A Constructive Density-Ratio Approach to Mutual Information Estimation: experiments in feature selection

Igor Braga

Institute of Mathematics and Computer Science
University of São Paulo, São Carlos-SP, Brazil
igorab@icmc.usp.br

Abstract. Mutual Information (MI) estimation is an important component of several data mining tasks (*e.g* feature selection). In classification settings, MI estimation essentially depends on the estimation of the ratio of two probability densities. Using a recently developed method of density-ratio estimation, which is constructive in nature, new estimators for MI can be derived. In this article, we consider one such new estimator — VMI — and compare it experimentally to previously proposed MI estimators. The first batch of experiments is conducted solely on mutual information estimation, and shows that VMI compares favorably to previous estimators. The second batch of experiments applies MI estimation to feature selection in classification tasks, evidencing that VMI leads to better feature selection performance. Combining the results of both experimental batches, we conclude that the development of improved density-ratio estimators can positively impact MI estimation and feature selection.

Categories and Subject Descriptors: I.5.1 [**Pattern Recognition**]: Models—*Statistical*; H.1.1 [**Models and Principles**]: Systems and Information Theory; G.1.9 [**Numerical Analysis**]: Integral Equations—*Fredholm equations*

Keywords: classification, density-ratio estimation, mutual information estimation, feature selection

1. INTRODUCTION

The mutual information between two random vectors X and Y is one of the most important concepts in Information Theory [Cover and Thomas 2006]. For continuous X and Y, it is often written as

$$I(X,Y) = \int_X \int_Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, dx \, dy, \tag{1}$$

where $p(x,y)$ is the joint probability density function of X and Y, and $p(x)$ and $p(y)$ are the marginal density functions associated with X and Y (resp.). Intuitively, mutual information measures how much information is shared by X and Y: if they are independent of each other, $p(x,y) = p(x)p(y)$ and then $I(X,Y) = 0$. On the other hand, if X and Y are the same random vector, the value of the mutual information achieves its upper bound — the differential entropy of X (or Y).

Mutual information plays an important role in data mining tasks like feature selection [Guyon and Elisseeff 2003] and Independent Component Analysis [Hyvärinen and Oja 2000]. For these tasks, it is typical for the distributions involved in MI calculation to be unknown. This way, it becomes important to develop methods of mutual information estimation using data sampled from these unknown distributions [van Hulle 2005; Kraskov et al. 2004; Suzuki et al. 2009].

This work is supported by grant #2009/17773-7, São Paulo Research Foundation (FAPESP). The author is grateful to Dr. Vladimir Vapnik and Rauf Izmailov for having introduced him to the subject of this article. He also thanks: Laís Pessine do Carmo and Prof. Maria Carolina Monard, for their help in setting this paper up; and the anonymous reviewers, for their valuable suggestions.

When Y is a categorical variable, the estimation of the mutual information essentially depends on the estimation of a finite number of ratios of probability densities [Sugiyama et al. 2011; Vapnik et al. 2014]. Previous work [Suzuki et al. 2009] has already attempted to cast MI estimation as density-ratio estimation. Still, the approach taken in this work is new. We consider an estimator — namely VMI — that differs from previous estimators in two aspects. The first one is in the form of the MI estimator itself, which is more robust. The other one is related to the method of density-ratio estimation employed in VMI, which can be proven to construct the real density ratio with high probability.

In this article, we experimentally evaluate this new approach to mutual information estimation. We first consider a set of synthetic two-dimensional models for which the real value of the mutual information is known. Since evaluation on real data is also desirable, we conduct a second batch of experiments considering the task of feature selection for classification.

The analysis of the results allows us to draw interesting conclusions about mutual information estimation and feature selection. Regarding MI estimation alone, the results on synthetic models corroborate the theoretical advantages of the new approach, as VMI outperformed other estimators. The feature selection experiment, in its turn, evidences that using VMI in a MI-based feature selection scheme improves upon the use of other MI estimators. Altogether, these results point to unexplored opportunities for improving MI estimation and feature selection by improving density-ratio estimation.

The remainder of this article is organized as follows. Section 2 describes the approach to mutual information estimation taken in this work as well as previous ones. As the new approach depends on density-ratio estimation, this topic is covered in Section 3. Section 4 reviews the Joint Mutual Information (JMI) feature selection method [Yang and Moody 1999] used in our experiments. Section 5 is devoted to the experimental evaluations. Section 6 concludes with the findings of this work and indications of future research.

2. MUTUAL INFORMATION ESTIMATION

The estimation of the mutual information $I(X,Y)$ based on a sample $(x_1,y_1),\ldots,(x_n,y_n) \overset{i.i.d}{\sim} p(x,y)$ is a long-standing problem in applied statistics. Several attempts of solving this problem have been made by considering the equality

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{2}$$

That is, $I(X,Y)$ is estimated by first estimating the differential entropies $H(X)$, $H(Y)$, and $H(X,Y)$, and then plugging these estimates into Expression (2). A potential problem with the entropy approach is that the errors in the estimation of the individual entropies do not necessarily cancel out.

