these three techniques considers how a feature would complement to the already selected features. In this paper, we propose a new feature selection technique which consider the relationship between features and the class as well as the relationship among a group of features. The details of the proposed algorithm are presented in the next section.

3. Methodology

The proposed feature selection method is based on the criteria of maximum relevance and maximum complementary (MRMC) of the feature. In our method, NN is employed for the calculation of the relevance and complementary score. NN is based on the concept of connectionism where several input nodes are connected with associated weights to several outputs nodes. We use a network with one hidden layer which is called Multi-layer perceptron (MLP). Given input of N features $F = \{f_1, f_2, ..., f_i, ..., f_N\}$, predict output of K classes $C = \{c_1, c_2, ..., c_K\}$. Fig. 1 depicts the neural network

architecture where b_1 is the bias and weights $W = \{w_{11}, w_{12}, ..., w_{Nj}\}$ where w_{11} represents a weight connect from f_1 to hidden node 1. The weights and bias are generated randomly from a univariate distribution. The network output node $\hat{y}_i$ can be calculated from the summation function [22]:

$$\hat{y}_i = g(\sum_{i=1}^{N} W^T f_i + b_1)$$

where $g(z)$ is a sigmoid activation function. In this study, a logistic function $g(z) = \frac{1}{1+e^{-z}}$ is used. The network tries to minimize the following cost function:

$$J(W) = -\frac{1}{N}\{\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^{(k)} \log(\hat{y}_i)^{(k)} + (1 - y_i^{(k)}) \log(1 - \hat{y}_i)^{(k)}\}$$

Two measurements, the relevancy score and complementary score are introduced below for calculating the feature's score.

3.1. Relevancy score

The relevancy score is designed to show how important each feature is to the overall network. By removing the feature node in the network then calculating the network's performance, the relevancy of the feature can be obtained such that if the clamped feature is important, the network performance will be significantly affected. First, the base network is constructed using all the features F and its performance is used as the base line. Next, the feature f_i is removed from the network. In order to remove the feature without disrupting the whole network, a static value is used. In this study, a mean value of the feature is used ($f_i = \bar{f}_i$). This network is referred as the relevancy network. After the feature is removed, the network performance is

re-calculated and evaluated with the base line performance. Figure 1 shows the architecture and concept of the base line network and the network with removed feature.

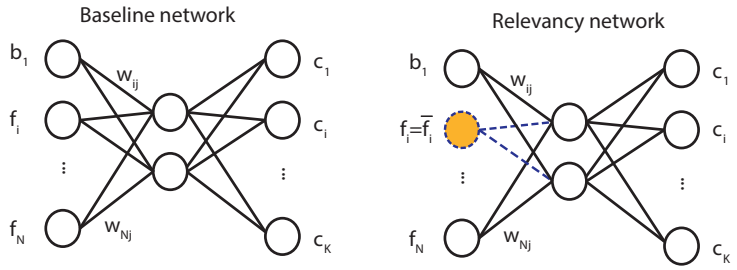

Figure 1: The architecture of the network with all features and the relevancy network

Given a set of feature F, the relevance of the feature Rel_{f_i} is:

$$Rel_{f_i} = 1 - \frac{P'(F|f_i = \bar{f}_i)}{P(F)} \tag{4}$$

where P is the generalized performance of the neural network using feature set F and P' is the generalized performance of the neural network using feature set F where feature f_i values are substituted by mean value of f_i. Note that the values of P and P' are between 0 and 1. The higher score of relevancy means the feature is more important. The score reflects the efficacy of the feature, should it be removed from the network. For example, $Rel_i = 0.7$ means that the absent of the feature f_i will lower the network's performance by 70%.

The relevance measurement only considers the relationship between a single input and the class. It does not consider the relationship between features i.e. redundancy and complementary. We enhance the clamping method by introducing another measurement to measure complementary of the features to the already selected feature set. Also, unlike other techniques which

consider redundancy measurement, MRMC considers feature complementary which to the best of our knowledge has not been used in other feature selection algorithms before.

3.2. Complementary score

The complementary score measures how much the feature complements the already selected features set. It also takes feature redundancy into account such that if the feature is redundant to the already selected features, the score should be low as it does not bring additional information to the classification. Firstly, the base line performance is obtained by constructing a network using all selected features S and calculating its performance. Next, a new feature f_i is added to the network. This network is referred as the complementary network. The architecture and concept of the base line network and the network with the new feature is shown in Figure 2.

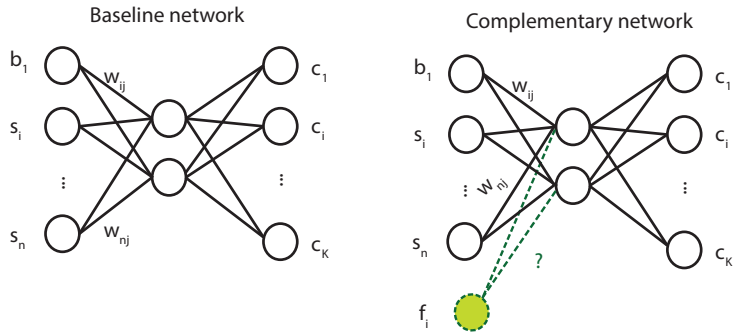

Figure 2: The architecture of the network with selected features and the complementary network

From Figure 2, it can be seen that the weights for the feature f_i needs to be obtained as they are not existed in the base line network. In our algorithm, we modify the construction of the complementary network such that

it partly uses the weights and biases from the base line network. We assume that the baseline network has already identified the correct weights for the already selected features. Thus, using the same weights and bias would help the network converges faster. This also reduces the possibility of the complementary network obtaining poor performance resulting from random initial weights. As the input and hidden nodes of the baseline network and the complementary network are different, the number of weights and biases are also different. The other weights and biases that are missing are generated randomly using the standard normal distribution with mean 0 and variance 1 scaled by the number of input nodes for bias and weights in the first layer and the number of hidden nodes in the second layer.

