The Effectiveness of Maximal Information Coefficients in Real-World Classification Tasks

Yanru CHEN*, Wanwan ZHENG**, Mingzhe JIN***

(Received July 7, 2021)

The maximal information coefficient is a measure that was proposed in 2011 and can detect non-linear relationships in experiments using artificial data. However, its effectiveness on real-world data has not been sufficiently demonstrated. In this study, various benchmark data sets from different fields were gathered to evaluate the effectiveness of the maximal information coefficient in real-world classification tasks. Distance-based discriminant analysis and support vector machine were adopted as classifiers. Accuracies and computational costs were employed to evaluate the results. Compared to the baselines including Euclidean distance, the Pearson correlation coefficient, cosine similarity and Spearman's rank correlation coefficient, the classification accuracy of the maximal information coefficient failed to show superiority and its computational costs were significantly higher than the other measures.

Key words : maximal information coefficient, classification, real-world data

1. Introduction

Modern analytics are becoming more indispensable as business becomes increasingly complex, in addition, decision-makers are required to act more rapidly and accurately based on available evidence. To detect the inner relationships within massive data, pattern recognition has become a hotspot in information science. In this field, most of the methods make use of statistical machine learning techniques which enable machines to process data, learn from data, and make decisions based on patterns in data. Due to the advancements of technology, dealing with large-scale data has become much easier. A significant application of multivariate data analysis, especially real-world data tasks, is to classify data for specific issues to achieve quantitative targets corresponding to each subject. Adoption of proper statistical measures contribute to increasing the classification accuracy.

On the other hand, the mutual relationship between different items builds a foundation that allows there to be a unity. For instance, the discovery of the interrelationships between genes, from massive genomic data, is an important research topic in modern biology. Generally, this kind of mutual relationship could be correlation, distance or similarity. Especially, in the era of information science, the large amount of data has become the most prominent characteristic of data.

Correlation indicates the statistical relationship between two variables (in bivariate data). The concept of correlation was first raised in the 19th century. It was suggested that all parts of an organization are connected or correlated to a certain extent. To measure the strength of correlation, the Pearson correlation coefficient was proposed, which has become one of the most famous dependence measures. In the following years, study of various dependence measures is one of the most basic aspects of science. A common goal of scientific research

* ann93101@outlook.com
** zhengwanw@hotmail.com
*** Culture and Information Science, Doshisha University, Kyoto
mjin@mail.doshisha.ac.jp

is to measure the correlations in data, and apply the corresponding correlation in other tasks, for example, outlier detection, feature selection and authorship attribution. However, the Pearson correlation coefficient has a weakness in that it can only detect linear relationships. Consequently, an applicable dependence measure for diverse situations is desired.

The maximal information coefficient (the MIC) is a new correlation measure proposed by Reshef *et al.* (2011), which is able to detect non-linear relationships and is considered to be a very useful complement to standard and rank correlation measures[1]. Due to its capability of capturing non-linear functional relationships, Speed (2011) regards the MIC as a promising correlation statistic for the 21st century[2].

The MIC has already been used in some fields. For example, Pang *et al.* (2012) adopted the MIC to measure dependence relationships between every pair of parameters from transcriptomics, proteomics, interactomics, as well as phenotypic and sequence-based data sets, revealing significant signaling between densely connected kinase clusters[3]. In astrostatistics and cosmology science, De Souza *et al.* (2014) made a comparison between the MIC and Spearman's rank correlation coefficient, and showed that the MIC is reliable when halo properties are weakly correlated[4]. Concentrating on the application of the MIC to assist a specific research field, Posnett *et al.* (2012) investigated the applicability of the MIC to software engineering data. This study was interested in the distributional properties of the MIC and its relative measures, the interpretation of the values, the stability under limited sample size, and the capability of identifying new relationships. The MIC was found to be potentially very useful for exploring new data sets and new hypothesis generation. On the other hand, according to this study, the MIC is sensitive to sample size and noise, and in some cases the MIC and its relative measures provide much less power than expected, with lower statistical power than other dependence measures. Fortunately, for larger sample sizes this flaw is less of a problem. However, the authors pointed out that the MIC is expensive to compute[5].

