Maximum Relevance Maximum Anti-Redundancy (mRmA) Feature Selection

Abdul Mannan[1], Kashif Javed[2], Serosh Karim Noon[1]

1. *Department of Electrical Engineering, NFC Institute of Engineering and Technology, Multan, Pakistan*
2. *Department of Electrical Engineering, University of Engineering and Technology, Lahore, Pakistan*

* **Corresponding Author:** Email: mannan@nfciet.edu.pk

Abstract

Filters represent a class of feature selection methods used to select a subset of useful features from high dimensional data on the basis of relevance and redundancy analysis. Maximum relevance minimum redundancy (mRMR) is a famous feature selection algorithm for microarray data [1]. The quotient based version of maximum relevance minimum redundancy (Q-mRMR) filter [1],[2] selects, at each iteration, the feature scoring maximum ratio between its class relevance and average redundancy over already selected subset. This ratio can be surprisingly large if the denominator i.e. redundancy term is very small, hence suppressing the effect of relevance and leads to the selection of features which can be very weak representatives of the class. This paper addresses this issue by presenting a maximum relevance maximum antiredundancy (mRmA) filter method. For mRmA the value of objective function is within reasonable limits for all values of relevance and redundancy, hence, making selection of appropriate features more probable. Our 10 fold cross validation accuracy results using naive Bayes and support vector machines (SVM) classifiers confirm that the proposed method outperforms both Q-mRMR and Fast Correlation based Filter (FCBF) methods on six datasets from various applications like microarray, image and physical domains.

Key Words: Feature selection, filter, Gene expression data, Image, Support vector machine, Naive bayes.

1. Introduction

In real world applications, datasets with very large number of features [3],[4] have become common these days. Analyzing such data for classification tasks not only questions the computational capability of the classifier but also adversely impacts its accuracy [3],[5]. However, not all features present in a dataset have the same worth. Those which help in class discrimination are relevant while those that do not provide additional information are redundant [6]. Since, the class discrimination capability of irrelevant features is poor, a method which reduces the size of the data by eliminating redundant and irrelevant features is highly desirable. One way to attain dimensionality reduction is feature selection; it selects a subset of features most relevant for the class variable while keeping their original meanings intact [3],[7]. A lot of research has been done on this topic over the last decade; researchers are continuously trying to find a strategy which can be generalized for all types of data.

Feature selection has applications in a wide variety of domains like biomedical [5], text [3] and images [4] etc. For microarray gene expression datasets where the number of features reach up to several thousand [8],[9] obvious advantage of selecting most important features/ genes saves computational cost in addition to making data more interpretable and improving classification accuracy by removing redundant features [10]. Among the feature selection algorithms proposed in the literature, filters are well-known. The quotient based maximum relevance minimum redundancy feature selection (Q-mRMR) method proposed by Peng et. al.[1],[2] is very popular especially in bioinformatics. The method searches for a subset having maximum class relevance and minimum inter-feature redundancy. Heuristic presented for obtaining this subset starts by picking the most relevant feature; the next feature is added on the basis of maximum ratio between its class relevance and average redundancy with the already selected feature(s). Peng et al [1] have demonstrated that Q-mRMR is, under various settings, more accurate than distance based mRMR. However, a shortcoming of Q-mRMR is that its objective function takes on unexpectedly large values for small values of redundancy. This, in certain situation, may compel the method to prioritize less relevant features over more relevant ones and hence can affect the classification accuracy of the generated subset. In this paper, we propose a method to overcome this issue using an anti-redundancy term instead of a redundancy term. Unlike Q-mRMR, the values taken by the objective function of our method are between zero and one and the features are selected in a balanced way giving equal priority to both relevance and redundancy. To investigate the usefulness of the

newly proposed method, we carry out experiments on six datasets from various application domains using Naive Bayes and support vector machine classifiers.

The rest of the paper is organized in five sections. In Section 2, work related to feature selection is presented. Section 3 describes our proposed algorithm along with its motivation. In Section 4, we present the experimental settings used for this work while results are presented in Section 5. We draw conclusions of this work in Section 6.

2. Related Work

Feature selection algorithms can broadly be divided in three classes; wrapper, filter and embedded depending on how a subset of the most relevant features is searched. Wrappers [6] search the feature space exhaustively with the help of a classifier. The feature subset producing the highest accuracy is finally selected; selecting a set of highly relevant and least redundant features is quite probable but over-fitting on training data, being dependent on the classifier used and large computational expense are considered to be their major disadvantages [1],[6].

Filters [3],[11] select a feature subset by maximizing an objective function which employs some metric for estimating the relevance and redundancy of features. The method is independent of a classifier and requires fewer computations than the wrapper [3]. It is the most popular method in feature selection community and has been extensively investigated in the literature [12],[13]. Unlike wrappers and filters, embedded methods integrate the task of feature selection and classification in one step [5]. For example, in recursive feature elimination (RFE), weights assigned by the support vector machines (SVM) classifier to each feature are used as ranking weights; features with the top weights are selected at each step [5].

The aim of feature selection is to obtain small subsets of features highly correlated with the class and uncorrelated with each other [14]. There can be a number of metrics used for calculating relevance and redundancy e.g. mutual information [15],[13], symmetric uncertainty [16] or Pearsons correlation coefficient [17]. Different metrics estimate relevance and redundancy differently thus resulting in different performances for the filters [11]. Peng et. al.[1], [2] propose mutual information for both relevance and redundancy in case of discrete data, whereas for continuous data, they use F-test correlation Quotient [2] to calculate class relevance and Pearsons correlation coefficient to determine redundancy between two features. Koller and Sahami [7] discover Markov Blanket of features on the basis of correlation and retain a set of relevant features through backward elimination. To make filters computationally more feasible, many filters select features in two stages and drop the low ranked features in their first step [16], [18]. For example Liu et. al.[16] proposed fast correlation based feature selection (FCBF) method which proceeds in forward direction and eliminates a feature if its class relevance is smaller than its redundancy with the most recently selected feature.