Two popular entropy estimators are the non-parametric k-NN estimator [Kraskov et al. 2004] and the parametric *Edgeworth*-expansion estimator [van Hulle 2005]. These estimators have some drawbacks of their own. To wit, there is no systematic way of selecting the best value of k in the k-NN estimator. Moreover, the Edgeworth estimator is based on the assumption that the densities $p(x,y)$, $p(x)$, and $p(y)$ are each normally distributed, which is often not satisfied in practice.

Motivated by the problems of these entropy estimators and the indirect nature of MI estimation through entropy estimation, a direct estimator of $I(X,Y)$ was proposed in [Suzuki et al. 2009]. The resulting estimator replaces Expression (1) by its empirical average

$$\widehat{I}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i,y_i)}{p(x_i)p(y_i)}. \tag{3}$$

Since the value of the ratio inside the log function is unknown, the authors proposed the KLIEP method [Sugiyama et al. 2008] for estimating this ratio using a sample drawn from $p(x,y)$. The resulting estimator was named *Maximum Likelihood Mutual Information* (MLMI).

In this work, we put forward a different direct approach to mutual information estimation. For the case where Y is discrete (classification), we will also stumble upon the unavoidable task of estimating the ratio of two probability densities. The difference from previous work is that we use an improved mutual information estimator and a constructive method of density-ratio estimation.

The reader might be wondering why do we need yet another MI estimator. The answer is that the estimator in Expression (3) will be very susceptible to errors in the estimation of the involved ratios, since $log(z)$ goes to $-\infty$ very fast when $z \to 0$ — Figure 1. Any estimation error of the ratios that may happen in this direction will be greatly magnified, causing a large error in MI estimation. Fortunately, there is an equivalent formulation for MI [Vapnik et al. 2014] which provides a more robust estimator by considering the better behaved function $z\,log(z)$ — Figure 1. Even though $z\,log(z)$ goes to ∞ faster than $log(z)$ when $z \to \infty$, the value of z will be bounded from above in MI estimation.

In order to arrive at this improved estimator, let us first rewrite $I(X,Y)$ as

$$I(X,Y) = \int_X \int_Y \frac{p(x,y)}{p(x)p(y)} \log \frac{p(x,y)}{p(x)p(y)}\, p(x)p(y)\, dx\, dy = E_X E_Y [r(x,y) \log r(x,y)], \tag{4}$$

where E is the expectation operator and $r(x,y) = \frac{p(x,y)}{p(x)p(y)}$ is a density-ratio function. Whenever Y takes only on a finite number of values $\{a_1, \ldots, a_m\}$, Expression (4) is written as

$$I(X,Y) = \sum_{i=1}^{m} p(a_i)\, E_X [r(x,a_i) \log r(x,a_i)]. \tag{5}$$

In this case, $r(x,a_i) = \frac{p(x|a_i)}{p(x)}$ can be considered as a density-ratio function that depends only on x.

Therefore, the problem of estimating the mutual information in classification settings is equivalent to estimating the value $I(X,Y)$ in Expression (5) when the densities $p(x,y)$, $p(x)$, and the probability $p(y)$ are unknown but a sample $(x_1,y_1),\ldots,(x_n,y_n) \overset{i.i.d}{\sim} p(x,y)$ is available. Denoting by n_i the number of elements from class a_i and considering that $n = n_1 + \ldots + n_m$, the probabilities $p(a_i)$ can be readily estimated by $p_i = \frac{n_i}{n}$. Using the values p_i and approximating the expectation in Expression (5) by its empirical average, we arrive at the following estimator for $I(X,Y)$

$$\widehat{I}(X,Y) = \frac{1}{n} \sum_{i=1}^{m} p_i \sum_{j=1}^{n} r(x_j,a_i) \log r(x_j,a_i). \tag{6}$$

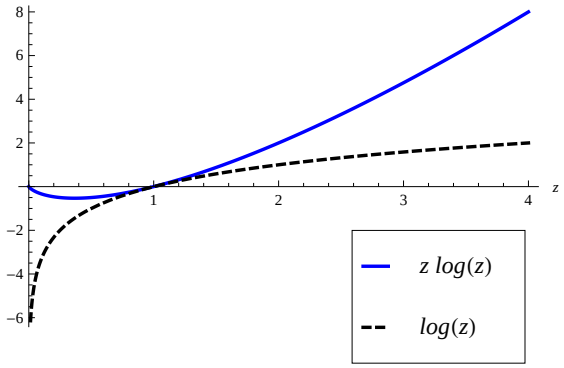


Fig. 1. Plot of functions $log(z)$ and $z\,log(z)$. In MI estimation, z will be an estimated ratio value. Hence, the estimated value of $z\,log(z)$ is less susceptible to estimation errors than $log(z)$ (see the explanation in the above paragraph).

When the n values of the m different density ratios $r(x_j, a_i)$ are known, the consistency of $\widehat{I}(X,Y)$ is guaranteed by the law of large numbers as $n \to \infty$. As these ratios are not known in advance, we show in the next section a constructive method for estimating them from data. Note that, when only two classes exist, just n values $r(x_1, a_1), \ldots, r(x_n, a_1)$ need to be estimated, since the other n values $r(x_1, a_2), \ldots, r(x_n, a_2)$ can be computed from the first ones by applying the law of total probability.