Given an already selected feature set S, the complementary of feature f_i to S can be calculated as:

$$Com_{f_i} = \frac{P(S \cup f_i)}{P(S)} - 1 \tag{5}$$

where $P(S \cup f_i)$ is the generalized performance of the complementary network and $P(S)$ is the generalized performance of the baseline network. The values of P are between 0 and 1. The complementary score reflects how much the new feature f_i contributes to the base line network. For example $Com_{f_i} = 0.1$ means by adding feature f_i, the performance of the network is improved by 10%.

3.3. Maximum relevance and maximum complementary score

The proposed algorithm ranks features based on the maximum relevance and maximum complementary score. After the relevancy and complementary scores are obtained, the relevance-complementary score (RC) can be

calculated as:

$$RC_{f_i} = Rel_{f_i} + Com_{f_i} \tag{6}$$

The feature is then selected based on the maximum RC score. From the algorithm, it can be seen that the complementary measurement can reduce the chance of selecting overlapping or redundant features. For example, given three features f_1, f_2, f_3 where $f_1 = f_3$ to represent overlapped feature and their relevance scores are expressed as $f_1 = f_3 > f_2$, if the clamping technique is used, the feature ranking will be f_1, f_3, f_2. However, by combing the complementary with relevancy, the ranking will be f_1, f_2, f_3. As the complementary score of f_3 should be zero, the RC score for f_2 will then be higher than f_3.

The steps of the MRMC algorithm are summarized in Figure 3 which are explained in detail below:

Step 1 : Normalize features value to [0 1] range. This step makes sure that features with larger values do not overwhelm features with smaller values. Set $S = \{\}$ and F contains all features.

Step 2 : Calculate the relevance score of all features f_i in F using (4). Note that the network is constructed using training data, then the generalized performance is calculated using validation data.

Step 3 : Select the first feature which has the maximum relevance score $f_s = \max_{f_i \in F} Rel(f_i)$.

Step 4 : Update S and F using equations (1) and (2).

Step 5 : Check if the size of feature set F is greater than 1. If Yes, go to Step 6. Otherwise, update S using $S = S \cup F$ and terminate the algorithm.

Step 6 : Calculate the complementary score for all features f_i in F using (5).

Step 7 : Calculate the RC score using (6).

Step 8 : Select feature f_s which has the maximum RC score $f_s = \{\max_{f_i \in F} RC(f_i)\}$. Go to Step 4.

4. Experimental results

This section presents evaluation results of MRMC against other feature selection methods as presented in Section 2. The experiments are carried out using two well-defined problems studied in [6] and four benchmark classification data sets including iris, breast cancer, cardiotocography, and chess which are obtained from UCI Machine Learning Repository [23] available at http://archive.ics.uci.edu/ml. The proposed algorithm is also evaluated using a real world data set which we have collected from multiple sensors used for predicting human activities.

The input features are discretized using bin 10 for calculating the MI of MRMR and NMIFS. For Clamping and MRMC, the number of hidden nodes is set to 2 $\times$ number of input nodes and the number of epochs is 300 whether or not the network converges. All experiments except the first and second experiments are carried out using 5-fold cross-validation where 3 folds are used for training, 1 fold for validation and 1 fold for testing. In this study, a balanced sampling is used where an equal number of positive and negative

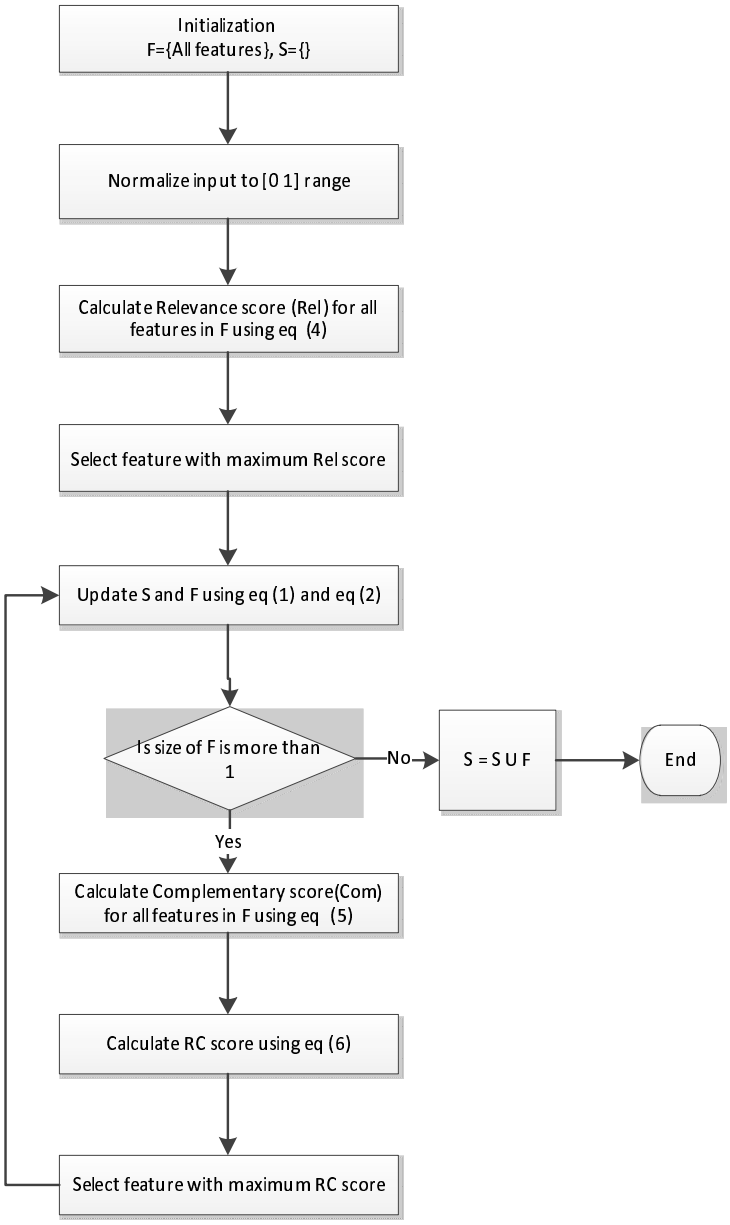

Figure 3: Flow chart of the MRMC feature selection algorithm

classes are randomly selected using a uniform distribution. The sizes of the training, validation and testing data of each fold used for different data set

are shown in Table 1.

