ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

On the Potential Inadequacy of Mutual Information for Feature Selection

Benoît Frénay, Gauthier Doquire and Michel Verleysen *

Université catholique de Louvain - ICTEAM/ELEN - Machine Learning Group
Place du Levant 3, 1348 Louvain-la-Neuve - Belgium

Abstract. Despite its popularity as a relevance criterion for feature selection, the mutual information can sometimes be inadequate for this task. Indeed, it is commonly accepted that a set of features maximising the mutual information with the target vector leads to a lower probability of misclassification. However, this assumption is in general not true. Justifications and illustrations of this fact are given in this paper.

1 Introduction

For a lot of machine learning and data mining applications, feature selection is a task of major importance. In particular, many regression or classification algorithms perform particularly bad when faced to high-dimensional data, due to the so-called *curse of dimensionality.* By reducing the dimensionality of the dataset while preserving the original features (by opposition to projection techniques), feature selection allows building efficient and easy-to-interpret models.

Filter methods, which are based on a statistical criterion to evaluate the relevance of a set of features, are often used in practice; this is mainly due to their low computational cost and their independence from any prediction model, in comparison to wrapper approaches which directly optimize the performances of a specific prediction model. Indeed, filter methods can be used prior to the construction of any prediction model.

As it is well-known, the mutual information (MI) [1] is a quantity measuring the dependency between two (groups of) random variables. Many reasons detailed below, including bounds relating it to the probability of classification error, made the MI criterion very popular for filter based feature selection [2]. However, despite its popularity, there exists a significant number of problems for which the MI should probably not be the criterion of choice. Indeed, the subset of features maximising the MI with a target class vector may not always minimise the probability of misclassification, which is often the final quantity of interest in real-world applications. The objective of the paper is to clearly point out and illustrate this fact. A sufficient condition for the MI criterion to be relevant for a certain problem is also given.

Section 2 briefly recalls basic notions about MI and presents the reasons why it is popular for feature selection. In Section 3, the possible inadequacy of the MI for this task is discussed, the potential problems are illustrated and a sufficient condition for optimality is given. Section 4 concludes the work.

*Gauthier Doquire is funded by a Belgian F.R.I.A grant.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2 Mutual Information

This section introduces mutual information in the context of feature selection.

2.1 Basic Definitions

Shannon's mutual information (MI) [1] measures the dependency between two discrete random variables X and Y. If X (resp. Y) takes on n_X (n_Y) possible values x_i (y_i) with probability $P(X = x_i)$ ($P(Y = y_i)$), MI is defined as

$$I(X;Y) = \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(X = x_i, Y = y_j) \log_2 \frac{P(X = x_i, Y = y_j)}{P(X = x_i)P(Y = y_j)} \tag{1}$$

where $P(X, Y)$ is the joint probability of X and Y. Equation (1) can be seen as the Kullback-Leibler divergence [1] between $P(X)\,P(Y)$ and the joint probability $P(X, Y)$. Knowing that the entropy of a discrete random variable is

$$H(X) = -\sum_{i=1}^{n_x} P(X = x_i) \log_2 P(X = x_i), \tag{2}$$

it is possible to show [1] from Eq. (1) that the MI can be rewritten as

$$I(X;Y) = H(Y) - H(Y|X) \tag{3}$$

with $H(Y|X)$ being the conditional entropy of Y given X. Similar definitions can derived for continuous variables, the sums being then replaced by integrals.

2.2 Use for Feature Selection

Since the seminal paper of Battiti [2], the MI criterion has been used extensively for filter feature selection as it possesses many desirable properties for this task.

First, as detailed in [2], the MI has a natural interpretation in terms of uncertainty reduction. Indeed, it is well known that the entropy of a random variable measures the uncertainty on the values taken by this variable. Let Y be a target class vector and X a (set of) feature(s). Equation (3) translates the fact that $I(X;Y)$ is the reduction of uncertainty about the value of Y once X is known; this appears to be a natural criterion for feature selection. Equation (1) can also be interpreted in the same way. If X and Y are independent, $P(X, Y) = P(X)P(Y)$ and the MI is zero. On the contrary, as the dependency between X and Y grows so does the divergence (1) and thus the MI.

Then, it is also stressed in [2] that MI has the advantage over other popular criteria (such as the correlation coefficient) that it is able to detect non-linear relationships between variables. Moreover, the MI criterion can naturally be defined for multivariate random variables, which again is not true for correlation. This property is of fundamental importance if greedy search procedures (such as forward or backward) have to be used to construct the feature subset.

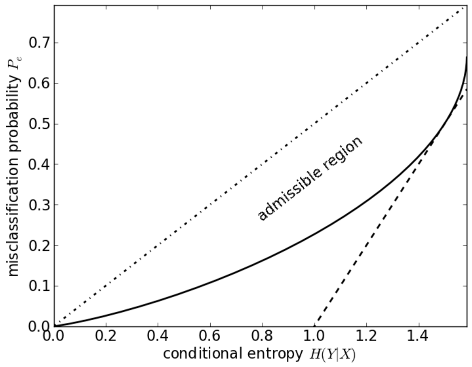

Figure 1: Weak Fano bound (dashed line), strong Fano bound (plain line) and Hellman-Raviv bound (dash-dotted line) on the probability of misclassification P_e for an optimal classifier with three balanced classes ($n_Y = 3$), with respect to the conditional entropy $H(Y|X)$. This figure is inspired by [4, 6].