Brown et. al.[19] present a unifying feature selection method based on mutual information. They conclude that for small datasets, joint mutual information (JMI) method proposed by Yang and Moody [20] is optimum in terms of accuracy, stability and flexibility. Mahmoud et. al.[18] and Apiletti et. al.[21] rank features on the basis of inter class overlapping score; they determine a mask for the complete dataset and search for the subset whose mask is the same as of reference mask.

Ooi et. al.[22] suggest finding an optimal subset of features by tuning an exponential parameter over relevance, redundancy and anti-redundancy terms. The anti-redundancy term is obtained by subtracting redundancy from 1. Relevance is calculated on the basis of F-score statistic between a feature and the class and redundancy between two features is calculated on the basis of Pearson correlation coefficient. They propose two variants, one is redundancy based (Equation 1) and the other is based on anti-redundancy (Equation 2). The feature selection heuristic is the same as proposed by Peng [2].

$$FS_R = \frac{(relevance)^{\alpha}}{(redundancy)^{1-\alpha}} \quad (1)$$

$$FS_A = (relevance)^{\alpha} * (antiredu\ ndancy)^{1-\alpha} \quad (2)$$

where FS_R is their proposed feature selection approach using relevance and redundancy whereas FS_A is the feature selection using relevance and anti-redundancy. Optimum value of the exponent α which yields highest accuracy is found empirically for various datasets. They further conclude that FS_A outperforms FS_R in general.

3. The Newly Proposed Maximum Relevance Maximum Anti-redundancy (mRmA) Feature Selection Algorithm

In this section, we propose a feature selection

method inspired by quotient based maximum relevance minimum redundancy (Q-mRMR) filter proposed by Peng et. al.[2]. Q-mRMR filter selects a feature if the ratio between its class relevance and average redundancy over the set of already selected features is maximum in that iteration. Equation 3 describes how its objective function selects the i^{th}feature.

$$Q_{mRMR} = max_{i \in list}\left[\frac{I(i,class)}{\frac{1}{|S|}\sum_{j \in S} I\,(i,j)}\right] \quad (3)$$

Were, $list$ refers to the list of features not selected so far, and S is the subset of already selected features at any iteration. $I(i,class)$ shows class relevance of a feature taken from the list and $I(i,j)$ shows redundancy between the i^{th} feature and the j^{th} feature which belongs to the subset of already selected features. The feature that produces the highest value of this ratio is selected. For Q-mRMR, if data is continuous, relevance is calculated as F-score between a feature and the class and average redundancy term is the average of Pearsons correlation coefficient calculated between the feature under consideration and all the features already selected. On the other hand, for discrete data, both relevance and redundancy are based on mutual information.

Since the redundancy term in Q-mRMR(Equation 3) is in the denominator, its very small value can result in a very large value of the objective function even for small values of relevance; this will force Q-mRMR to select features that are weakly related to the class. A feature with low redundancy with the existing set but of lower relevance for the class will be assigned a higher score as compared to a highly relevant feature having moderate redundancy with the selected features.

To address this issue, we propose an anti-redundancy based maximum relevance minimum redundancy (mRmA) feature selection algorithm which is a variant of Q-mRMR. In Section 3.1, we elaborate this idea with the help of a motivating example. The anti-redundancy term is obtained by subtracting the redundancy term from 1 and the objective function maximizes product of relevance and anti-redundancy terms as given in Equation 4.

$$mRmA = max_{i \in list}\left[I(i,class) \times \frac{1}{|S|}\sum_{j \in S}[1 - I(i,j)]\right] \quad (4)$$

3.1 A Motivating Example

Q-mRMR does not fairly rank features when the average redundancy with respect to the already selected feature set is very small. Since this term is in the denominator, a small value generates a large value for the fraction and the role of relevance term in numerator is suppressed. To understand this idea, lets take the help of a synthetic dataset shown in Table 1. The data is continuous with 3 features, 6 instances and 2 classes. As shown in Figure 1, the class is clearly segregated by features x and y because there is no overlap in the values of features for class 0 and class 1. This indicates that these two features have good class relevance. Pearson correlation coefficient which is used by both Q-mRMR and mRmA assigns x a score equal to 0.72 while y has a score 0.84. The third feature z is weekly related with class; this is because its values for both classes overlap each other more than 60% of the time which results in its small class relevance equal to 0.19.

Both Q-mRMR and mRmA start by first selecting the top ranked feature which is y for the dataset shown in Table 1. Out of x and z, Q-mRMR employs the search technique described in Equation 3 for the selection of the next feature while mRmA uses Equation 4. Q-mRMR selects feature z while our proposed algorithm mRmA selects x as the second feature. Feature x is highly relevant for the class and has a reasonably small redundancy equal to 0.3728 with the already selected subset (which is feature y at the moment). On the other hand, z has small class relevance and very small redundancy equal to 0.0925 with already selected feature y. The objective function of Q-mRMR assigns a higher score to z and selects it. The final ranking generated by Q-mRMR is < y, z, x > whereas mRmA prefers x over z. The ranked list of features of mRmA is given by < y, x, z >. This shows that in certain cases Q-mRMR prefers weakly relevant features on strongly relevant features which can lead to degradation in the accuracy of selected subset.