For the real world problems, the feature selection methods are evaluated using NN. The number of hidden nodes is set to 2 $\times$ number of inputs and the number of epochs is set to 300. For each size of input, 10 models are constructed and the best one is selected using validation data. The test data is then applied to obtain the classification results. The validation data is also used to determine the size of features. The number of features is selected at the point where there is no significant improvement when more features are added. The data normality is tested and the appropriate test e.g. paired T-test or the wilcoxon Signed Ranks test is applied. In addition, the four algorithms are compared using statistical tests at 95% confidence interval. First, the data are tested for their normality using the shapiro-wilk test. If the data is normal distribution, then the ANOVA is used, otherwise a Friedman test is used. For the ANOVA test, if the mauchly result is significant, then the greenhouse-geisser test is reported.

Table 1: Characteristics and data partition of different data sets used in the study

Data set	# Features	# Classes	Data type	# Sample	# Training	# Validation	# Testing
Nonlinear AND	14	2	Real	500	500	-	-
Nonlinear AND with partly overlapped features	17	2	Real	500	500	-	-
Iris	4	3	Real	150	90	30	30
Cancer-1992	9	2	Integer	699	288	96	96
Cancer-1995	30	2	Real	569	252	84	84
Cardiotocography-fetal	21	3	Real	2126	315	105	105
Cardiotocography-morp	21	10	Real	2126	300	100	100
Chess	36	2	Categorial	3196	1830	610	610
Multi-sensor activity recognition	141	12	Real	39328	15120	5040	5040

4.1. Experiment 1: Nonlinear AND problem

In this experiment, a well-defined problem which the correct features are known is studied. We use a nonlinear AND problem previously studied in [12, 6]. There are 14 features in this problem. The first five features f_1 to f_5 are generated randomly from an exponential distribution with mean 10. These features represent irrelevant features. The next six features f_6 to f_{11} are relevant features generated randomly from a uniform distribution range [-1 1]. The next three features f_{12} to f_{14} are redundant features (fully overlapped features) where $f_{12} = f_9$, $f_{13}f_{10}$, $f_{14} = f_{11}$. The class label is determined by:

$$f(x) = \begin{cases} C_1 & \text{If} \quad f_6 * f_7 * f_8 > 0 \quad \text{AND} \quad f_9 + f_{10} + f_{11} > 0 \\ C_2 & \text{If} \quad f_6 * f_7 * f_8 < 0 \quad \text{AND} \quad f_9 + f_{10} + f_{11} < 0 \end{cases} \tag{7}$$

According to this problem, the optimal feature set is $\{f_6, f_7, f_8, [f_9 \text{ or } f_{12}], [f_{10} \text{ or } f_{13}], [f_{11} \text{ or } f_{14}]\}$. The set of 500 data samples is generated randomly from a uniform distribution. The class label for each data sample is determined using equation (7). Feature selection algorithms described in Sections 2 and 3 are applied on the data set. For Clamping and MRMC which require the validation data set, the 500 training data set is used. Table 2 presents the ranking results using these algorithms.

Table 2: Feature rankings of four feature selection methods (Nonlinear AND)

Algorithm	Feature rankings													
MRMR	f_{11}	f_9	f_1	f_2	f_4	f_3	f_{10}	f_5	f_6	f_8	f_7	f_{14}	f_{12}	f_{13}
NMIFS	f_{11}	f_9	f_{10}	f_6	f_8	f_7	f_3	f_4	f_2	f_1	f_5	f_{14}	f_{12}	f_{13}
Clamping	f_8	f_7	f_6	f_9	f_{11}	f_{12}	f_{14}	f_{10}	f_{13}	f_4	f_5	f_1	f_2	f_3
MRMC	f_8	f_7	f_6	f_9	f_{11}	f_{10}	f_{14}	f_{12}	f_{13}	f_4	f_5	f_1	f_3	f_2

From Table 2, it can be seen that only NMIFS and MRMC can identify the correct set of features. The first important feature ranked by MRMR and NMIFS is f_{11} and by Clamping and MRMC is f_8. This is expected as MRMR and NMIFS selects the first feature using maximum MI. Similarly, Clamping and MRMR use the same measurement to select the first feature. MRMR cannot detect the irrelevant feature where it ranks f_1 as the third important feature. Clamping correctly select the first five features, however it fails to detect that f_{12} is the redundancy of f_9 and f_{14} is the redundancy of f_{11}. According to this result, it can be seen that NMIFS gives the emphasis on detecting redundancy where it puts redundant features f_{12}, f_{13}, f_{14} at the end of the rank. On the contrary, MRMC gives emphasis on complementary where all irrelevant features are put at the end.

4.2. Experiment 2: A nonlinear AND problem with partly overlapped features

This experiment aims to show the superior ability of MRMC over the other three algorithms where it can select the correct features set from the data set which contains irrelevant, complete overlapped and partly overlapped features.

We use the same data set as generated in experiment 1. However, we introduce another three features f_{15} to f_{17} which represent partly overlapped features. Feature f_{15} is set to $f_{15} = f_6 * f_7$ which overlaps the feature f_6 and f_7. Feature f_{16} is set to $f_{16} = f_9 + f_{10}$ which overlaps the feature f_9 and f_{10}. Feature f_{17} is set to $f_{17} = f_8 * f_{11}$ which overlaps the feature f_8 and f_{11} but has no relationship to the classes. From this example, it can be seen that the relevant features are f_6 to f_{16}. f_{15} is the overlap of feature f_6 and f_7. However, it is better to select f_{15} and treat f_6 and f_7 as redundant as f_{15}

contains information from f_6 and f_7, therefore by selecting f_{15} the feature space would be smaller. The same reason also applies for selecting f_{16} over f_9 and f_{10}. The optimal subset of this data set is $\{f_8, [f_{11}$ or $f_{14}], f_{15}, f_{16}\}$.