3. CONSTRUCTIVE DENSITY-RATIO ESTIMATION

Several settings for the problem of density-ratio estimation have been proposed [Sugiyama et al. 2011]. Here we focus on a recently developed one, which is distinguished by being constructive, *i.e* for an increasing amount of data, the method provides solutions that converge in probability to the real density ratio *regardless* of the choice of metric used for evaluating the distance between the solutions and the real density ratio. For brevity, we omit the derivations. The reader is referred to [Vapnik et al. 2014] for details. Hereafter, we consider a random vector $X = (X^1, \ldots, X^d)$, although the notation used will be that of random variables.

In the realm of mathematical statistics, the density function $p(x)$ of X (if it exists) is defined as the derivative of the cumulative distribution function $P(x)$ of X: $p(x) = \frac{dP(x)}{dx}$. Let us consider two probability densities $p(x)$ and $q(x)$. When $q(x) > 0$, the density-ratio function between $p(x)$ and $q(x)$ is defined as

$$r(x) = \frac{dP(x)/dx}{dQ(x)/dx} = \frac{p(x)}{q(x)}. \tag{7}$$

From the first equality in Expression (7), the problem of estimating the density ratio from data is the problem of solving the integral equation

$$\int_{-\infty}^{x} r(t)\, dQ(t) = P(x) \tag{8}$$

when the distribution functions $P(x)$ and $Q(x)$ are unknown but samples $x_1, \ldots, x_\ell \overset{i.i.d}{\sim} P(x)$ and $x'_1, \ldots, x'_n \overset{i.i.d}{\sim} Q(x)$ are given. The constructive setting of this problem is to solve Expression (8) using the empirical (multidimensional) cumulative distribution functions[1]

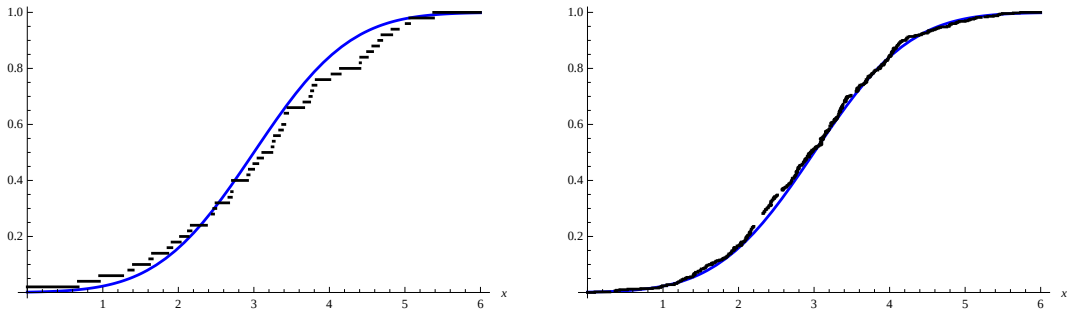
$$P_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \prod_{k=1}^{d} \theta(x^k - x_i^k) \quad \text{and} \quad Q_n(x) = \frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{d} \theta(x^k - x_i^{k'}),$$

instead of the actual cumulative distributions $Q(x)$ and $P(x)$. It is known that any cumulative distribution is well-approximated by the empirical cumulative distribution function and that fast convergence takes place [Vapnik 1998, Section 4.9.3] — Figure 2.

Solving the integral equation in Expression (8) using approximations to its right hand side and to its integral operator is *ill-posed* [Vapnik 1998, Section 1.12], which means that these approximations can lead to large deviations in the final solution $r(x)$. In order to solve it properly, the regularization method must be used. In accordance with this method, the following minimization problem, parameterized by $\gamma > 0$, can be considered for obtaining estimates of $r(x)$ at the "denominator points" $x'_1, \ldots, x'_n$ [Vapnik et al. 2014]:

$$\underset{r(x'_1),\ldots,r(x'_n)}{\arg\min} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} r(x'_i)\, \theta(x - x'_i) - \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i) \right\|_{L^2}^2 + \gamma \int r(t)^2\, dQ_n(t) \right]. \tag{9}$$

[1] Step-function $\theta(t)$ is defined as $\theta(t) = \begin{cases} 1, & \text{if } t \geq 0 \\ 0, & \text{otherwise.} \end{cases}$



(a) 50-point ECDF (b) 500-point ECDF

Fig. 2. Cumulative distribution functions (in blue) of a univariate Gaussian distribution and its empirical distribution (ECDF, step function in black) based on points drawn from the same distribution.