3.2 Understanding mRmA

To further highlight the differences between the working of the objective functions of Q-mRMR and mRmA, we refer to Figures 2, 3 and 4. The values of objective function for Q-mRMR ranges from 0 to a very large number while those of mRmA lies in the [0 1] range. We scale the values of the Q-mRMR objective function between 0 and 1. This allows us to plot both the objective function values on the same scale, thus providing insight into the working of Q-mRMR and mRmA. In both images shown in Figure 2, relevance is plotted against redundancy. The values taken by the objective functions for Q-mRMR and mRmA

are shown as shades of varying intensities. Using the convention of image processing, complete black means zero and complete white refers to value equal to 1 [23]. All gray shades in between these two extremes represent values between 0 and 1. In each image, bottom left corner is the origin, relevance increases by moving to the right on horizontal axis and redundancy increases by moving up on vertical axis; both vary from 0 to 1 with a step of 0.1. It can be seen in Figure 2(a) that the last row, which belongs to very small values of redundancy, is significantly brighter than the rest of the image. This forces the feature selection algorithm to select features having low redundancy by giving less importance to class relevance.

$$As\ redu\ ndancy \to 0$$

$$Objective\ Fu\ nction(Q_{mRMR}) \to \infty$$

In contrast to Q-mRMR, uniformly varying gray values pertaining to redundancy = 0 in Figure 2(b) show that the objective function of our proposed mRmA becomes equivalent to class relevance based feature selection for such small values of redundancy

$$As\ redu\ ndancy \to 0$$

$$Objective\ Fu\ nction(Q_{mRMR}) \to rel(f)$$

curves/lines for Q-mRMR and mRmA each, with redundancy plotted against the values taken by the objective functions for a given value of relevance. We vary relevance from 1 to 0 with a step of 0.1. For mRmA, the blue straight lines indicate the variations in its objective function. It can be seen that all these lines are linearly decreasing from a maximum value to zero. This maximum value is directly proportional to relevance meaning that for large values of relevance the objective function generates values in a large range and vice versa, which is quite intuitive.

For relevance equal to zero, the value generated by objective function is also zero. On the other hand, the red curves in the Figure 3 correspond to Q-mRMR. These lines indicate the non-linearly decreasing behavior of the objective function with increasing redundancy. The maximum value is set by the relevance but the fall is so sharp that the objective function saturates (starts generating very small and almost equal values) as the redundancy increases just beyond zero.

It can be further seen in Figure 3 lines corresponding to various values of relevance are significantly gapped away for mRmA whereas these are indistinguishable for Q-mRMR. Moreover, if we move along x-axis from redundancy = 1 to redundancy = 0 the vertical gap between the values taken by the objective function of mRmA keeps on increasing but this gain is very small for Q-mRMR. Finally if we draw a vertical line passing through a certain value of redundancy, we can see that the mRmA is expected to select features with greater value of class relevance than Q-mRMR.

This renders Q-mRMR inappropriate for cases where some features, irrespective of relevance, have very small values of average redundancy with the already selected subset. In other words, Q-mRMR prefers features with very low redundancy over better alternatives i.e. features having large value of relevance and a small value of redundancy. The same concept is described in Figure 4, where relevance is plotted on x-axis, redundancy on y-axis and the values taken by the objective functions of the two methods along z-axis.

3.3 Applications of Proposed Feature Selection Method

The proposed feature selection method can be used in various practical applications including microarray, image processing, signal processing, speech and geological domains. The data from numerous applications exhibit different characteristics. For example, microarray data contain expression levels of haundreds of thousands of genes involved to describe a certain phenotype. The goal is to rule out genes having no contribution or the ones which are exact or approximate copies of other genes. That is where feature selection comes in. Using our proposed feature selection method, we can easily select the most relevant and least redundant genes.

Another interesting application is image processing where selecting important attributes from a shape descriptor can be challenging. An example is a dataset containing pixel data of handwritten digits. Out of thousands of pixels, only few hundred are the best descriptor of the digit. Using our feature selection method, we can eliminate attributes (pixel information) which are either weakly related to the target class (i.e. handwritten digit) or redundant.

In Geology, the scientists are usually interested in exploring presence of petroleum reservoirs under earth's crust. For that purpose, 2 or 3 dimensional arrays of sensors are placed on ground that record the reverberation of sound from various objects (e.g. rocks, water reservoirs and petroleum reservoirs etc.). One can only be interested in the subset of sensors' data relevant to the petroleum

reservoirs (the target class in this case) which of course can be done through feature selection.

4. Experiments

This section describes the experimental settings which we used to carry out our comparisons. Rankings generated by the Q-mRMR, FCBF and the newly proposed mRmA methods are investigated and compared.

4.1 Classifiers

We tested the performance of the three feature selection methods using two classifiers i.e. Support Vector Machines (SVM) [5] and Naive Bayes [24]. Both these classifiers are widely used in the literature [25],[26] due to their simplicity of implementation and efficiency [27]. For all experiments, we have used the implementation of both classifiers given in MATLAB [28] and WEKA toolbox [29]

4.1.1 Support Vector Machines (SVM) Classifier

Support Vector Machines (SVM) [5] is a well-known classifier used to discriminate between two classes. For linearly separable data, linear kernel is used and positive and negative examples are segregated using a line equidistant from the support vectors in each class. For cases where data cannot be separated by a single line, a non linear kernel (e.g. RBF) can be used [27]. The original SVM classifier can classify between two classes only but the concept can easily be extended for multiple classes [27]. In this work, we use one versus all SVM classifier i.e. out of many classes, one class is considered as positive and the rest are assumed to be negative.

4.1.2 Naive Bayes (NB) Classifier

The Naive Bayes classifier is based on Bayes rule [27], which calculates posterior conditional probability of a class given a sample assuming features are conditionally independent.

4.2. Datasets

We used six datasets in our experiments; four from the microarray, one from image and one from physical domain. The summary of the data sets is given in Table 2. Datasets contain either continuous real numbers or positive integers. For the datasets, number of classes range from 2 to 4, number of features from 40 to 11,340 and number of instances from 62 to 5,000.