Table 3: Feature rankings of four feature selection methods (Modified nonlinear AND)

Algorithm	Feature rankings																
MRMR	f_{16}	f_{11}	f_1	f_2	f_4	f_5	f_3	f_{15}	f_9	f_8	f_{10}	f_6	f_7	f_{17}	f_{14}	f_{12}	f_{13}
NMIFS	f_{16}	f_{11}	f_6	f_9	f_8	f_7	f_{10}	f_3	f_4	f_2	f_1	f_{14}	f_5	f_{15}	f_{12}	f_{17}	f_{13}
Clamping	f_{15}	f_8	f_{14}	f_{11}	f_9	f_{12}	f_4	f_{10}	f_{13}	f_6	f_{16}	f_1	f_{17}	f_3	f_5	f_7	f_2
MRMC	f_8	f_{15}	f_{16}	f_{11}	f_{14}	f_6	f_{10}	f_4	f_{12}	f_9	f_{13}	f_1	f_7	f_3	f_{17}	f_5	f_2

The result from Table 3 shows that only MRMC can produce the correct feature set. Only two features (f_{16}, f_{11}) are selected correctly by MRMR. The next five features selected by MRMR are irrelevant features. Clamping can select the first three features (f_{15}, f_8, f_{14}) correctly. However, the fourth feature (f_{11}) is the redundant of the third feature (f_{14}). This is because Clamping cannot detect overlap or redundant features. NMIFS can identify the first two features correctly. However, it selects f_6 and f_7 instead of f_{15} which makes the feature set larger. It also fails to detect that f_9 is the redundant feature of f_{16}.

4.3. Experiment 3: Iris data set

This data set widely used in classification literatures [24, 25] contains three classes of the type of Iris plant: Setosa, Versicolor, and Verginica. There are 50 samples each class. One class is linearly separable from the others. Two classes are not linearly separable. This data set has four features including sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm). The four feature selection algorithms are applied on the data set

and the mean classification accuracy of the test set is presented in Figure 4.

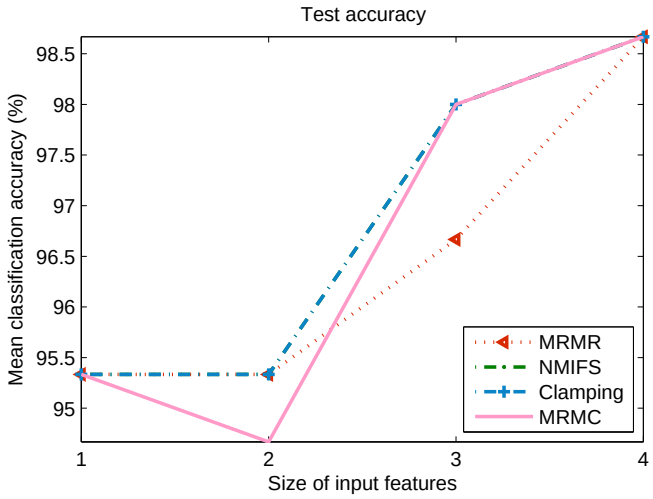

Figure 4: Mean classification accuracy of test data on four FS algorithms (Iris data set)

From Figure 4, all algorithms select the first feature correctly. MRMC does not correctly select the second feature in all folds and MRMR does not correctly select the third feature, therefore this slightly affects classification accuracy. The size of features, test classification accuracy and standard deviation are shown in Table 4. The results show that there is no statistical significance in classification accuracy among different feature selection algorithms (p=1.00).

Table 4: Feature sets selected by four feature selection algorithms and mean test accuracy (Iris data set)

Algorithm	No. of features	Mean test accuracy (%)	Standard Deviation
MRMR	1	95.333333	1.8257419
NMIFS	1	95.333333	1.8257419
Clamping	1	95.333333	2.9814240
MRMC	1	95.333333	2.9814240

4.4. Experiment 4: Wisconsin diagnostic breast cancer data set

This data set is used extensively in previous works [26, 10]. The breast cancer data set was obtained from the University of Wisconsin Hospitals, Madison [27]. This data set was collected in 1992 and we shall refer this data set Cancer-1992. It contains 9 integer-valued features such as clump thickness, uniformity of cell size, uniformity of cell shape, bland chromatin, etc. The values for each feature range between 1 and 10. There are 699 samples with 65.5% benign and 34.5% malignant cases. There are 16 samples with missing attribute values. In this study, 0 is used to replace any missing values. The mean classification accuracy on test data is shown in Figure 5.

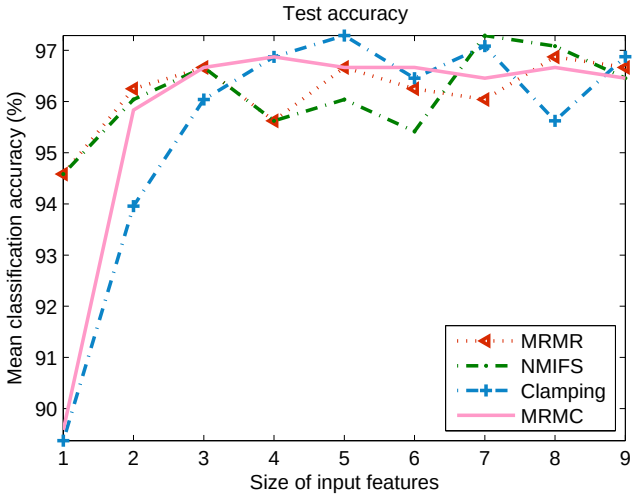

Figure 5: Mean classification accuracy of test data by four FS algorithms (Breast cancer data 1992 set)

From Figure 5, the accuracy of Clamping and MRMC starts lower than MRMR and NMIFS. MRMR, NMIFS and MRMC reach similar accuracy when 3 features are used. Clamping reaches the highest accuracy when 5

features are used. The accuracy of MRMR and MRMC remain steady after 3 features. The number of features used for each algorithm is shown in Table 5.

