NEW METHODS FOR FINDING ASSOCIATIONS IN LARGE DATA SETS: GENERALIZING THE MAXIMAL INFORMATION COEFFICIENT (MIC)

Tomasz M. Ignac[1,2], *Nikita A. Sakhanenko*[2], *Alexander Skupin*[1,2] *and David J. Galas*[1,2]

[1]Luxembourg Centre for Systems Biomedicine
University of Luxembourg
7, Avenue des Hauts-Fourneaux
L-4362 Esch-sur-Alzette
[2]Institute for Systems Biology
401 Terry Avenue North, Seattle, Washington 98109, USA
{tomasz.ignac, alexander.skupin}@uni.lu
{nsakhanenko, dgalas}@systemsbiology.org

ABSTRACT

We propose here a natural, but substantive, extension of the MIC. Defined for two variables, MIC has a distinct advance for detecting potentially complex dependencies. Our extension provides a similar means for dependencies among three variables. This itself is an important step for practical applications. We show that by merging two concepts, the interaction information, which is a generalization of the mutual information to three variables, and the normalized information distance, which measures informational sharing between two variables, we can extend the fundamental idea of MIC. Our results also exhibit some attractive properties that should be useful for practical applications in data analysis. Finally, the conceptual and mathematical framework presented here can be used to generalize the idea of MIC to the multi-variable case.

1. INTRODUCTION

Data sets that represent measurements on complex systems often embody functional relationships between variables that are difficult to discover. In a complex system the dependency between measured variables can have a functional form that is itself complex, and can therefore be difficult to detect by standard methods. The maximal information coefficient (MIC) represents an interesting new approach [1] to measuring dependency between two random variables. It is able to capture a wide range of functional associations, which makes it a particularly useful tool for exploring large, complex data sets, and thus it is especially appropriate for investigating large biological data sets with many variables. For example, the yeast based data set discussed in [2] contains 225 variables representing genetic markers and 374 yeast strains leading obviously to a potentially huge number of genetic interaction.

In Speed's commentary on reference [1] he pointed out that an interesting challenge is presented by this new approach. The challenge is the generalization of MIC, which he called a ”correlation for the 21st century” [3], to more than two variables. In this paper, we take up this challenge by making use of our early work [2] where we used the concept of interaction information [4, 5], a multivariate generalization of mutual information, to find a suitable binning of a continuous variable while preserving the relationships between the other variables. We use the interaction information idea as a framework for extending MIC from two to three variables. We propose substituting a normalized information distance for mutual information, which is used in MIC, as the key measure of dependence. The approach we propose also offers a clear conceptual starting point for extending the theory of MIC much farther.

The central aim of this paper is to provide a theoretical framework for future work on constructing measures of associations between three variables. There are two general approaches to this problem: the first is based on information theory methods like MIC and interaction information. The second is the more traditional approach similar to the partial correlation coefficient which is an extension of the linear correlation between variables X_1 and X_2 while a third variable Y is fixed at some value [3]. In the current paper we focus our attention on the first approach; however, we briefly discuss possibilities of alternative solutions of the problem.

2. THE MAXIMAL INFORMATION COEFFICIENT

Here we present a brief description of the process of calculating the MIC. A more detailed discussion can be found in [1]. Let X_1 and X_2 be two continuous random variables, and let $\mathcal{D}$ be a set of pairs drawn from the joint probability distribution $P(x_1, x_2)$. Then, the value $MIC(\mathcal{D})$, which stands for the MIC computed for the sample set $\mathcal{D}$, is the maximal possible mutual information [6] between all possible binnings of these variables. More precisely, let X_1' and X_2' be binned (discretized) versions of X_1 and X_2. Based on the data set $\mathcal{D}$, we can approximate the mutual information, $I(X_1'; X_2')$, between X_1' and X_2'. Subsequently, $I(X_1'; X_2')$ is normalized by $\log(min(|X_1'|, |X_2'|))$, where $|X_i'|$, $i = 1, 2$, is the number of states (bins) of X_i'. Finally, the $MIC(\mathcal{D})$ is the maximal normalized mutual information estimated from all possible binnings X_1 and X_2. An algorithm for approximating this value was presented in the supplementary materials of [1]. An improvement of this algorithm will be one of the most important topics of our future work.

The basic concept behind the idea of MIC is simply that the mutual information is a good measure of association between two variables, and maximizing this discretized measure by choice of binnings produces MIC. It simply finds the most informative binning of the two variables of interest. Unfortunately, however, the use of mutual information in practical applications can be difficult. There are two sources of these difficulties: 1) estimation of mutual information from data is difficult, especially for continuous random variables, and 2) mutual information itself is not a normalized measure; thus, the interpretation of the results may be sometimes problematic [3, 6].