Microarray datasets are CLL-SUB-111 [30], Colon [31], Leukemia [32], and TOX-171 [33]. CLL-SUB-111 is a microarray dataset to identify two genetic subtypes of B-cell chronic lymphocytic leukemia (B-CLL). It contains 111 instances out of which 100 represent either of the two types of B-CLL and the remaining 11 samples are from healthy controls and hence three classes. This is a high dimensional dataset containing 11,340 features and the data is continuous varying from 0 to 1. Colon is another microarray gene expression dataset indicating whether a sample comes from tumor biopsy or not (2-classes). It has 2,000 features, 62 instances and the data contains multi-valued integer feature values ranging from 1 to 5. Leukemia is a cancer gene expression dataset containing 7,129 features and two classes, either AML or ALL. The original dataset is divided in 34 training and 38 test instances. We join them together to construct a data of 72 patients [2]. The data is continuous ranging from 0 to 1. TOX-171 is toxicology dataset which makes use of clinical chemistry and expression data from liver of rats 48 hours after inducing three types of toxicants namely alpha-naphthyl-isothiocyanate, dimethylnitrosamine and N-methylformamide. Three classes represent the mechanism of toxicity of each compound and the fourth class represents samples from untreated controls. Data is continuous with 171 samples and 5,748 features.

GINA [4] is an agnostic learning hand written digit dataset. The task is to distinguish two digits odd number from two digits even number i.e. two classes. Since only one digit (least significant) is enough to tell whether the number is even or odd, at least 50% of the data is redundant. This is a sparse dataset containing continuous numbers between 0 and 1. For our analysis, we converted it to binary by setting threshold equal to zero i.e. the values equal to zero in original dataset are kept zero and all non zero values are converted to 1. The operation is the same as it is used to convert a gray scale image to black & white [23]. The dataset contains 970 features and 3,153 instances. Waveform [34] dataset contains large number of instances each being a combination of two of the three base waves with noise (mean=0 and variance=1) added. Dataset has 40 attributes and the output is one of three possible resultant waves. Data contains integers ranging from 1 to 5.

4.3. Performance Evaluation

We evaluated the performance of three feature selection methods on the basis of 10 fold cross validation accuracy on nested subsets of top 20 features of each dataset. In order to find the 10

fold cross validation accuracy, the data was randomly divided in 10 equal portions, nine were used for training the classifier and then it was tested on the left out portion; the process was repeated ten times giving each portion a chance to become testing data. Finally, the average gives 10 foldcross validation accuracy with given number of features. The experiment was repeated 20 times for each dataset and for each classifier.

5. Results

In this section, we present the results obtained when FCBF, Q-mRMR and mRmA were tested on six datasets using SVM and Naive Bayes classifiers. For the FCBF filter, we used its implementation given in FEAST toolbox [19]. The subset generated by the method may not contain 20 features for all datasets; to combat this issue we concatenated the selected and rejected subsets and picked top 20 features from the combined set.

5.1 CLL-SUB-111 Dataset

CLL-SUB-111 is the highest dimensional dataset used in our experiments. Figure 5(a) shows that for CLL-SUB-111 dataset, our proposed feature selection method is throughout better than Q-mRMR and FCBF using SVM classifier. It can also be seen in the first entry of Table 3 that mRmA has achieved the highest accuracy for 15 nested subsets of features, which is much greater than the other two competitors. mRmA demonstrates its superiority even for naive Bayes classifier too. As shown in Figure 5(b) it produces far better accuracy than the other two methods, FCBF is the second best and Q-mRMR is the poorest. Moreover, as shown in the first entry of Table 4 mRmA produces the highest accuracy for 18 nested subsets of features on this dataset with naive Bayes.

5.2 Colon Dataset

Colon is another microarray gene expression dataset for which our proposed method outperforms FCBF and Q-mRMR. Using SVM classifier (refer to Figure 6(a)) mRmA is the best for small number of features but later on all three methods saturate to a constant value of accuracy. As shown in Table 3, mRmA is on top for 16 nested subsets of features. Using naive Bayes classifier (Figure 6(b)) mRmA is better than Q-mRMR initially and then they all saturate with Q-mRMR a bit better than the other two methods. Table 4 reveals that Q-mRMR and mRmA achieve the best accuracy for equal number of nested subsets of features.

5.3 Leukemia Dataset

As shown in Figure 7(a), mRmA reaches to the highest accuracy with minimum number of features and is close to FCBF. mRmA is, however, far better than Q-mRMR for the dataset. Table 3 indicates that our proposed method mRmA is able to generate highest accuracy for all 19 nested subsets of increasing features. Using naive Bayes classifier, mRmA is just better than FCBF and is significantly superior to Q-mRMR (refer to Figure 7(b)). Table 4 shows that out of 19 nested feature subsets, mRmA produces highest accuracy for 17 times.

5.4 TOX-171 Dataset

Figure 8(a) indicates that for the TOX-171 dataset, FCBF and Q-mRMR produce almost similar accuracy for all nested subsets. However, mRmA is throughout better than the other two. Further, mRmA produces highest accuracy (refer to Table 3) for 14 nested subsets while FCBF is the best only 4 times and Q-mRMR is only twice. FCBF is the best method for the dataset using naive Bayes classifier, mRmA is slightly inferior while Q-mRmR is the poorest of all (refer to Figure 8(b)). Table 4 shows that FCBF produces the highest accuracy for 16 nested subsets whereas our proposed mRmA is the best for 5 times but is 14 times better than Q-mRMR.