Eventually, the use of MI is also supported by the existence of bounds relating the probability of misclassification P_e for an optimal classifier to the conditional entropy $H(Y|X)$. More specifically, Fano [3] derived two lower bounds on P_e. The weak Fano bound states that

$$H(Y|X) \leq 1 + P_e \log_2(n_Y - 1) \tag{4}$$

where n_Y is the number of classes, whereas the strong Fano bound is

$$H(Y|X) \leq H(P_e) + P_e \log_2(n_Y - 1). \tag{5}$$

The two above upper bounds on $H(Y|X)$ can be inverted to obtain lower bounds on P_e. It is important to notice that the weak bound (4) is useless in binary classification problems, since it cannot be inverted to get a lower bound on P_e when $n_Y = 2$. Moreover, the bound (4) is generally much looser than the bound (5), especially if P_e is small, which is precisely the situation of interest for classifier design [4]. However, the strong bound (5) on P_e is less easy to manipulate in practice since it has no closed-form and must be solved numerically. An upper bound on P_e is also given by the Hellman-Raviv inequality [5]

$$P_e \leq \frac{1}{2} H(Y|X). \tag{6}$$

As can be seen in Figure 1 inspired by [4, 6], decreasing the conditional entropy decreases both the upper and the lower bound on P_e, motivating the use of this criterion for feature selection. Since $H(Y)$ is a constant value for a given classification problem, Equations (4), (5) and (6), together with Equation (3), also give a justification to the maximisation of the MI for feature selection.

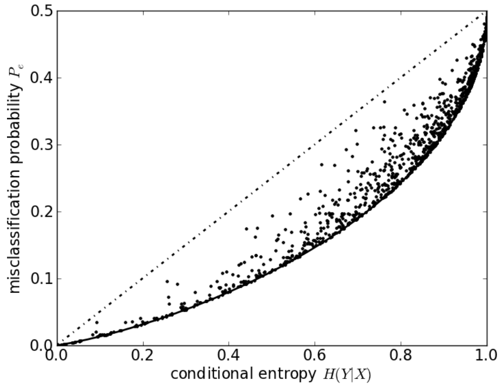

Figure 2: Examples of pairs $\langle H(Y|X), P_e \rangle$ corresponding to random binary classification problems with two binary features. The strong Fano bound (plain line) and the Hellman-Raviv bound (dash-dotted line) on P_e are shown.

3 Potential Inadequacy of Mutual Information

As mentioned in Section 2, the actual goal in many applications is to minimise the probability of misclassification. In other words, the utility of a feature subset can be quantified using P_e, which is a tight lower bound for the misclassification probability of any classifier. Using Figure 1 and Equations (3), (4), (5) and (6), several papers [4, 6, 7] conclude (i) that minimising MI is equivalent to minimising P_e and (ii) that MI can therefore be used equivalently for feature selection. This section shows that both claims are not necessarily true.

3.1 Relationships Between Misclassification Probability and Entropy

Figure 2 shows (i) the strong Fano bound and the Hellman-Raviv bound for P_e in terms of $H(Y|X)$ and (ii) several examples of actual pairs $\langle H(Y|X), P_e \rangle$. The pairs correspond to random binary classification problems with two binary features. For each problem, both P_e and $H(Y|X)$ are computed exactly, which is possible since all necessary probabilities are known. The problems are drawn as follows: (i) the values $P(Y = y)$ and $P(X = x|Y = y)$ are randomly drawn from the uniform distribution $\mathcal{U}(0, 1)$, (ii) these values are normalised to enforce $\sum_y P(Y = y) = 1$ and $\sum_x P(X = x|Y = y) = 1$ for each y and (iii) probabilities $P(X)$ and $P(Y|X)$ are computed using marginalisation and Bayes' theorem.

Figure 2 shows that the pairs $\langle H(Y|X), P_e \rangle$ are scattered between the strong Fano lower bound and the Hellman-Raviv upper bound. Moreover, it is possible to find two pairs such that the entropy $H(Y|X)$ decreases and the probability of misclassification P_e increases (and *vice versa*). In other words, contrary to what is often claimed, it is not sufficient to reduce $H(Y|X)$ in order to reduce P_e. It suggests that minimising MI may not be sufficient, which is illustrated below.

3.2 Illustration of Mutual Information Failure for Feature Selection

Let us now review a simple, artificial example of mutual information failure. In a context of disease diagnosis, two classes are distinguished with prior

$$P(Y) = \begin{pmatrix} 0.316 & 0.684 \end{pmatrix} \tag{7}$$

where columns correspond to possible values of $Y \in \{0, 1\}$. Furthermore, two tests are available to classify a new patient, whose binary outcomes are denoted $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1\}$. However, the practician can only perform one of these tests. In terms of feature selection, he has to select the best feature.

