Information-based Feature Selection

Farzan Farnia, Abbas Kazerouni, Afshin Babveyh
Email: {farnia,abbask,afshinb}@stanford.edu

1 Introduction

Feature selection is a topic of great interest in applications dealing with high-dimensional datasets. These applications include gene expression array analysis, combinatorial chemistry and text processing of online documents. Using feature selection brings about several advantages. First, it leads to lower computational cost and time. Less memory is needed to store the data and less processing power is needed. Feature selection helps improve the performance of the predictors by avoiding overfitting. It can also capture the underlying connection between the data. And perhaps the most important aspect, it can break through the barrier of high-dimensionality.

To select the most relevant subset of features, we need a mathematical tool to measure dependence among random variables. In this work, we use the concept of mutual information. Mutual information is a well-known dependence measure in information theory. For any arbitrary pair of discrete random variables, $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, *Mutual Information* is defined as

$$I(X;Y) \;=\; \sum_{x\in\mathcal{X},\, y\in\mathcal{Y}} p_{X,Y}(x,y)\, \log \frac{p_{X,Y}(x,y)}{p_X(x)\, p_Y(y)}. \tag{1}$$

The paper is organized as follows. In section 2 the method of Maximum-Relevance Minimum-Redundancy (MRMR) is presented along with Maximum Joint Relevant (MJR) method. In section 3, we present our method to solve the feature selection problem. Section 4 presents the result of our algorithm tested on Madelon dataset. Finally, section 5 discusses the conclusion.

2 Mutual Information as a tool for Feature Selection

As discussed earlier, mutual information is a powerful tool in measuring relevance among random variables. Hence, it can be a useful mathematical tool to find and select relevant features. In other words, if our goal is to select no more than k features an optimal task is to solve

$$\arg\max_{|S|=k}\; I(X_S;Y), \tag{2}$$

where $X_S = \{X_i : i \in S\}$. However, as k gets larger our estimation of mutual information becomes less accurate. It is because for large k's we do not have enough samples to estimate mutual information accurately. Hence, the objective function in (2) should be modified so that it becomes estimable by available samples. In the next sections, we first discuss a past approach to solve this issue and then propose a new solution to improve such approaches.

2.1 Max-Relevance Min-Redundancy (MRMR) approach

As mentioned earlier, we aim to identify the most relevant subset of features whose size is limited to a given factor. Note that this is not the same as characterizing the k best features with the most individual mutual information to the target Y. In fact, different features may share redundant information on the target. Thus, redundancy is another important factor to be considered in feature selection. To balance the trade-off between relevance and redundancy, the following modified objective function (MRMR) has been suggested in [2]:

$$\Phi(X_S, Y) \;=\; \frac{1}{|S|}\sum_{i\in S} I(X_i;Y) - \frac{1}{|S|^2}\sum_{i,j\in S} I(X_i;X_j). \tag{3}$$

Here, the first term measures the average relevance of features to the target, while the second term measures average pairwise redundancy among selected features. Therefore, maximizing $\Phi(X_S, Y)$ leads to identifying a well-characterizing feature subset whose total information on the target is close to the optimal feature subset's. To maximize this objective, they used an inductive approach where first the most informative feature is chosen, and then next features are inductively added by solving the following at every step:

$$\arg\max_{X_j\in X\setminus S_m} I(X_j;Y) - \frac{1}{m-1}\sum_{X_i\in S_m} I(X_j;X_i). \tag{4}$$

2.2 Maximum Joint Relevance

Although MRMR is a well-known feature selection method, there are several applications where the test error rate never goes below some large thresholds like 34% which seems quite unsatisfactory. Note that (3) includes only up to pairwise interactions. By considering higher order interactions we can become able to select a more informative feature subset which in turn results in smaller error rates. To this end, *Maximum Joint Relevant* (MJR) algorithm changes the inductive rule of (4) to a more sensitive one [3]:

$$\arg\max_{X_j\in X\setminus S_m} \sum_{X_i\in S_m} I(X_j, X_i\,;\, Y). \tag{5}$$

Nevertheless, we may again encounter the issue of lack of enough samples to estimate the second order mutual information appeared in the above formulation. As a matter of fact, a considerable number of third order empirical marginals may become too small, and thus it requires a more accurate estimation of mutual information than the empirical one. Therefore, in next section we are going to propose a new algorithm to estimate mutual information with higher accuracy. As an important advantage, this estimation technique reduces the required sample size to estimate mutual information within the same accuracy.