Table 5: Feature sets selected by four FS algorithms and mean test accuracy (Breast cancer 1992 data set)

Algorithm	# Selected features	Mean test accuracy (%)	Standard Deviation
MRMR	3	96.666667	2.8905077
NMIFS	4	95.625000	2.0036858
Clamping	8	95.625000	1.1410887
MRMC	2	95.833333	1.6470196

Based on the mean test accuracy, the algorithms' performances can be expressed as Clamping<NMIFS<MRMC<MRMR. The statistical test indicate that there is no significant difference between the four algorithms (Chi Square(3)=1.826, p=0.609). When we look at the number of features used in each algorithm, it can be seen that MRMC uses the smallest number of features. Hence, MRMC is the most optimum algorithm for this data set.

We also evaluate the proposed algorithm on another breast cancer data set which was collected in 1995. It is composed of 30 real-valued input features computed from a digitalized image of cell nucleus such as radius, texture, smoothness, mean, standard error, etc. to determine whether the cell is malignant or benign. The data set contains 357 benign and 212 malignant samples. The mean classification accuracy of the test data set for all four algorithms are shown in Figure 4.

From Figure 6, the first feature selected by Clamping and MRMC has lower accuracy then the feature selected by MRMR and NMIFS. However, using two selected features by MRMC, the accuracy is significantly improved.

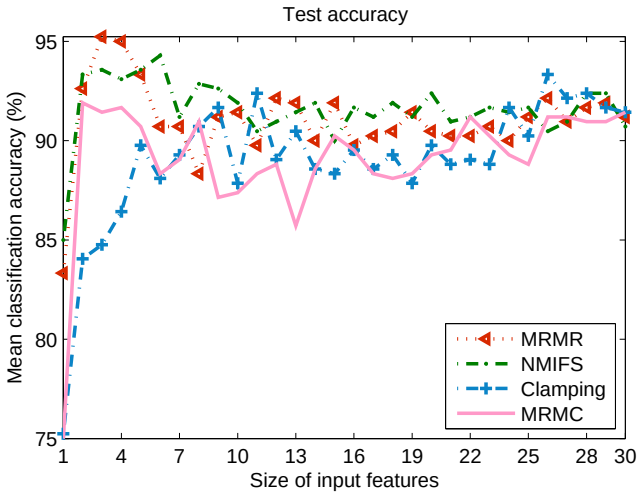

Figure 6: Mean classification accuracy of test data by four FS algorithms (Breast cancer 1995 data set)

MRMR and NMIFS provide similar performances on this data set. The number of features and performances of the four algorithms are shown in Table 6.

Table 6: Feature sets selected by four FS algorithms and mean test accuracy (Breast cancer data 1995 set)

Algorithm	# Selected features	Mean test accuracy (%)	Standard Deviation
MRMR	4	95.000000	2.7147034
NMIFS	1	85.000000	13.0573044
Clamping	2	84.047619	8.7319623
MRMC	2	91.904762	2.7147034

The test accuracy for each algorithm is shown in Table 6. Based on the test accuracy, the algorithms' performances can be expressed as $Clamping < NMIFS < MRMC < MRMR$. The normality test shows that the data have normal distribution. The results indicate that there is no statistical sig-

nificant between each algorithm ($F(1.322, 5.288) = 2.273, p = 0.192$). From Table 6, it can be seen that NMIFS uses the lowest number of features. Therefore, it can be concluded that NMIFS is the optimum method on this data set.

4.5. Experiment 5: Cardiotocography data set

This data set used in [28] contains the measurement of fetal heart rate (FHR) and uterine contraction features e.g. minimum FHR histogram, percentage of time with abnormal long term variability, etc. on cardiotocograms classified by expert obstetricians. The data set contains 21 input features which are classified into 10 types of morphologic patterns or 3 fetal states. The data set has the unbalanced class distribution.

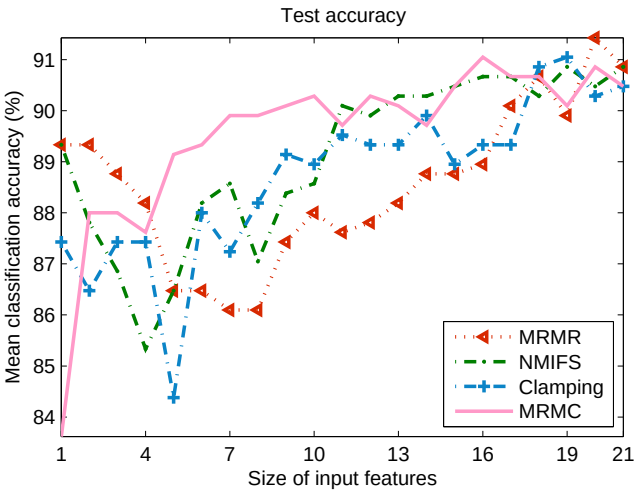

Figure 7: Mean classification accuracy of test data by four FS algorithms (Cardiotocography-Fetal data set)

The average classification accuracy of 10-class morphologic patterns and 3-class fetal states are shown in Figs. 7 and 8, respectively. From Figure 7,