Denoting by $\vec{r}$ the $n \times 1$ vector $[r(x'_1), \ldots, r(x'_n)]^\top$ and by $\vec{1}$ the $\ell \times 1$ vector $[1, \ldots, 1]^\top$, the optimization problem in Expression (9) has the following form in vector-matrix notation

$$\arg\min_{\vec{r}} \left[\frac{1}{2}\, \vec{r}^\top V'' \vec{r} - \frac{n}{\ell}\, \vec{r}^\top V' \vec{1} + \frac{\gamma}{n}\, \vec{r}^\top \vec{r} \right]. \tag{10}$$

This density-ratio estimation method is termed DRE-V. The elements of the matrices V'' and V' come from the expansion of the first term (norm) in Expression (9), and are computed from sampled data. For a fixed value of γ, the optimization problem in Expression (10) can be constrained to take into account the positivity of $\vec{r}$, in which case a standard quadratic optimization routine can be used to solve it. For the unconstrained problem, the minimum of this functional may be computed faster by just solving a system of linear equations ($O(n^3)$ time complexity in the worst case for a dense system).

Obtaining good estimates of the density ratio using finite samples depends on the proper selection of the regularization parameter γ. In DRE-V, this selection is carried out by cross-validation on several candidate values of γ [Vapnik et al. 2014, Section 7]. A special feature of the unconstrained optimization problem in Expression (10) is that a leave-one-out cross-validation procedure can be leveraged with the same computational complexity of solving the problem for a single value of γ, for this optimization problem has the same structure as that of the Regularized Least-Squares method [Rifkin 2006]. For the experiments in Section 5, we exploit this special feature for selecting γ. The constrained problem is used only to obtain the final solution.

From now on, we employ the name VMI to refer to the method of MI estimation that uses DRE-V to estimate the density ratios in Expression (6). This estimator has some advantages over the Edgeworth and k-NN estimators, for it is non-parametric and its parameter γ can be optimized using the available data. Moreover, VMI uses a method of density-ratio estimation that was experimentally shown to outperform KLIEP [Vapnik et al. 2014], the latter being the estimator used in MLMI.

4. FEATURE SELECTION BASED ON MUTUAL INFORMATION

In this section, we describe how mutual information estimation can be used as a component of a feature selection scheme in classification tasks. Remind that in feature selection the goal is to use only a fraction of the original d features used to describe the training set $(x_1, y_1), \ldots, (x_n, y_n)$, $x_i = [x_i^1, \ldots, x_i^d]$. The need for such procedure may have different reasons, the most common ones being: 1) only a small portion of the features are relevant to discriminate the classes; or 2) too many features are available, rendering the training phase of a classifier computationally unfeasible.

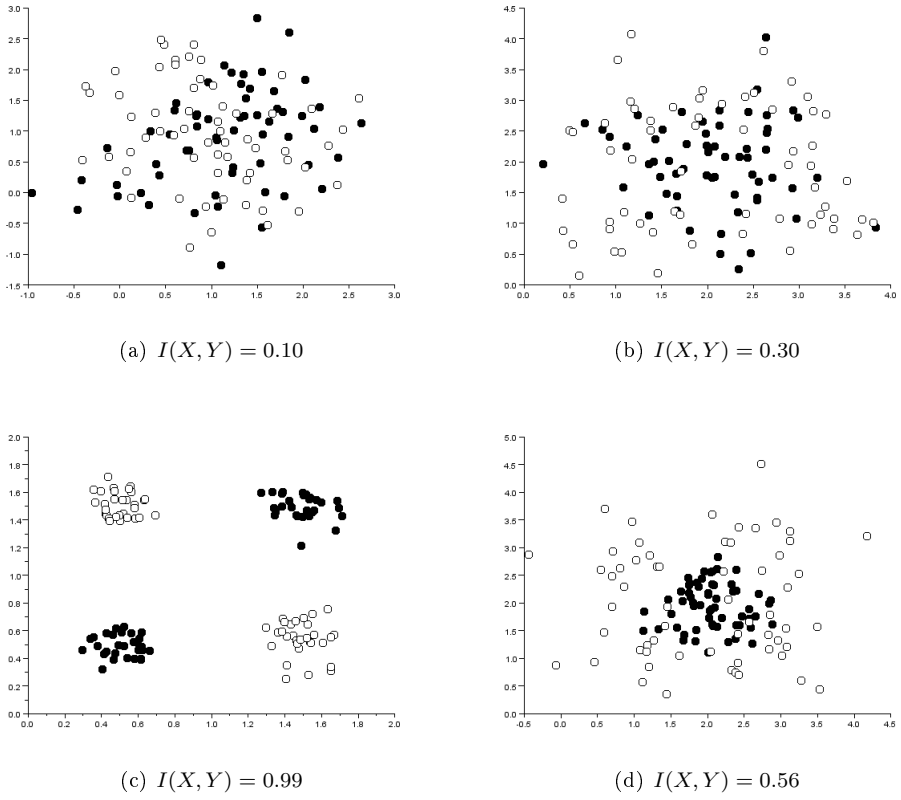
From the *theoretical* point of view, this problem breaks down into two stages:

(1) Among the original d features, select the k features that provide the largest mutual information towards the target variable Y;
(2) Train a classifier using the training set restricted to the k selected features.

Taking mutual information as a criterion of feature set importance is justified for it bounds the probability of erroneous classification of the optimal decision rule — the so-called *Bayes function*. A higher value of $I(X,Y)$ implies a smaller probability of erroneous classification of the optimal decision rule. To illustrate this relationship, consider Figure 3. Each one of the four scatter plots depicts two-class data points sampled from a different artificial two-dimensional feature model. Since the model used to generate the data is known, the mutual information $I(X,Y) = I([X^1, X^2], Y)$ can be calculated in each case. This illustration shows that, as mutual information increases, the clearer is the separation between the points from the two classes.