Through experimentation, the practician discovers that the conditional distributions of X_1 and X_2 with respect to Y are

$$P(X_1|Y) = \begin{pmatrix} 0.417 & 0.104 \\ 0.583 & 0.896 \end{pmatrix} \text{ and } P(X_2|Y) = \begin{pmatrix} 0.991 & 0.479 \\ 0.009 & 0.521 \end{pmatrix} \tag{8}$$

where rows correspond to values of X_i and columns correspond to values of Y. Hence, using marginalisation and Bayes' theorem, one obtains the posteriors

$$P(Y|X_1) = \begin{pmatrix} 0.649 & 0.231 \\ 0.351 & 0.769 \end{pmatrix} \text{ and } P(Y|X_2) = \begin{pmatrix} 0.489 & 0.008 \\ 0.511 & 0.992 \end{pmatrix} \tag{9}$$

where rows correspond to values of Y and columns correspond to values of X_i. On one hand, the test with outcome X_1 allows discriminating between both classes, but there is an important error probability ($P_e = .351$ if $X_1 = 0$ and $P_e = .231$ if $X_1 = 1$). On the other hand, the test with outcome X_2 allows discriminating almost perfectly when it is positive ($P_e = .008$ if $X_2 = 1$), but it is almost useless when it is negative ($P_e = .489$ if $X_2 = 0$).

Using the first test, one obtains $P_e = 0.255$ and $I(X_1; Y) = 0.089$. Using the second test, one obtains $P_e = 0.316$ and $I(X_2; Y) = 0.236$. Here, the MI is significantly larger using X_2. However, P_e is also larger, which means that selecting X_2 based on mutual information leads here to an increase in error. This phenomenon is not rare: using pairs of random problems drawn as explained in the previous subsection, about 20% of the pairs violate the common belief that increasing mutual information decreases the misclassification probability.

Figure 3 illustrates the example. Each pair $\langle H(Y|X), P_e \rangle$ stands between the Fano and Hellman-Raviv bounds. It is clear that $I(X_2; Y) = H(Y) - H(Y|X_2)$ is larger than $I(X_2; Y) = H(Y) - H(Y|X_2)$, whereas $P_e(X_2)$ is larger than $P_e(X_1)$.

3.3 Conditions of Optimality

According to the above discussion, mutual information seems to be more a heuristic than a never-failing criterion. However, it is possible to guarantee whether MI is valuable or not. Indeed, let us define two feature subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ which must be compared. If the value of the Hellman-Raviv bound for $\mathcal{X}_1$ is smaller than the value of the strong Fano bound for $\mathcal{X}_2$, then an increase in mutual information always leads to a decrease in misclassification probability.

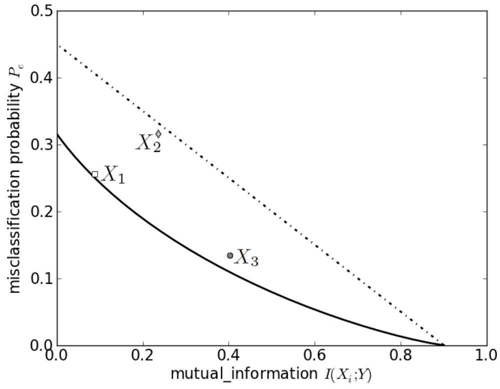

Figure 3: Example of mutual information failure for feature selection, with the strong Fano bound (plain line) and the Hellman-Raviv bound (dash-dotted line).

In the above example, the Fano bound for X_1 is $P_e \geq .250$, whereas the Hellman-Raviv bound for X_2 is $P_e \leq .332$. Here, it is not possible to guarantee that mutual information is a relevant criterion to choose between X_1 and X_2. Figure 3 also shows an other candidate X_3 for which the Hellman-Raviv bound is $P_e \leq .249$. Here, the new feature X_3 is guaranteed to be a better choice.

4 Conclusion

This paper shows that mutual information is not necessarily an optimal criterion to select features, if the actual goal is to achieve minimal probability of misclassification. The behaviour of mutual information is described and related to Fano and Hellman-Raviv bounds. An example of MI failure is given, which shows that increasing MI can sometimes increase the misclassification probability as well.

References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 99th edition, August 1991.

[2] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5:537–550, 1994.

[3] R. Fano. *Transmission of Information: A Statistical Theory of Communications.* The MIT Press, Cambridge, MA, 1961.

[4] J. W. Fisher, M. Siracusa, and T. Kihn. Estimation of signal information content for classification. In *Proceedings of DSP/SPE 2009*, 2009.

[5] M. E. Hellman and J. Raviv. Probability of error, equivocation and the chernoff bound. *IEEE Transactions on Information Theory*, 16:368–372, 1970.

[6] G. Brown. An information theoretic perspective on multiple classifier systems. In *Proceedings of MCS 2009*, pages 344–353, Berlin, Heidelberg, 2009. Springer-Verlag.

[7] U. Ozertem, D. Erdogmus, and R. Jenssen. Spectral feature projections that maximize Shannon mutual information with class labels. *Pattern Recogn.*, 39:1241–1252, July 2006.