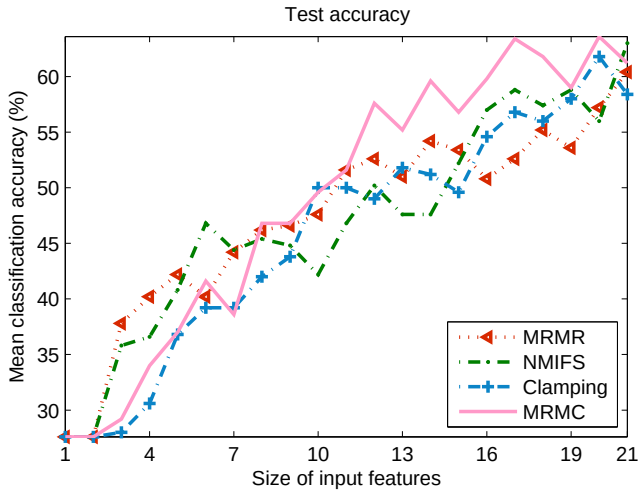

Figure 8: Mean classification accuracy of test data by four FS algorithms (Cardiotocography-Morp data set)

the classification accuracy of MRMC starts at the lowest but continues improving as the number of features increases. The classification accuracy of MRMR and NMIFS are high when using one feature. The classification accuracy of NMIFS falls to the lowest point when 5 features are used. The performance of MRMC is better than the other 3 algorithms when 6 to 10 features are used. From Figure 8, all feature selection algorithms produce similar accuracy trend. The classification accuracy is improved when more number of features is used. The performance of MRMC is superior than the other 3 algorithms when 12 to 17 features are used.

Table 7 shows the number of features selected by each algorithm, the mean classification accuracy on test data and the standard deviation on the cardiotocography data set for classifying 3 fetal states. Based on the test classification accuracy, the performance of each algorithm can be expressed as $MRMC < Clamping = NMIFS < MRMR$. The results indi-

Table 7: Feature sets selected by four FS algorithms and mean test accuracy (Cardiotocography-Fetal data set)

Algorithm	# Selected features	Mean test accuracy (%)	Standard Deviation
MRMR	18	90.666667	1.7036708
NMIFS	15	90.476190	2.0203051
Clamping	21	90.476190	1.5058465
MRMC	4	87.619048	3.5634832

cate no statistical significance between accuracy obtained by four algorithms ($F(1.045, 4.18) = 6.711, p = 0.058$). From Table 7, it can be seen that MRMC selects the lowest number of features. Therefore, it can be concluded that MRMC is the optimum method on this data set.

Table 8: Feature sets selected by four FS algorithms and mean test accuracy (Cardiotocography-Morp data set)

Algorithm	# Selected features	Mean test accuracy (%)	Standard Deviation
MRMR	21	91.428571	1.5058465
NMIFS	16	90.666667	1.2417528
Clamping	21	87.428571	4.1184282
MRMC	15	83.619048	6.8146834

Table 8 shows the results of four methods on classifying 10 morphologic patterns of cardiotocography data set. Based on the mean classification accuracy on the test data, the performance of the algorithms is expressed as $MRMC < NMIFS < Clamping < MRMR$. The shapiro-wilk is applied to test data normality. The results show that there is no statistical significant in accuracy among four algorithms ($F(3, 12) = 0.278, p = 0.840$). Among the four algorithms, it can be seen that MRMC uses the lowest number of features. Therefore, it can be concluded that MRMC is the optimum feature selection method for this data set.

4.6. Experiment 6: Chess data set

The chess data set used in [29, 30] contains sequences of chess-description for chess end game. The data set consists of 36 categorical-input features to classify if the White can win or cannot win. The class distribution is 52% win and 48% cannot win. The equal class distribution is used and the number of training, validation, and testing data is shown in Table 2. The data set uses a string to represent the board-description e.g. f, l, n, etc. which we convert these into integer values e.g. f=1, l=2, n=3, etc. The mean classification accuracy of the test data set is shown in Figure 9. The classification result of each algorithm using the number of features determined by validation data is presented in Table 9.

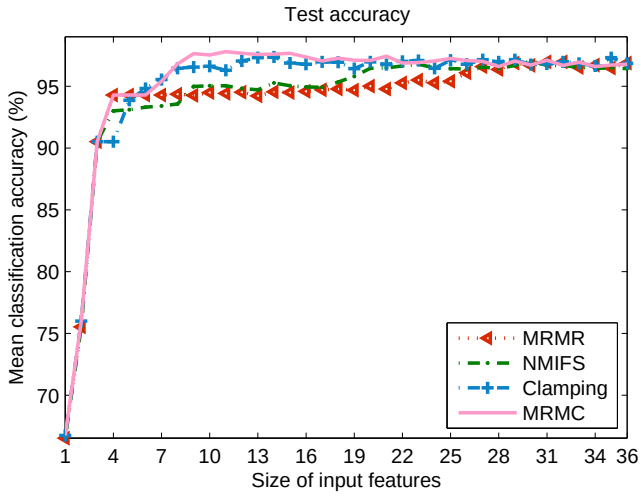

Figure 9: Mean classification accuracy of test data by four FS algorithms (Chess data set)

From Figure 9, the performance of all algorithms increases when using more features. When we observe the features selected by each algorithm, it is found that the first three features selected are the same. Generally, the

performance of Clamping and MRMC is better than that of MRMR and NMIFS in this data set. MRMC performance is better than Clamping when 8 to 21 features are used. All algorithms reach similar accuracy when 29 and more features are used.

Table 9: Feature sets selected by four FS algorithms and mean test accuracy (Chess data set)

Algorithm	# Selected features	Mean test accuracy (%)	Standard Deviation
MRMR	27	96.590164	1.2234825
NMIFS	20	96.459016	1.2835134
Clamping	14	97.377049	0.9628967
MRMC	10	97.540984	0.8439041

Based on the mean classification accuracy of test data, the algorithms' performance can be expressed as $NMIFS < MRMR < Clamping < MRMC$. The result reveals that there is no statistical significant difference among the algorithms ($p = 0.054$). Based on the number of features used in each algorithm, MRMC uses the lowest number while MRMR uses the highest number of features. Therefore, we can conclude that MRMC is the most optimum algorithm for this data set.