Given enough time and data, it would be possible to estimate mutual information for each combination of d features taken k at a time and, afterwards, choose the k-combination with the largest estimated value of MI. However, this is unfeasible in practice. Data is usually scarce, which impairs k-dimensional density-ratio estimation when k is large. In addition, even if we had enough data for estimating the density ratio for a single combination of k features, the time it would take to run this procedure for all possible combinations of features is prohibitive for typical[2] values of k and d.

[2]For instance, for $k = 4$ and $d = 100$, around 4 million combinations should be investigated. The number goes up to around 400 billion combinations if $k = 4$ and $d = 1000$.



(a) $I(X,Y) = 0.10$ (b) $I(X,Y) = 0.30$

(c) $I(X,Y) = 0.99$ (d) $I(X,Y) = 0.56$

Fig. 3. Mutual information and a data sample of 120 points divided into two classes (circle/solid circle) for different two-dimensional models.

Thus, the problem of feature selection as posed in this section requires a lot of engineering, in which restrictions imposed by the real world play a prominent role. All sorts of *heuristics* have been employed over the years to deal with this problem [Guyon and Elisseeff 2003]. Some of them may not even have an explicit connection to mutual information. In this work, we employ a heuristic that does not give up on mutual information entirely: it tries to estimate mutual information for low-order combinations of features (say, pairs of features), and, afterwards, selects k features based on these estimates. In [Brown et al. 2012], several methods following this heuristic are investigated. The conclusion was that the Joint Mutual Information (JMI) method [Yang and Moody 1999] was one of the best methods of this category. Hence, we use JMI in our feature selection experiments described in the next section.

The JMI procedure — which is a kind of forward selection — goes as follows. After the value of mutual information $I([X^i, X^j], Y)$ is estimated for every pair of features (i, j) using the training set $(x_1, y_1), \ldots, (x_n, y_n)$, the pair with the largest $I([X^i, X^j], Y)$ is singled out to compose the initial set of selected features S. If $k > 2$, the method iterates over the remaining features, adding to S the feature j that maximizes

$$\sum_{i \in S} I([X^i, X^j], Y).$$

The running time of this procedure is dominated by the estimation of the mutual information for each pair of features. As datasets with 10000 features figure in our experimental evaluation, the whole procedure would be very time consuming. In these cases only, we rely on one-dimensional mutual information estimation to reduce our working set of features. More precisely, we restrict the application of JMI to the 50 features with the largest value of the estimated $I(X^i, Y)$.

5. EXPERIMENTS

The present experimental evaluation was conducted to guide the evaluation of the new approach of mutual information estimation based on both Expression (6) and the constructive density-ratio estimation method. For this, we use the simplest constructive density-ratio estimator: DRE-V [Vapnik et al. 2014]. For comparison, we evaluate our approach against the popular k-NN and Edgeworth (EDGEW) approaches and the MLMI [Suzuki et al. 2009] estimator.

First, we conducted experiments in 10 synthetic two-dimensional datasets. The 2D estimation case is important, since it provides the basis for a number of existing feature selection algorithms, as mentioned in Section 4. Using synthetic data allows us to compare the obtained estimates with the real value of mutual information. Due to lack of space, a detailed description of these synthetic models is provided online[3]. We consider datasets of size $n = 40, 100, 200$ and 400. For each sample size, 20 different samples are used for estimating $I(X, Y)$. The relative estimation error $\frac{\hat{I} - I_{real}}{I_{real}}$ of each method is averaged over these 20 samples.

Table I reports the results of these experiments. In Table II, we present a summary for each method of how many times their average MI estimation error was the smallest for a given dataset and sample size, and also how many times it exceeded 50%. Note that VMI achieves the highest counts in "Best" and the lowest in "50%", with the exception of sample size 40, where all methods tend to achieve poor results. The MLMI method delivers the worst results in most cases. The Edgeworth estimator starts delivering reasonable results only when the sample size reaches 400. Finally, the k-NN estimators deliver very similar performances for the tested values of k. It is worth remembering that, even if the k-NN estimator could obtain better performance with a larger value of k, there is no way of selecting the best value of k in mutual information estimation.