5.5 GINA Dataset

Using SVM classifier and thresholded GINA dataset, mRmA outperformed Q-mRMR (Figure 9(a)) while being very close to FCBF. For small subsets of nested features, the superiority of mRmA over Q-mRMR is more significant than with larger subsets. For 17 out of 19 nested subsets of features, mRmA produces the highest accuracy (Table 3). Using naive Bayes classifier, FCBF performs the best by remaining very close to mRmA. They both, however, are better than Q-mRMR initially but all three methods saturate to maximum accuracy as the size of nested features subset increases beyond 8. As it is mentioned in Table 4, both FCBF and mRmA generate the highest accuracy 11 times in 19 trials.

5.6 Waveform Dataset

For the waveform dataset, both FCBF and mRmA are better than Q-mRMR for small nested subsets of features and saturate to a higher accuracy than obtained by Q-mRMR (refer to Figure 10(a)). Table 3 indicate that mRmA and FCBF are comparable for this dataset generating highest accuracy for 16 and 12 times respectively. For

naive Bayes classifier, all three methods are comparable in the beginning (refer to Figure 10(b)) but mRmA settles to a slightly larger value of accuracy than the other two methods. The second last entry of Table 4 establishes the superiority of our proposed mRmA method as it generates highest accuracy for 17 nested subsets of features which is much greater than obtained through Q-mRMR and FCBF.

5.7 Overall Performance

For all datasets using both SVM and naive Bayes classifiers, our proposed mRmA achieved the best overall performance for increasing number of nested feature subsets. The final entry of Table 3 shows that for all six datasets, mRmA method achieved the highest accuracy 85% of the times whereas for FCBF this is only 19% and is the least for Q-mRMR i.e. only around 8%. Similarly, the last row of Table 4 indicates that mRmA attains the highest accuracy in around 70% of the cases, the runner up is once again FCBF (around 33%) and Q-mRMR is the best method for only 19% of the nested feature subsets. However, our proposed algorithm is equally accurate or inferior to Q-mRMR for datasets where there is "no small relevance and very small redundancy issue" mentioned in detail in Section 3.

6. Conclusion

In this paper, we have shown that irrespective of relevance the objective function of Quotient based maximum relevance minimum redundancy (Q-mRMR) [1] filter produces surprisingly large values for very small values of redundancy. Such large values favor those features which have low redundancy with the selected features but are less relevant for the class.

Due to this shortcoming, Q-mRMR can select features which are highly irrelevant to the class. To address this issue, we propose a new algorithm named maximum relevance maximum anti-redundancy (mRmA). It replaces the redundancy term in the Q-mRMR with anti-redundancy and the division operation with multiplication. Experiments were conducted on six real world datasets from various domains using features generated by the three methods i.e. FCBF, Q-mRMR and mRmA with 10 fold cross validation. Results confirm that mRmA performs better than both FCBF and Q-mRMR for SVM and Naive Bayes classifiers. This shows that replacing redundancy by anti-redundancy term in Q-mRMR improves its accuracy on a variety of datasets from various application domains including physical, image and microarray.

Acknowledgements

The authors are thankful to the Department of Electrical Engineering at NFC Institute of Engineering and Technology, Multan, Pakistan for providing computational resources for this work.

References

[1] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.

[2] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.

[3] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[4] Javed, K., Babri, H. A., & Saeed, M. (2012). Feature selection based on class-dependent densities for high-dimensional binary data. IEEE Transactions on Knowledge and Data Engineering, 24(3), 465-477.

[5] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1), 389-422.

[6] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.

[7] Koller, D., & Sahami, M. (1996). Toward optimal feature selection. Stanford InfoLab.

[8] Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, 5(Oct), 1205-1224.

[9] Yu, L., & Liu, H. (2004, August). Redundancy based feature selection for microarray data. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 737-742). ACM.

[10] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

[11] Javed, K., Babri, H. A., & Saeed, M. (2014). Impact of a metric of association between two variables on performance of filters for binary data. Neurocomputing, 143, 248-260.

[12] Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial intelligence, 97(1), 245-271.

[13] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.

[14] Hall, M. A., & Smith, L. A. (1999, May). Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. In FLAIRS conference (Vol. 1999, pp. 235-239).

[15] Doquire, G., &Verleysen, M. (2013). Mutual information-based feature selection for multilabel classification.Neurocomputing,122, 148-155.

[16] Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 856-863).

[17] Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1996). Numerical recipes in C (Vol. 2). Cambridge: Cambridge university press.

[18] Mahmoud, O., Harrison, A., Perperoglou, A., Gul, A., Khan, Z., Metodiev, M. V., &Lausen, B. (2014). A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. BMC bioinformatics, 15(1), 274.

[19] Brown, G., Pocock, A., Zhao, M. J., &Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. Journal of machine learning research, 13(Jan), 27-66.

[20] Yang, H., & Moody, J. (1999, June). Feature selection based on joint mutual information. In Proceedings of international ICSC symposium on advances in intelligent data analysis (pp. 22-25).

[21] Apiletti, D., Baralis, E., Bruno, G., & Fiori, A. (2007, August). The painter's feature selection for gene expression data. In Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE (pp. 4227-4230). IEEE.

[22] Ooi, C. H., Chetty, M., &Teng, S. W. (2006). Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data. BMC bioinformatics, 7(1), 320.

[23] Gonzalez, R. C. (2009). Digital image processing: Pearson Education India.

[24] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). IBM.

[25] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and information systems, 34(3), 483-519.

[26] Kira, K., & Rendell, L. A. (1992, July). A practical approach to feature selection. In Proceedings of the ninth international workshop on Machine learning (pp. 249-256).

[27] Alpaydin, E. (2014). Introduction to machine learning. MIT press.