3 Adaptive Maximum Joint Relevant

In this section, we propose the *Adaptive Maximum Joint Relevant* (AMJR) feature selection algorithm to tackle the instability problem in MJR. Similar to MJR, we use the criterion in (5) to iteratively select the most relevant features. However, we propose a new scheme to estimate the mutual informations which stabilize the algorithm in small training set regimes. We build our estimation technique based on functional estimation method proposed in [4]. Specifically, in order to

estimate $I(X_j, X_i; Y)$ at each step, we have to estimate the joint entropies according to the following identity:

$$I(X_j, X_i; Y) = H(X_j, X_i) + H(Y) - H(X_j, X_i, Y). \tag{6}$$

In order to describe the estimation method in AMJR, consider for example, estimating $H(X_j, X_i)$. Following from [4], first the empirical joint distribution of (X_j, X_i) is computed according to

$$\hat{P}_{a,b} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{(X_j, X_i)^{(t)} = (a, b)\}, \tag{7}$$

where n is the size of training set and $(X_j, X_i)^{(t)}$ is the joint value of t^{th} training example. Note that a and b are assumed to take value in some finte sets $\mathcal{A}$ and $\mathcal{B}$, respectively. Now, assuming that $P_{a,b}$ is the true joint probability of (X_j, X_i) at point (a, b), the true joint entropy would be

$$H(X_j, X_i) = - \sum_{a \in \mathcal{A},\ b \in \mathcal{B}} P_{a,b} \log P_{a,b}. \tag{8}$$

In order to provide the estimator $\hat{H}(X_j, X_i)$ of $H(X_j, X_i)$, one naive way is substitute each $P_{a,b}$ in (8) with its estimate $\hat{P}_{a,b}$. This method which is used in MJR, is in fact the source of instability on the performance since most of the estimated probabilities are very small. In AMJR, we consider two cases for the estimated joint probabilities:

- If $\hat{P}_{a,b} \geq \frac{\log n}{n}$, we use it as an estimation of $P_{a,b}$ in (8).
- If $\hat{P}_{a,b} < \frac{\log n}{n}$, first we fit a polynomial f of order $\lfloor \log n \rfloor$ to the function $x \log x$ in the interval $(0, \frac{\log n}{n})$. Then, we use $f(\hat{P}_{a,b})$ as an estimation for $P_{a,b} \log P_{a,b}$ in (8).

As we see in Section 4, the approximation polynomial f introduces stability to the algorithm and improves its performance. Consequently, the estimation of $H(X_j, X_i)$ in AMJR would be

$$\hat{H}(X_j, X_i) = -\left(\sum_{\hat{P}_{a,b} \geq \frac{\log n}{n}} \hat{P}_{a,b} \log \hat{P}_{a,b} + \sum_{\hat{P}_{a,b} < \frac{\log n}{n}} f(\hat{P}_{a,b}) \right). \tag{9}$$

Similarly, the estimations $\hat{H}(X_j, X_i, Y)$ and $\hat{H}(Y)$ are provided for $H(X_j, X_i, Y)$ and $H(Y)$, respectively. Finally, the mutual information is estimated as

$$\hat{I}(X_j, X_i; Y) = \hat{H}(X_j, X_i) + \hat{H}(Y) - \hat{H}(X_j, X_i, Y). \tag{10}$$

4 Numerical Results

In this section we provide numerical results to confirm our theoretical analysis. We perform different feature selection and classification methods on the dataset "Madelon" released in NIPS 2003 feature selection challenge [5]. This data set consists of 2000 samples each containing 500 continuous input features and one binary output response. Here we have used 1400 samples (70%) as the training set and used the other 600 samples (30%) as the test set.

In order to explore the effect of sample size on different feature selection methods, we quantize the input space into 3 and 5 levels, uniformly. Thus, we have two scenarios. In the first one, the input features are quantized separately into three levels which corresponds to the large training set regime

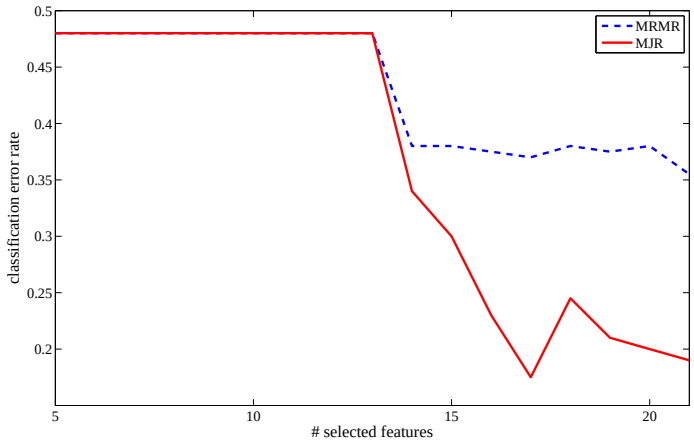

Figure 1: SVM classification error for 3-level quantization of input space.