4.7. Experiment 7: Multi-sensor activity recognition data set

We collect raw sensor data of accelerometer, gyroscope, heart rate monitor, light, temperature, altimeter, and barometer from 12 elderly people performing 12 activities of daily livings including walking, feeding, exercising, reading, watching TV, washing dishes, sleeping, ironing, feeding, scrubbing, wiping, and brushing teeth. The participants wear the sensors on their wrists and heart rate monitor on their chests. The data set consists of 141 real-valued input features. The classification accuracies of the test data set for

all algorithms are shown in Figure 10. The performance of the algorithms is reported in Table 10.

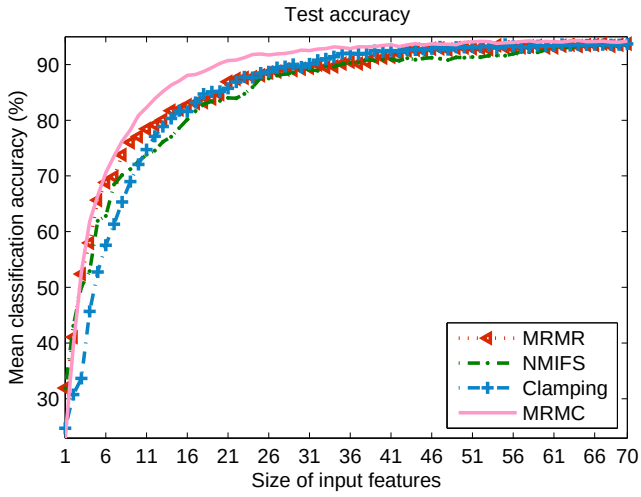

Figure 10: Mean classification accuracy of test data by four FS algorithms (Multi-sensor activity recognition data set)

From Figure 10, MRMC has better accuracy than that of the other algorithms when four or more features are used. The large different accuracy is noticeable when 10 and 34 features are used. Clamping performance start lower than other algorithms. However, its accuracy is similar to MRMR and NMIFS when more than 14 features are used. MRMR and NMIFS start at the same accuracy. However, NMIFS performance drops when 3 and 25 features are used.

Based on the mean classification accuracy on test data, the algorithms' performance can be expressed as $MRMR < Clamping < NMIFS < MRMC$. The results indicate that there is no significant difference among the four algorithms ($F(1.474, 5.895) = 1.417, p = 0.301$). Based on the num-

Table 10: Feature sets selected by four FS algorithms and mean test accuracy (Multi-sensor activity recognition data set)

Algorithm	# Selected features	Mean test accuracy (%)	Standard Deviation
MRMR	64	93.313500	0.4858261
NMIFS	66	93.662700	0.5337766
Clamping	62	93.611120	0.8777444
MRMC	50	94.027800	0.6026319

ber of features used in each algorithm, it can be seen that MRMC only uses 50 while the other use over 60 features. Therefore, we can conclude that MRMC is the most optimum algorithm for this data set.

5. Discussion and conclusion

The summary of the experiments is presented in Table 11. The optimal feature selection algorithms of each data set is based on the statistical result and the number of features. Based on 8 experiments, MRMC is the optimum feature selection algorithm in general. It is able to obtain high classification result using the minimum number of features. NMIFS is the second optimum algorithm.

The results from experiments 1 and 2 show that MRMC is capable of detecting completely overlapped and partial overlapped features. In other experiments, the result also shows that MRMC can be used on various data types i.e. categorical, real, and integer values. The performance of MRMC is not as good as NMIFS in the breast cancer-1995 data set. This is due to the fact that the first feature selected by MRMC normally has lower classification accuracy. The difference in accuracy between NMIFS and MRMC are about 10% when 1 feature is selected. When looking at other data sets,

Table 11: Optimum FS algorithm on each data set

Data set	Optimum feature selection algorithm
Nonlinear AND	NMIFS, MRMC
Nonlinear AND with partly overlapped features	MRMC
Iris	MRMR, NMIFS, Clamping, MRMC
Cancer-1992	MRMC
Cancer-1995	NMIFS
Cardiotocography-fetal	MRMC
Cardiotocography-morp	MRMC
Chess	MRMC
Multi-sensor activity recognition	MRMC

the differences are about 5% or less. This implies that if another algorithm obtains a significantly higher accuracy than MRMC when using 1 feature, then that algorithm would be more optimum for that data set provided the number of input features is small.

The result of cardiotocography data set indicate that MRMC is the most optimum algorithm among the four algorithms. It achieves good accuracy while using the smallest number of features. Experiment 6 demonstrates that MRMC also works well with categorical data. In experiment 7, we evaluate the proposed algorithm with the data set with a large number of inputs. The result shows that MRMC is much superior in which it uses less features than other algorithms while achieving the highest accuracy. When comparing MRMC with Clamping, it can be seen that by introducing a complementary measurement, the performance of the algorithm is better. For example, in the breast cancer 1995 data set, using the same number of features, MRMC can obtain higher accuracy. Overall, the obtained experimental results show that the main advantage of MRMC over other state-of-the-art techniques is

the number of selected features.

From this study, it can be seen that using the clamping algorithm to detect the most important feature may not give the correct result. This affects the performance of MRMC as it uses the same criteria to select the first feature. As forward search is used, the performance of the feature selection algorithm depends on the first selected feature. Therefore, in case of the feature selection of a small set of features, using MRMC may not guarantee good results. However, when the number of features increases, MRMC is superior than the other three algorithms. This is due to the fact that although the first feature selected by the clamping algorithm may not always be the most important but it is important i.e. the second or third most important feature, and by using complementary measurement, the correct subset of features can later be identified.

In this paper, we propose a new feature selection algorithm based on maximum relevance maximum complementary based on neural network. We evaluate the proposed methods on well-defined problems and real world data sets containing small to larger set of features (N=4 to 100+). The study is carried out using 5-fold cross validation. The algorithms performance is evaluated empirically using statistical tests at 95% confidence interval. We show that in general MRMC provides a good result comparing to the other three algorithms such that it can select smaller set of features. We also demonstrate that the complementary measure introduced improves the performance of the clamping algorithm. The main advantages of the proposed technique are the capability of selecting smaller set of features and especially works well when it is used with a large number of features. Nowadays, a

large volume of data is available due to the development of sensors and technologies. It is essential to be able to select only the important features in order to reduce the costs such as computation, storage, and battery and to increase performances of any expert and intelligent systems. The proposed algorithm can be applied to any type of data, has most potential when it is used with data which has a large number of features.