[28] Duin, R. P. W., Juszczak, P., Paclik, P., Pekalska, E., De Ridder, D., Tax, D. M. J., &Verzakov, S. (2000). A matlab toolbox for pattern recognition. PRTools version, 3, 109-111.

[29] Jelinek, H. F., Depardieu, C., Lucas, C., Cornforth, D. J., Huang, W., & Cree, M. J. (2005, November). Towards vessel characterization in the vicinity of the optic disc in digital retinal images. In Image Vis ComputConf (pp. 2-7).

[30] Haslinger, C., Schweifer, N., Stilgenbauer, S., Döhner, H., Lichter, P., Kraut, N., &Abseher, R. (2004). Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. Journal of Clinical Oncology, 22(19), 3937-3949.

[31] Cho, S. B., & Won, H. H. (2003, January).

Machine learning in DNA microarray analysis for cancer classification. In Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19 (pp. 189-198). Australian Computer Society, Inc..

[32] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., &Yakhini, Z. (2000). Tissue classification with gene expression profiles. Journal of computational biology, 7(3-4), 559-583.

[33] James, A. P., &Dimitrijev, S. (2012). Feature selection using nearest attributes. arXiv preprint arXiv:1201.5946.

[34] Bailey, J., Manoukian, T., &Ramamohanarao, K. (2003, November). A fast algorithm for computing hypergraph transversals and its application in mining emerging patterns. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 485-488). IEEE

Table 1: A synthetic dataset and feature ranking obtained through Q-mRMR and mRmA

X	Y	Z	Class	Class Relevance			Redundancy with Feature Y		Feature Ranking	
				X	Y	Z	X	Z	Q-mRMR	mRmA
0.1	0.4	0.4	1	0.72	0.84	0.19	0.3728	0.0925	y, z, x	y, x, z
0.6	0.5	0.9	1							
0.8	0.6	0.5	1							
0.9	0.0	1.0	0							
0.9	0.2	0.0	0							
1.0	0.3	0.4	0							

Table 2: Summary of datasets

No.	Dataset	Application	Features	Instances	Classes	Type of Data
1	CLL-SUB-111	Microarray	11,340	111	3	Continuous
2	Colon	Microarray	2,000	62	2	Integer
3	Leukemia	Microarray	7,129	72	2	Continuous
4	TOX-171	Microarray	5,748	171	4	Continuous
5	GINA	Image	970	3,153	2	Continuous
6	Waveform	Physical	40	5,000	3	Integer

Table 3. The best accuracy count of each feature selection method using SVM classifier

Dataset	Winner Count		
	FCBF	Q_mRMR	mRmA
CLL-SUB-111	0	4	15
Colon	3	2	16
Leukemia	3	0	17
TOX-171	0	1	19
GINA	4	2	14
Waveform	12	0	16
Winning Percentage	**19.29%**	**7.89%**	**85.08%**

Table 4: The best accuracy count of each feature selection method using naïve Bayes classifier

Dataset	Winner Count		
	FCBF	Q_mRMR	mRmA
CLL-SUB-111	1	0	18
Colon	2	11	11
Leukemia	11	6	11
TOX-171	5	3	17
GINA	16	2	5
Waveform	3	1	17
Winning Percentage	**33.33**	**20.17**	**69.29**

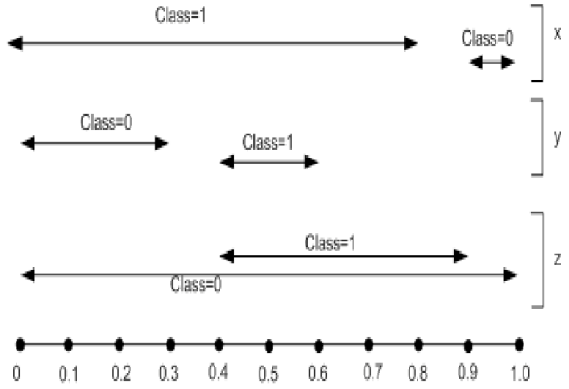


Figure 1: The features of the dataset given in Table 1 and their class discrimination capability

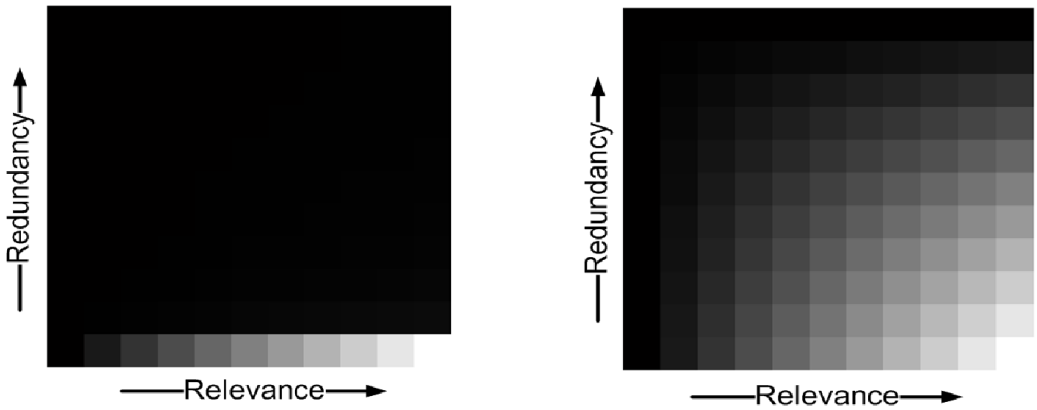


Figure 2: Values of objective function taken by Q-mRMR and mRmA for various values of relevance and redundancy represented as grayscales