(since each level happens too many times and we have small number of probabilities to estimate). In the second scenario, the input features are quantized separately into 5 levels. The later scenario corresponds to a small training set regime where there are a large number of probabilities to estimate.

Figure 1 compares the misclassification error of MRMR and MJR feature selection algorithms for different number of features. Here, SVM is used as the classification method and the input space is quantized into 3 levels. Since this scenario corresponds to large training set regime, the MJR outperforms MRMR as depicted in the figure.

In Fig. 2, the SVM misclassification error of MJR and AMJR has been compared for different number of selected features. Here, the input space is quantized into 5 level which corresponds to the small training set scenario. As depicted in this figure, MJR has unstable performance in this scenario while AMJR shows stable and better performance. This figure confirms our theoretical analysis of instability of MJR and shows that our proposed method (AMJR) removes the instability problem almost completely.

The advantage of the proposed method AMJR method is further described in Fig. 3. In this figure, the SVM misclassification error of AMJR and MRMR methods are compared for differnt number of selected features. Here, the input space is quantized into 5 levels (small training set regime). As depicted in this figure, AMJR substantially outperforms MRMR for any number of

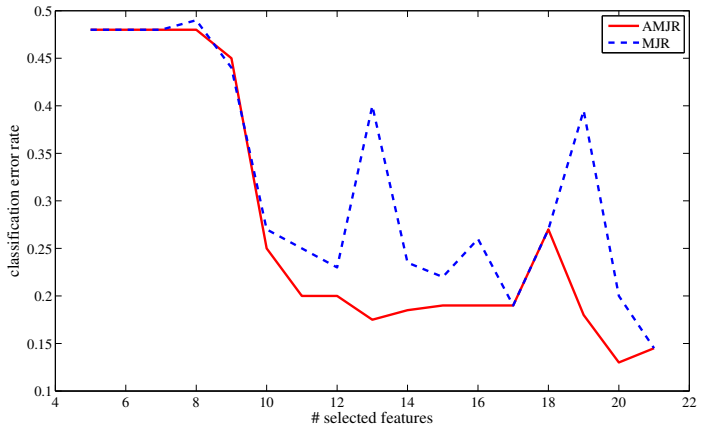

Figure 2: SVM classification error for 5-level quantization of input space.

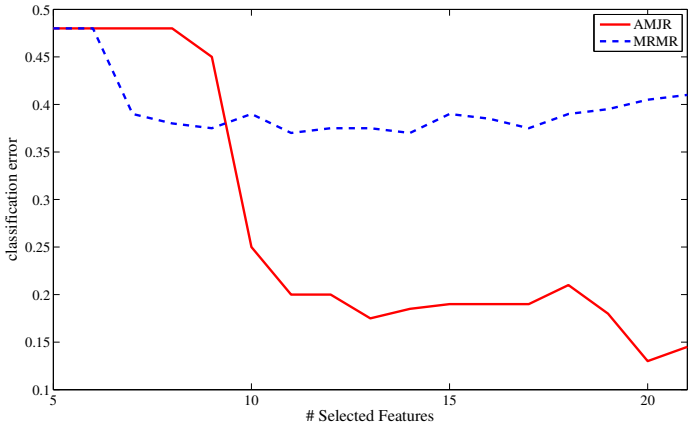

Figure 3: SVM classification error for 5-level quantization of input space.

selected features.

It worth mentioning that other than SVM, we have also repeated the above experiments for logistic regression and classification trees and the same relative results were obtained. Since our focus is on comparing the feature selection algorithms (and not the classification methods), and also due to the lack of space, the results for these methods are not provided here.

5 Conclusion

Feature selection is an indispensable part of solution when dealing with high-dimensional datasets. One powerful tool to address this problem is mutual information. A common approach is to use Maximum Relevance Minimum Redundancy (MRMR) approach to solve the feature selection problem. In this paper, based on insight from information theory, a new objective function is used. Also, a novel mutual information estimator is used enabling us to discretize the data into finer levels. Combining the novel mutual information estimator with the new objective function, an error rate 3 times lower than that of MRMR is demonstrated.

References

[1] T. Cover, and J. Thomas. "Elements of information theory", John Wiley & Sons, 2012.

[2] H. Peng, H. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." Pattern Analysis and Machine Intelligence, IEEE Transactions on 27.8, 2005, 1226-1238.

[3] H. Yang, and J. Moody. "Data Visualization and Feature Selection: New Algorithms for Non-gaussian Data." NIPS. 1999.

[4] J. Jiao, K. Venkat, Y. Han, T. Weissman, "Minimax Estimation of Functionals of Discrete Distributions", available on arXiv. 2014.

[5] Available online: http://www.nipsfsc.ecs.soton.ac.uk/datasets