MIC cannot be treated as an approximation of the mutual information between X and Y, however. On the other hand, there are two theorems [1] showing that if X and Y are independent or if $X = f(Y)$, then the $MIC(\mathcal{D})$ converges to 0 or 1 respectively with the size of $\mathcal{D}$ going to infinity. What is more, MIC offers a natural normalization of obtained results.

3. INTERACTION DISTANCE

To extend MIC we have to have a multivariate generalization of the mutual information. The natural choice is the conditional mutual information $I(X_1; X_2|Y)$. However, it can be easily shown that this is not the best solution. For example, $I(X_1; X_2|Y)$ equals zero when either all the three variables are mutually independent or when X_1 and X_2 are independent given Y. These are two completely different situations, which should be differentiated by a good measure, but are not distinguished by conditional mutual information.

In [2] we used the concept of interaction information, $I(X_1; X_2; Y)$ [4, 5, 7], which is defined as a difference between $I(X_1; X_2|Y)$, the mutual information between X_1 and X_2 given Y, and $I(X_1; X_2)$. The value of $I(X_1; X_2; Y)$ can be either positive or negative, as can be seen from the range of values of these two terms. A positive value suggests a synergy between X_1 and X_2; i.e., both variables together contain more information about Y than separately. A negative value suggests redundancy between X_1 and X_2.

While the interaction information appears to provide the path to a natural extension of MIC, we note that $I(X_1; X_2; Y)$ is symmetric: for example, $I(X_1; X_2; Y) = I(X_1; Y; X_2)$. Consequently, $I(X_1; X_2; Y)$, a single number describing a relationship between three variables, still does not capture all possible associations between the variables. This is illustrated by the Example 3.1.

A better generalization of MIC can be achieved by replacing the mutual information measure, $I(X_1; X_2)$, with the normalized information distance, $d(X_1; X_2)$, and then extending $d(X_1; X_2)$ to three variables using the concept of interaction information. The normalized information distance [8, 9] is a metric defined as

$$d(X_1; X_2) = \frac{\max[H(X_1|X_2), H(X_2|X_1)]}{\max[H(X_1), H(X_2)]},$$

which can be rewritten to

$$d(X_1; X_2) = \frac{\max[H(X_1), H(X_2)] - I(X_1; X_2)}{\max[H(X_1), H(X_2)]}.$$

Here, $H(\cdot)$ stands for the entropy.

The normalized information distance was defined in [8] in terms of Kolmogorov complexity [10]. However, it can be easily adapted to the Shannon's formalism [6]. We want to point out that this distance offers an alternative approach for the normalization of the mutual information between X_1 and X_2.

The extension to three variables $d(X_1; X_2.Y)$ can then be made in the same way as in the case of the interaction information:

$$d(X_1; X_2.Y) = d(X_1; X_2|Y) - d(X_1; X_2).$$

We call this quantity the interaction distance by analogy, even though it is not a metric, as opposed to the two-variable form. Here $d(X_1; X_2|Y)$ is a conditional version of the normalized distance, i.e., $d(X_1; X_2|Y) =$

$$\frac{\max[H(X_1|Y), H(X_2|Y)] - I(X_1; X_2|Y)}{\max[H(X_1|Y), H(X_2|Y)]}.$$

It can be shown that $d(X_1; X_2|Y)$ is a metric. The proof of that property will be presented in the extended version of this paper. Here, we want to note that the proof for the Shannon's form of the distance is relatively simple; i.e., it is an extension of the original proof [8], that $d(X_1; X_2)$ is a metric. This extension is based on the fact that:

$$H(X, Y|Z) = H(X|Z) - H(Y|X, Z).$$

Unfortunately, the Kolmogorov counterpart of this property does not hold exactly [11]. Therefore, it is unclear if the Kolmogorov's version of $d(X_1; X_2|Y)$ is a metric.

To prove usefulness of the interaction distance, we need to describe basic properties of $d(X_1; X_2.Y)$. In order to do that, we need to go back to a theorem which is a key result of [7]. This theorem describes behavior of the interaction information in the context of various relationships between the three variables. Here, we present a lemma that is a counterpart of this theorem. The detailed proof is omitted here, due to space limitations. The lemma can be treated as a corollary of the theorem.

Lemma 3.1 *The following three properties hold for arbitrary functions f and g:*

1. $-1 \leq d(X_1; X_2.Y) \leq 1$.

2. $d(X_1; X_2.Y) = -1$ *if and only if* X_1, X_2 *are independent and* $X_i = f_j(X_j, Y), i, j = 1, 2$.