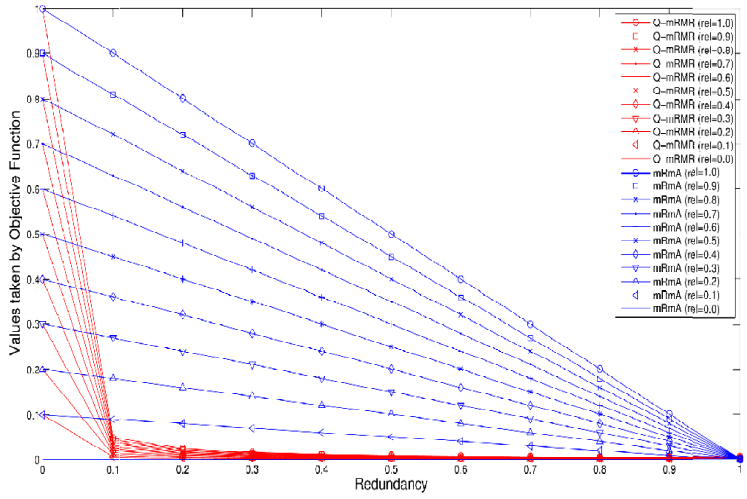


Figure 3: Objective function of Q-mRMR and mRmA plotted verses redundancy. Each line/curve represents a different value of relevance

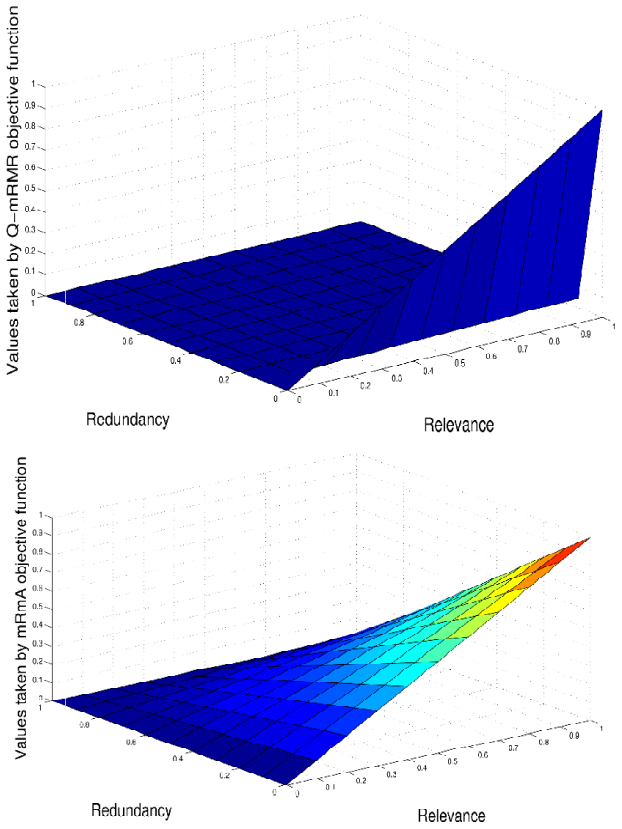
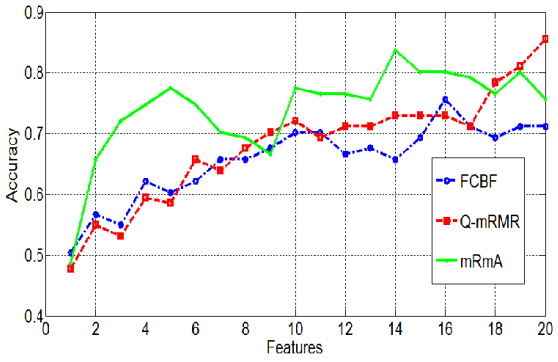
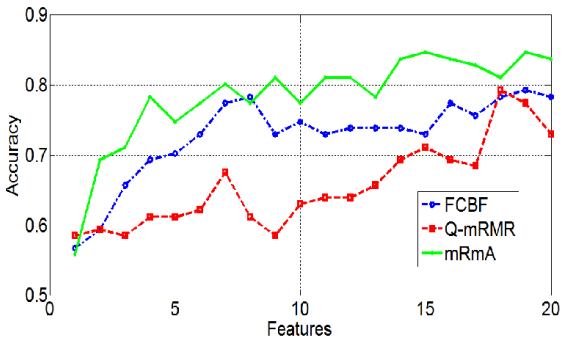


Figure 4: Values of objective function taken by Q-mRMR and mRmA for various values of relevance and redundancy shown as 3D plot.

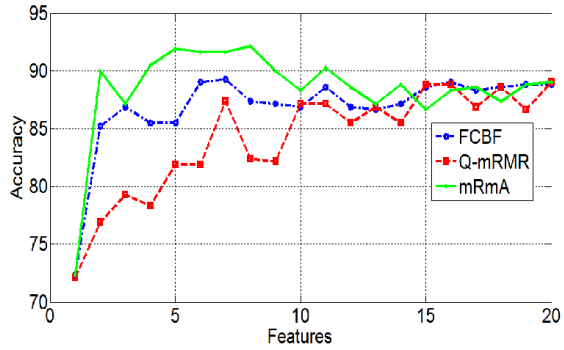


(a) SVM

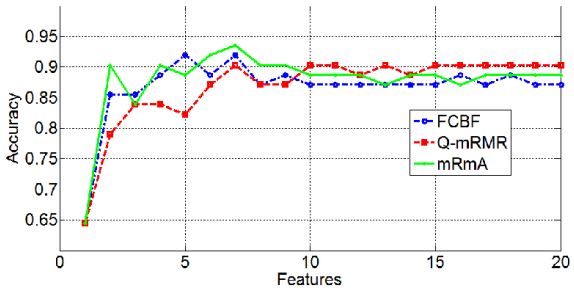


(b) Naïve Bayes

Figure 5: 10 fold cross validation accuracy of CLL-SUB-111 dataset.