[3] `http://sites.google.com/site/igorabmi/`

Table I. Mean and standard deviation of the relative mutual information estimation error $\frac{\hat{I}-I_{real}}{I_{real}}$

Model	Real MI	n	VMI	MLMI	EDGEW	3-NN	5-NN	7-NN
1	0.10	40	0.10 (0.11)	0.08 (0.06)	0.08 (0.11)	0.20 (0.10)	0.17 (0.14)	0.21 (0.13)
		100	0.03 (0.02)	0.14 (0.19)	0.05 (0.03)	0.10 (0.04)	0.10 (0.08)	0.13 (0.10)
		200	0.03 (0.02)	0.10 (0.04)	0.04 (0.02)	0.08 (0.06)	0.07 (0.04)	0.05 (0.04)
		400	0.02 (0.01)	0.05 (0.02)	0.04 (0.01)	0.05 (0.04)	0.03 (0.03)	0.03 (0.03)
2	0.23	40	0.10 (0.07)	0.09 (0.09)	0.10 (0.05)	0.19 (0.18)	0.30 (0.13)	0.34 (0.16)
		100	0.08 (0.06)	0.10 (0.05)	0.12 (0.04)	0.08 (0.07)	0.13 (0.08)	0.16 (0.10)
		200	0.05 (0.03)	0.11 (0.03)	0.12 (0.03)	0.08 (0.04)	0.09 (0.06)	0.10 (0.07)
		400	0.03 (0.02)	0.04 (0.04)	0.13 (0.01)	0.07 (0.04)	0.05 (0.04)	0.06 (0.05)
3	0.25	40	0.16 (0.11)	0.10 (0.08)	0.11 (0.11)	0.16 (0.13)	0.19 (0.12)	0.24 (0.14)
		100	0.06 (0.05)	0.08 (0.06)	0.08 (0.06)	0.11 (0.11)	0.09 (0.07)	0.11 (0.09)
		200	0.04 (0.03)	0.06 (0.03)	0.05 (0.05)	0.05 (0.04)	0.08 (0.05)	0.08 (0.06)
		400	0.04 (0.03)	0.04 (0.03)	0.04 (0.04)	0.05 (0.04)	0.05 (0.03)	0.06 (0.04)
4	0.03	40	0.04 (0.05)	0.05 (0.05)	0.10 (0.09)	0.14 (0.12)	0.12 (0.10)	0.11 (0.10)
		100	0.02 (0.04)	0.10 (0.28)	0.04 (0.04)	0.08 (0.07)	0.07 (0.07)	0.06 (0.05)
		200	0.02 (0.02)	0.06 (0.08)	0.04 (0.04)	0.07 (0.05)	0.05 (0.03)	0.05 (0.03)
		400	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.05 (0.05)	0.05 (0.04)	0.04 (0.03)
5	0.25	40	0.16 (0.10)	0.36 (0.60)	0.13 (0.16)	0.21 (0.16)	0.16 (0.11)	0.20 (0.12)
		100	0.10 (0.06)	0.15 (0.05)	0.05 (0.04)	0.11 (0.08)	0.05 (0.05)	0.06 (0.05)
		200	0.07 (0.04)	0.10 (0.06)	0.05 (0.04)	0.07 (0.07)	0.07 (0.06)	0.06 (0.05)
		400	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)	0.05 (0.04)	0.04 (0.03)	0.03 (0.03)
6	0.90	40	0.21 (0.31)	0.05 (0.05)	0.16 (0.11)	0.07 (0.04)	0.11 (0.10)	0.19 (0.20)
		100	0.06 (0.04)	0.09 (0.02)	0.10 (0.06)	0.04 (0.03)	0.04 (0.02)	0.04 (0.02)
		200	0.04 (0.03)	0.05 (0.02)	0.04 (0.03)	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)
		400	0.03 (0.02)	0.06 (0.04)	0.05 (0.04)	0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
7	0.07	40	0.05 (0.04)	0.09 (0.04)	0.09 (0.09)	0.14 (0.10)	0.11 (0.10)	0.11 (0.10)
		100	0.04 (0.02)	0.18 (0.35)	0.05 (0.04)	0.09 (0.07)	0.06 (0.05)	0.06 (0.05)
		200	0.03 (0.01)	0.08 (0.10)	0.04 (0.03)	0.05 (0.04)	0.05 (0.04)	0.05 (0.04)
		400	0.03 (0.01)	0.05 (0.03)	0.03 (0.03)	0.05 (0.04)	0.04 (0.03)	0.04 (0.02)
8	0.67	40	0.17 (0.17)	0.08 (0.07)	0.15 (0.16)	0.13 (0.09)	0.11 (0.10)	0.21 (0.14)
		100	0.08 (0.07)	0.08 (0.06)	0.12 (0.08)	0.09 (0.08)	0.11 (0.07)	0.10 (0.07)
		200	0.06 (0.04)	0.03 (0.03)	0.08 (0.06)	0.08 (0.05)	0.07 (0.05)	0.07 (0.04)
		400	0.03 (0.03)	0.05 (0.02)	0.10 (0.04)	0.04 (0.03)	0.05 (0.03)	0.06 (0.03)
9	0.33	40	0.27 (0.06)	0.38 (0.37)	0.20 (0.13)	0.32 (0.20)	0.39 (0.16)	0.43 (0.13)
		100	0.15 (0.08)	0.23 (0.08)	0.20 (0.26)	0.19 (0.09)	0.24 (0.10)	0.26 (0.07)
		200	0.12 (0.04)	0.10 (0.09)	0.15 (0.11)	0.14 (0.09)	0.16 (0.07)	0.19 (0.05)
		400	0.10 (0.02)	0.25 (0.04)	0.08 (0.09)	0.11 (0.05)	0.11 (0.05)	0.12 (0.05)
10	0.24	40	0.19 (0.05)	0.31 (0.51)	0.20 (0.06)	0.32 (0.21)	0.31 (0.15)	0.38 (0.17)
		100	0.16 (0.06)	0.23 (0.04)	0.21 (0.05)	0.20 (0.08)	0.23 (0.09)	0.22 (0.07)
		200	0.09 (0.05)	0.19 (0.08)	0.19 (0.03)	0.11 (0.07)	0.14 (0.08)	0.16 (0.08)
		400	0.06 (0.03)	0.23 (0.05)	0.18 (0.01)	0.08 (0.04)	0.10 (0.05)	0.11 (0.04)