3. $d(X_1; X_2.Y) = 1$ *if and only if* $X_i = g_i(Y) = f_j(X_j), i, j = 1, 2$.

The first property simply shows the range of the interaction distance. The limit values are obtained if and only if certain functional associations between the three variables occur. The second property implies that Y is fully determined by two independent variables. The third statement describes situation when X_1 and X_2 are determined by Y. In more general setting, see [4, 5], when the relation between variables is not functional, the negative values of the interaction distance suggest synergy of X_1 and X_2; while the positive distance means redundancy of the information of these two variables with respect the conditioning variable Y. This includes also a case when X_1 and X_2 are independent given Y. Note that, in contrast to the interaction information, the information distance is negative in the case of a synergy between X_1, X_2, and positive if the variables are redundant.

Example 3.1 Here we present an example that demonstrates the difference between the interaction information and the interaction distance. Let X_1 and X_2 be two independent, binary, random variables such that $P(X_i = 0) = 0.5, i = 1, 2$. Let us define a third variable Y in as follows:

- $Y = 0$ for $X_1 = 0$ and $X_2 = 0$;
- $Y = 1$ for $X_1 = 1$ and $X_2 = 0$;
- $Y = 2$ for $X_1 = 0$ and $X_2 = 1$;
- $Y = 3$ for $X_1 = 1$ and $X_2 = 1$.

Note that the triple X_1, X_2, Y fulfills requirements of the Lemma 3.1, point 2. This is an obvious case of synergy between X_1 and X_2; i.e., on one hand, knowledge about the state of only one of these two variables leaves the state of Y uncertain. On the other hand, when states of both X_1 and X_2 are known, the state of Y becomes certain.

Let us now analyze the behavior of the interaction information and interaction distance in this case. Clearly, $I(X_1; X_2) = 0$; subsequently, elementary calculations reveal that $I(X_1; X_2|Y) = H(X_i|Y) = 0$ and thus
$I(X_1; X_2; Y) = 0$. Since X_1 and X_2 are independent $d(X_1; X_2) = 1$; then, we can show that $d(X_1; X_2|Y) = 0$. To this end, we note that

$$I(X_1; X_2|Y) = H(X_i|Y) \text{ for } i = 1, 2,$$

Thus, it follows, from the definition of the conditional distance, that $d(X_1; X_2|Y) = 0$. Hence, we have

$$d(X_1; X_2.Y) = d(X_1; X_2|Y) - d(X_1; X_2) = 0 - 1 = -1.$$

Note that in the example $d(X_1; X_2|Y)$ looks like zero over zero. We treat it as zero since this is the special case when the conditional mutual information is equal to the conditional entropy. Thus, $d(X_1; X_2|Y) = 0$ for arbitrarily small value of these quantities. Hence, in the limit, we obtain the zero interaction distance.

The immediate advantage of using $d(X_1; X_2.Y)$ is its ability to capture a broader spectrum of relations than the interaction information itself. A positive value of the interaction information indicates synergy between variables [4, 5]. However, the above example shows that the reverse is not always true. To understand the difference between the interaction distance and the interaction information we need to go back to the key theorem of [7]. From this theorem it follows that in situations similar to that in our example the interaction information equals $H(X_i|Y)$. Hence, when the conditional entropy tends to zero the interaction information follows. We can see that the distance is independent from the values of the entropies and conditional entropies of X_1 and X_2. This allows us to detect associations that cannot be captured by the interaction information itself. The price for this capacity is that the distance is not symmetric, and not a metric. Thus, sometimes, one may need to consider three cases of conditioning by all variables.

4. CONCLUSIONS AND DISCUSSION

One of the directions for future work is exploration of the question of statistical power of MIC. However, this is not specific to the three (or more) variable case. There are two main directions on future problems. The first is the practical implementation of the interaction distance. The second one is to find possible alternatives for the interaction distance. Further work will involve using this present framework, used to generalize to three variables, to extend MIC to multi-variable cases.

4.1. Implementation

For a given data set the information distance, *d*, is approximated from the data in a manner very similar to the MIC algorithm [1]. The details of these operations are beyond the scope of this short paper and will be presented in its extended version. We have simply adapted the algorithm presented in [1]. In short, we estimate $d(X_1; X_2.Y)$ in two steps. In the first step, we maximize $d(X_1; X_2)$ and in the second step we maximize $d(X_1; X_2.Y)$. Note that the first step is a simple adjustment of the existing MIC algorithm. On the other hand, the second step is more complex, as it requires taking into account the conditional variable Y. The simplest solution here is to impose an equi-partition on the values of Y; c.f., supplementary materials of [1]. It is unclear if this solution is generally optimal in practical applications. In [1] a similar approach is used where one of the variables is equi-partitioned.