Since we have found that the performance of the algorithm is affected by the selection of the first feature, future research will be focusing on the identification of the first feature in order to improve the MRMC performance. According to the experiment, the MI-based technique is good at obtaining the first feature. The future work will be carried out to integrate the MI-based technique with MRMC. Another interesting research direction is to study feature selection by considering a group of features rather than a single feature i.e. relevancy of a group of features. This idea arises when we observe the case when there are more than one feature with equal importance. In order to correctly identify the feature, the next important feature needs to be considered. Finally, since we can see the potential of the complementary concept, it is interesting to see how to apply this concept in other feature selection approaches.

References

[1] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, I. Korhonen, Activity classification using realistic data from wearable sensors, Information Technology in Biomedicine, IEEE Transactions on 10 (2006) 119–128.

[2] J. Ward, P. Lukowicz, G. Troster, T. Starner, Activity recognition of assembly tasks using body-worn microphones and accelerometers, Pattern Analysis and Machine Intelligence, IEEE Transactions on 28 (2006) 1553–1567.

[3] M. Ermes, J. Parkka, J. Mantyjarvi, I. Korhonen, Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions, Information Technology in Biomedicine, IEEE Transactions on 12 (2008) 20–26.

[4] O. Banos, M. Damas, H. Pomares, A. Prieto, I. Rojas, Daily living activity recognition based on statistical feature quality group selection, Expert Systems with Applications 39 (2012) 8013 – 8021.

[5] H. Peng, F. Long, 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 27 (2005) 1226–1238.

[6] S. Cang, H. Yu, Mutual information based input feature selection for classification problems, Decision Support Systems 54 (2012) 691 – 698.

[7] Y. Zhang, Z. Zhang, Feature subset selection with cumulate conditional mutual information minimization, Expert Systems with Applications 39 (2012) 6078 – 6088.

[8] R. Setiono, H. Liu, Neural-network feature selector, Neural Networks, IEEE Transactions on 8 (1997) 654–662.

[9] W. Wang, P. Jones, D. Partridge, Assessing the impact of input features in a feedforward neural network, Neural Computing & Applications 9 (2000) 101–112.

[10] M. M. Kabir, M. M. Islam, K. Murase, A new wrapper feature selection approach using neural network, Neurocomputing 73 (2010) 3273 – 3283.

[11] Y. Saeys, I. Inza, P. Larraaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.

[12] P. Estevez, M. Tesmer, C. Perez, J. Zurada, Normalized mutual information feature selection, Neural Networks, IEEE Transactions on 20 (2009) 189–201.

[13] J.-X. Peng, S. Ferguson, K. Rafferty, P. D. Kelly, An efficient feature selection method for mobile devices with application to activity recognition, Neurocomputing 74 (2011) 3543 – 3552.

[14] A. Dalton, G. Olaighin, Comparing supervised learning techniques on the task of physical activity recognition, Biomedical and Health Informatics, IEEE Journal of 17 (2013) 46–52.

[15] M. Zhang, A. Sawchuk, Human daily activity recognition with sparse representation using wearable sensors, Biomedical and Health Informatics, IEEE Journal of 17 (2013) 553–560.

[16] A. Khan, Y.-K. Lee, S. Lee, T.-S. Kim, A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer, Information Technology in Biomedicine, IEEE Transactions on 14 (2010) 1166–1172.

[17] J. Varkey, D. Pompili, T. Walls, Human motion recognition using a wireless sensor-based wearable system, Personal and Ubiquitous Computing 16 (2012) 897–910.

[18] H.-H. Hsu, C.-W. Hsieh, M.-D. Lu, Hybrid feature selection by combining filters and wrappers, Expert Systems with Applications 38 (2011) 8144 – 8150.

[19] C. E. Shannon, A mathematical theory of communication, SIGMOBILE Mob. Comput. Commun. Rev. 5 (2001) 3–55.

[20] C. O. Sakar, O. Kursun, F. Gurgen, A feature selection method based on kernel canonical correlation analysis and the minimum redundancy-maximum relevance filter method, Expert Systems with Applications 39 (2012) 3432 – 3437.

[21] A. Verikas, M. Bacauskiene, Feature selection with neural networks, Pattern Recognition Letters 23 (2002) 1323 – 1335.

[22] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Inc., New York, NY, USA, 1995.

[23] K. Bache, M. Lichman, UCI machine learning repository, 2013. URL: `http://archive.ics.uci.edu/ml`.

[24] J. G. Dy, C. E. Brodley, S. Wrobel, Feature selection for unsupervised learning, Journal of Machine Learning Research 5 (2004) 845–889.

[25] P. Zhong, M. Fukushima, Regularized nonsmooth newton method for

multi-class support vector machines, Optimization Methods and Software 22 (2007) 225–236.

[26] A. Antos, B. Kégl, T. Linder, G. Lugosi, Data-dependent margin-based generalization bounds for classification, J. Mach. Learn. Res. 3 (2003) 73–98.

[27] K. P. Bennett, O. L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, 1992.

[28] D. Ayres-de campos, J. Bernardes, A. Garrido, J. Marques-de s, L. Pereira-leite, Sisporto 2.0: A program for automated analysis of cardiotocograms, Journal of Maternal-Fetal and Neonatal Medicine 9 (2000) 311–318.

[29] B. Kijsirikul, S. Sinthupinyo, K. Chongkasemwongse, Approximate match of rules using backpropagation neural networks, Machine Learning 44 (2001) 273–299.

[30] I. Cohen, F. Cozman, N. Sebe, M. Cirelo, T. Huang, Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction, Pattern Analysis and Machine Intelligence, IEEE Transactions on 26 (2004) 1553–1566.