Table II. (Best) Count of datasets in which each method achieved the smallest error. ($> 50\%$) Count of datasets in which each method achieved $\frac{\hat{I}-I_{real}}{I_{real}} > 50\%$.

	n	VMI	MLMI	EDGEW	3-NN	5-NN	7-NN
Best	40	3	5	3	0	0	0
	100	8	1	1	2	2	1
	200	7	2	2	1	1	1
	400	8	2	5	1	1	1
$> 50\%$	40	6	5	5	6	6	6
	100	3	6	5	4	4	4
	200	1	4	4	2	4	4
	400	1	5	3	2	2	2

Now we proceed to the second batch of experiments, this time conducting feature selection in classification tasks. We considered 10 binary classification datasets[4], and a 10-fold cross-validation procedure to generate 10 pairs of training and test sets. Each feature in the training set was normalized to have zero mean and unit variance. The scale factors obtained for the training set were also applied for normalizing the test set. Normalized values greater than 3 or less than -3 were set to 3 and -3 (resp.) in both training and test sets.

[4]See footnote 3.

Table III. Mean and standard deviation of the balanced error of an SVM classifier after the selection of the best 5 or 10 features according to several methods. Also shown are the size of the dataset (n), the number of features (d), the proportion of minority-class examples (% Min.), and the SVM balanced error using all features (All Feat.).

Dataset	n	% Min.	d	All Feat.	k	VMI	MLMI	EDGEW	k-NN
Arcene	200	44	10000	0.14 (0.04)	5	0.25 (0.08)	0.25 (0.09)	0.41 (0.06)	0.26 (0.09)
					10	0.19 (0.09)	0.23 (0.09)	0.40 (0.05)	0.23 (0.09)
Lung-Uterus	250	50	10936	0.07 (0.05)	5	0.09 (0.03)	0.11 (0.05)	0.13 (0.06)	0.07 (0.05)
					10	0.07 (0.03)	0.07 (0.03)	0.10 (0.03)	0.06 (0.05)
Ovary-Kidney	458	43	10936	0.03 (0.03)	5	0.04 (0.02)	0.06 (0.04)	0.18 (0.11)	0.03 (0.03)
					10	0.03 (0.02)	0.05 (0.03)	0.08 (0.03)	0.03 (0.03)
Biodeg	1055	34	41	0.13 (0.03)	5	0.19 (0.03)	0.32 (0.03)	0.42 (0.03)	0.19 (0.03)
					10	0.18 (0.03)	0.27 (0.04)	0.28 (0.04)	0.15 (0.03)
Climate	540	9	18	0.11 (0.07)	5	0.10 (0.08)	0.09 (0.06)	0.11 (0.08)	0.10 (0.07)
					10	0.12 (0.09)	0.09 (0.06)	0.16 (0.12)	0.09 (0.08)
Ionosphere	351	36	34	0.06 (0.03)	5	0.08 (0.05)	0.16 (0.07)	0.10 (0.05)	0.14 (0.05)
					10	0.07 (0.04)	0.08 (0.04)	0.08 (0.05)	0.11 (0.03)
Parkinson's	195	25	22	0.07 (0.07)	5	0.13 (0.10)	0.15 (0.10)	0.13 (0.08)	0.10 (0.06)
					10	0.11 (0.08)	0.08 (0.08)	0.09 (0.09)	0.13 (0.09)
WPBC	194	24	33	0.34 (0.09)	5	0.41 (0.08)	0.37 (0.09)	0.33 (0.12)	0.37 (0.11)
					10	0.34 (0.12)	0.29 (0.12)	0.33 (0.09)	0.32 (0.11)
WDBC	569	37	30	0.02 (0.02)	5	0.06 (0.03)	0.05 (0.02)	0.10 (0.05)	0.05 (0.02)
					10	0.04 (0.02)	0.05 (0.02)	0.05 (0.03)	0.03 (0.02)
Sonar	208	47	60	0.14 (0.07)	5	0.27 (0.09)	0.22 (0.08)	0.27 (0.08)	0.22 (0.09)
					10	0.16 (0.08)	0.18 (0.09)	0.22 (0.11)	0.18 (0.09)

The mutual information between each pair of features and the target class was estimated using a fixed mutual information estimation method and the normalized training examples for the dataset. Then, 5 (or 10) features were selected by the JMI procedure described in Section 4, and an SVM classifier[5] was obtained from the normalized training set restricted to those selected features. The balanced error rate[6] of the resulting classifier was verified on the normalized test set restricted to the same 5 (or 10) features. To add perspective, we also report the balanced error rate for the classifier obtained on all features.

