From: Igor Braga
To: Editor(s) of the JIDM-KDMile special issue
Subject: JIDM-KDMile paper submission

ABOUT THE EXTENSION

First I would like to thank the KDMile and JIDM organizers for having arranged this great opportunity to communicate our research results. Also, I would like to thank the anonymous reviewers of KDMile for their helpful comments and suggestions.

This submission is an extension to the previous short paper (4 pages) submitted to KDMile. The improvements to the previous version follow two lines.

1) *Addition of justification for the proposed techniques.* As the previous version is a short paper, a lot of explanation had to be left out. It is my hope that the paper has now enough motivation for the appreciation of reviewers. The main modifications here were the addition of an introductory and concluding section, more explanation about mutual information estimation, and several illustrative pictures.
2) *New batch of experiments on feature selection using real data.* There is an entire new section on feature selection and how the proposed techniques can be leveraged to perform this task. The experimental section is augmented with an evaluation of mutual information estimation as part of a feature selection scheme in classification. The results of the two batches of experiments allow us to draw interesting conclusions about mutual information estimation and feature selection.

RESPONSE TO THE KDMILE REVIEWS

REVIEW 1

The main problem is the lack of justification for the techniques that are empirically analyzed.

We hope that this expanded version has helped mitigate this problem.

Consider Expression (3)... why is it written as r(X|Y)? Is it supposed to be a conditional density?

No. We fixed the notation in the final version of the KDMile paper to avoid confusion.

Also, it is not clear to me that Expression (3), as an interpretation of I(X,Y), is of any use to the paper. How is it used?

Actually, former Expression (3), now Expression (5), is essential to this work. As it is now explained in Section 2, it is the basis for an estimator that is more robust to errors in the estimation of the density-ratio.

Another point is that the expression at the top of Page 3 is very mysterious to me. Is it just a simple statement of facts that are known to any statistical book (as it seems to me), or something new (where?).

That is right. It is simply the mathematical expression for the multidimensional empirical cumulative distribution function. The difference is that textbooks generally keep themselves to the one-dimensional expression. Pictures were added to illustrate the empirical cumulative distribution function.

Finally, I did not quite understand the meaning of "> 50%" in Table I: relative average error, but relative to what?

Errors in mutual information estimation are reported relative to the real mutual information. That is, *err = (Iest – Ireal)/Ireal*. When this error was greater than 50%, it was counted in the bottom rows of former Table I, now Table II.

REVIEW 2

The performance must be evaluated with large real datasets.

We hope that the experiments on feature selection have helped mitigate this problem.

REVIEW 3

The paper lacks of motivation and needs extra explanations.

We hope that this expanded version has helped mitigate this problem.

They present the number of the datasets that a estimator perform better than the other, but how much better?

On the occasion, the full table of results was made available online. Now it is in the paper. However, the description of the synthetic models is still online. See: https://sites.google.com/site/igorabmi/.

A conclusion section with future directions are also missing.

We hope that this expanded version has helped mitigate this problem.

REVIEW 4

In fact, the experiments were performed with k *only* in {3,5,7}. So, the use "irrespective of k" is a bit misleading.

Not so much. We followed the consistency rule of k-NN classification for determining the maximum value of k. One of the consequences of this rule is that if we use k=ln(n), where n is the sample size, then the k-NN classifier is consistent. Since our largest sample size in the first batch of experiments is 400, then k=ln(400) justifies the choice of 7 (the next odd number) as the maximum value of k. Formally, this justification does not apply because we are not playing with k-NN classification, so it is just an heuristic. Nevertheless, we removed the “irrespective of k” in the text for it is too strong.

Also, it would be interesting if the author comments on the computational complexity of the different methods analyzed.

We hope that this expanded version has helped mitigate this problem.