Note that by maximizing $d(X_1; X_2|Y)$ and $d(X_1; X_2)$ separately different discretizations of X_1 and X_2 can be obtained. Consequently, an important direction for future research is to define context-dependent discretization of a random variable: such discretization will be different when we change contextual variables of the variable of interest.

The complexity and time for the computations could be a key issue for future applications. We have performed some preliminary tests on the yeast data set, mentioned above, with 225 binary variables and 375 samples. We calculated the interaction distances $d(X_i; X_j.Y)$, where i, j went across all the pairs of the 225 variables and Y was an additional variable that represents phenotype. The phenotype variable was binned into four states. The running time on a laptop was about ten seconds. We artificially generated similar set with 1000 samples and 1200 variables: the running time in this case was about nine minutes. Even if we take into account that we may need to test various binnings, the running time should be acceptable. The detailed discussion of this issue will be considered in future papers.

4.2. Alternatives

The problem with the interaction distance is the large number of samples required to obtain a sound estimation of $d(X_1; X_2.Y)$. Even for the MIC between two variables we need relatively large number of samples: the minimal practical size of $\mathcal{D}$ is about 100. Since the interaction distance involves three variables, an order of magnitude more samples will be required here. In many practical applications we may not be able to collect a sufficient amount of observations. Thus, we need to find an alternative. This could be based on the idea that underlies the partial correlation between X_1 and X_2 given Y, denoted by $\rho_{X_1 X_2 \cdot Y}$. This is the correlation of residuals of X_1 and X_2 calculated from the linear regression of X_1 given Y and X_2 given Y.

In our future research we want to develop a similar approach but replace the correlation of the residuals by a measure or a statistical test that can capture more than only linear relationships. A good candidate for such a measure seems to be the distance correlation introduced by Szekely [12].

In the Science perspective commenting on [1] a challenge was presented [3]: "MIC is a great step forward, but there are many more steps to take." To take the first step we have proposed here a relatively natural, but substantive, extension of the MIC for detecting potentially complex associations among three random variables. This itself is an important step for practical applications. We have shown that by merging two concepts, the interaction information, which is a generalization of the mutual information to three variables, and the normalized information distance, which measures informational sharing between two variables, we are able to extend the fundamental idea of MIC. The interaction distance we propose exhibits some attractive properties that should also be useful for practical applications in many aspects of data analysis and the framework presented here can be used to generalize to the multi-variable case. The technical details of our method will be a topic of a future publication.

5. ACKNOWLEDGEMENTS

This work was supported by the ISB-Luxembourg Program, and by the FIBR program of NSF (0527023).

Nikos Vlassis from the Luxembourg Centre for Systems Biomedicine for discussions.

6. REFERENCES

[1] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 64, pp. 1518–1524, 2011.

[2] N. A. Sakhanenko and D. J. Galas, "Interaction information in the discretization of quantitive phenotype data," in *Proceedings of the 8th International Workshop on Computational Systems Biology*, Zurich, Switzerland, June 2011.

[3] T. Speed, "A correlation for the 21st century," *Science*, vol. 334, no. 64, pp. 1502, 2011.

[4] A. Jakulin and I. Bratko, "Testing the significance of attribute interactions," in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, June 2004.

[5] A. Jakulin and I. Bratko, "Quantifying and visualizing attribute interactions: An approach based on entropy," *http://arxiv.org/abs/cs.AI/0308002 v3*, 2004.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, NY, USA, 1991.

[7] T. Tsujishita, "On triple mutual information," *Advances in Applied Mathematics*, vol. 16, no. 3, pp. 269–274, 1995.

[8] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The similarity metric," *IEEE Transactions on Information Theory*, pp. 863 – 872, Sep. 2003.

[9] D. J. Galas, M. Nykter, G. W. Carter, N. D. Price, and I. Shmulevich, "Biological information as set-based complexity," *IEEE Transactions on Information Theory*, vol. 56, pp. 667 – 677, Feb. 2010.

[10] A. N. Kolmogorov, "Three approaches to the definition of the concept quantity of information (russian)," *Problemy Peredachi Informacii*, vol. 1, pp. 3 – 11, 1965.

[11] P. Gacs, J. T. Tromp, and P. M. B. Vitanyi, "Algorithmic statistics," *IEEE Transactions on Information Theory*, vol. 47, pp. 2443 – 2463, Sep. 2001.

[12] G. J. Szekely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 27692794, 2007.