Here, the application of the k-NN estimator is *distinct* from that of the previous set of experiments. We can treat k as just another parameter to be selected in the parameter selection procedure for SVM. This way, we optimized the parameter k not for mutual information estimation per se, but directly for feature selection. We could have applied the same procedure to estimate the parameters of MLMI and VMI, yet we chose not to do so in order to evaluate whether better mutual information estimation leads to better feature selection.

The full set of results is presented in Table III. By ordering the mutual information estimators from the lowest (best) balanced error to the highest (worst) one in each row of the table, we arrive at the following average ranking for the methods: 2.0 for k-NN, 2.2 for VMI, 2.4 for MLMI, and 3.4 for EDGEW. This ranking puts the feature selection scheme using the k-NN estimator as the best one. This result alone should come with no surprise, since the parameter k is being optimized directly for feature selection. Also, notice that the best performing method for a fixed dataset often varies.

Now we focus on just those cases that rely only on mutual information estimation, that is, VMI, MLMI, and EDGEW. It is clear that the Edgeworth estimator leads to the worst classification results. According to the average ranking, VMI has a slight advantage over MLMI. However, the average ranking discards the magnitude of the differences in classification error. This way, let us compare the results of VMI and MLMI where the largest discrepancies occurred: in favor of VMI, Ionosphere and Biodeg; and in favor of MLMI, Sonar and WPBC. Notice that the discrepancy in performance when VMI loses is not as large as the opposite. Thus, VMI led to safer feature selection than MLMI.

[5]For SVM, we use an RBF kernel. The parameters C and σ were selected using grid-search [Braga et al. 2013].
[6]The balanced error gives the same weight to the errors within each class.

6. CONCLUSION

In this work we investigated VMI, a new mutual information estimation method that was observed experimentally to be more accurate than previous estimators like MLMI, Edgeworth, and k-NN. Along with this observation, experiments that employed MI estimation for feature selection evidenced that better feature selection can be achieved by using VMI instead of MLMI or Edgeworth.

These results allows us to conclude that better mutual information estimation is a way of achieving better feature selection. However, as the experiments with the k-NN estimator indicate, it is not the only way: by optimizing the parameter k directly for feature selection, the k-NN estimator obtained the best results among the four investigated methods. As the k-NN estimator was not the best mutual information estimator in our experiments, these results point to a class of feature selection methods that rely on the *identification* of feature sets that have a large value of mutual information towards the target class. This identification task comprises another research line on feature selection, as it is different from (and may be simpler than) the estimation of the *value* of MI.

For now, we are left with the question: by considering better mutual information estimation, can we hope to construct state-of-the-art methods of feature selection? By using a more principled density-ratio estimation method, it is possible to achieve better mutual information estimation using the approach taken in this work. Taking into account that the particular estimator used in our experiments is the simplest instance of the constructive density-ratio estimation method, exploring more sophisticated instances is a good opportunity to check this question.

REFERENCES

Braga, I., do Carmo, L. P., Benatti, C. C., and Monard, M. C. A note on parameter selection for support vector machines. In *MICAI '13: Proceedings of the 2013 Mexican International Conference on Artificial Intelligence*. Mexico City, Mexico, pp. 233–244, 2013.

Brown, G., Pocock, A., Zhao, M.-J., and Luján, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research* 13 (1): 27–66, 2012.

Cover, T. and Thomas, J. *Elements of Information Theory*. Wiley-Interscience, 2006.

Guyon, I. and Elisseeff, A. An introduction to variable feature selection. *Journal of Machine Learning Research* vol. 3, pp. 1157–1182, 2003.

Hyvärinen, A. and Oja, E. Independent Component Analysis: algorithms and applications. *Neural Networks* 13 (4-5): 411–430, 2000.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical Review E* 69 (6): 1–16, 2004.

Rifkin, R. M. *Everything Old is New Again: a Fresh Look at Historical Approaches in Machine Learning*. Ph.D. thesis, MIT-Sloan School of Management, USA, 2006.

Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS '07: Advances in Neural Information Processing Systems 20*. Vancouver, Canada, pp. 1–8, 2008.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2011.

Suzuki, T., Sugiyama, M., and Tanaka, T. Mutual information approximation via maximum likelihood estimation of density ratio. In *ISIT '09: Proceedings of the 2009 IEEE International Symposium on Information Theory*. Seoul, South Korea, pp. 463–467, 2009.

van Hulle, M. M. Edgeworth approximation of multivariate differential entropy. *Neural Computation* 17 (9): 1903–1910, 2005.

Vapnik, V., Braga, I., and Izmailov, R. A constructive setting for the problem of density ratio estimation. In *SDM '14: Proceedings of the 2014 SIAM International Conference on Data Mining*. Philadelphia, USA, pp. 434–442, 2014.

Vapnik, V. N. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Yang, H. H. and Moody, J. Data visualization and feature selection: new algorithms for nongaussian data. In *NIPS '99: Advances in Neural Information Processing Systems*. Denver, USA, pp. 687–693, 1999.