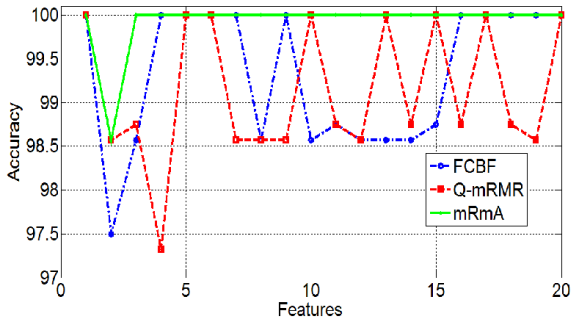


(a) SVM

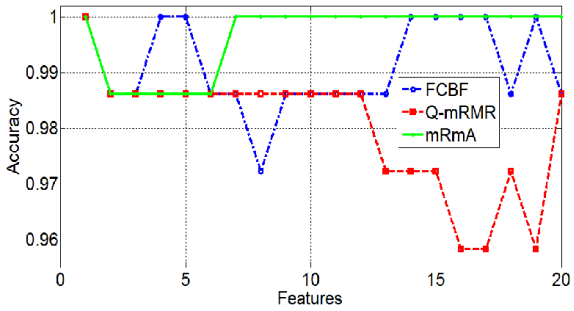


b) Naïve Bayes

Figure 6: 10 fold cross validation accuracy of Colon dataset.

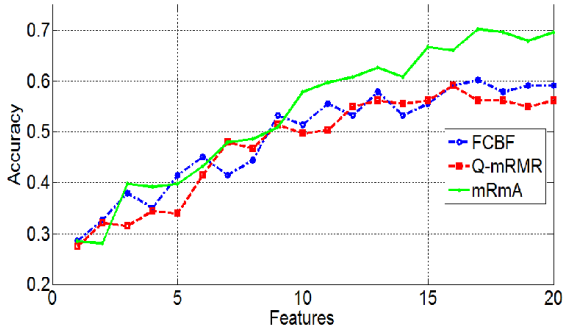


(a) SVM

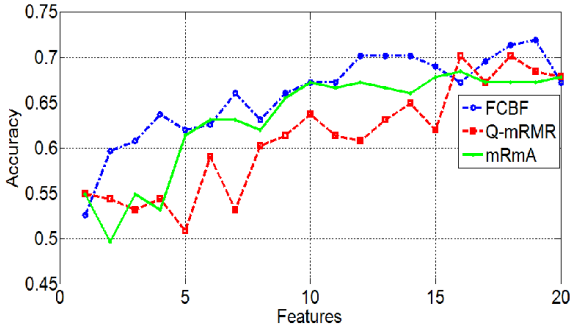


(b) Naïve Bayes

Figure 7: 10 fold cross validation accuracy of Leukemia dataset

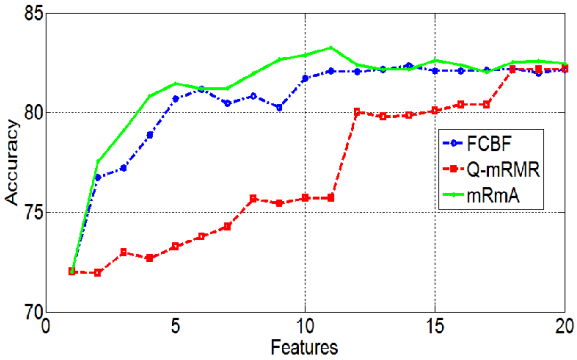


(a) SVM

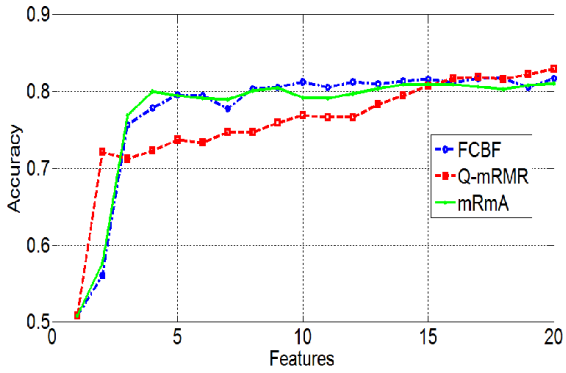


(b) Naïve Bayes

Figure 8: 10 fold cross validation accuracy of TOX-171 dataset.

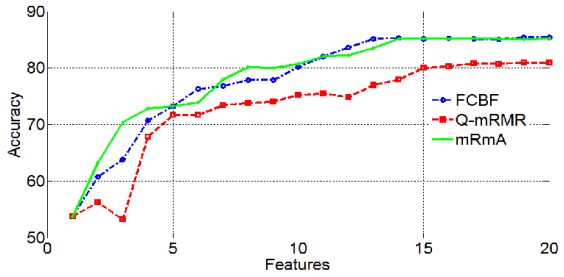


(a) SVM

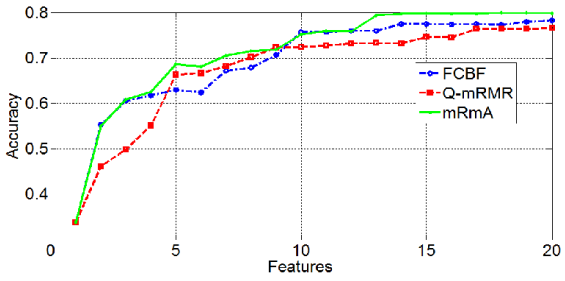


(b) Naïve Bayes

Figure 9: 10 fold cross validation accuracy of GINA dataset



(a) SVM



(b) Naïve Bayes

Figure 10: 10 fold cross validation accuracy of Waveform dataset