Although the MIC has been reported to perform well in most current functional simulations and has been applied in a few fields for experiments, the appropriateness of using the MIC has never been probed, and the performance of this measure in real-world tasks has not been well studied. For a novel measure, people often become curious about two aspects: which kind of classification method it suits, and which characteristics of data it is compatible with. In this study a wide variety of real-world benchmark data sets from different fields were used to evaluate the classification effectiveness of the MIC. Additionally, the processing time was examined to assess the MIC's computational cost.

2. The MIC and the Baselines

The main idea of computing a MIC is based on the recognition that if there is some correlation between two variables, then through the grid search technique on the scatter plot of these two variables, the distribution of data in the grid could indicate a correlation.

Let D be a set of ordered pairs. For the grid G, let $D|_G$ denote the probability distribution induced by data D on the cells of G, and let $I(-)$ denote mutual information. Let $I^*(D,x,y) = \max_G I(D|_G)$, where the maximum is taken over all x-by-y grids G which may have empty rows or columns. The MIC is defined as follows:

$$\mathrm{MIC}(D) = \max_{xy<B(|D|)} \frac{I^*(D,x,y)}{\log_2 \min\{x,y\}} \qquad (1)$$

where B is a growing function satisfying $B(n) = O(n)$. Generally, $B(n) = n^{0.6}$ is suggested. The value of the MICs ranges from 0 to 1.

Previous research has verified that the MIC is capable of detecting a wide variety of associations whether they are functional or not. According to Reshef *et al.* (2013), the MIC has two main properties: generality and equitability[6]. Generality means that with enough samples, the MIC can capture a wide range of associations not limited to specific function types. Equitability means that the MIC can give similar scores

to equally noisy relationships of different data types.

Although the MIC is capable of detecting non-linear relationships with generality and equitability, the already examined superiorities of the MIC over the existing measures are based on artificial simulations, and few studies focused on the validity of using the MIC in real-world tasks. In this empirical study, widely used measures including Euclidean distance, the Pearson correlation, cosine similarity and Spearman's rank correlation were chosen as the baselines for comparison.

3. The Current Study

In this section, a total of 30 data sets from various fields were employed to compare the performance of the MIC with other frequently used measures in multivariate data analysis.

3.1 *Data*

Most of the data sets were downloaded from the UCI Machine Learning Repository[7], KEEL data set repository[8] and OpenML Machine Learning Repository[9]. All the data sets are numerical and were collected from various fields including medicine science, biology, commerce, gene expression, robotics and chemistry.

Most of these data sets are unbalanced, the number of instances ranges from 47 to 1424, and the number of attributes ranges from 3 to 10935. According to the conclusion in the original paper[1], it is assumed that the MIC would perform better in classification tasks for data with more attributes.

Therefore, these experiment data sets were divided into two groups: those with less than 100 attributes (18 cases) and those with over 100 attributes (12 cases).

3.2 *Data preprocessing*

Some data sets with missing values or sparse data were preprocessed by deleting the instances including missing values; for each cleaned data set, less than 5% of cases were deleted.

In multivariate analysis, data sets usually have different dimensions and orders of magnitude. If the raw data sets are directly used for analysis, the attributes with higher values in the comprehensive analysis would be amplified, and oppositely the attributes with lower values would be weakened. Therefore, in order to ensure the reliability of experimental results, it is necessary to standardize the original data. The z-score standardization was conducted. The equation is:

$$z = \frac{x - \bar{x}}{s} \tag{2}$$

where $\bar{x}$ is the arithmetic mean, and S is the standard deviation.

Generally, for a distance measure, the nearer the items are, the more similar attributes they may possess; but for a correlation measure, the higher the value is, the closer the relationship will be. Consequently, to validly compare the measures to each other, all the dependence measures were transformed by subtracting them from one (in this way, zero indicates the highest possible correlation, just as zero distance represents existing at the same place). Thus after this transformation, it is valid to evaluate these measures against each other. In the following experiments, the distances and various correlations between items were computed for all data sets. Due to the z-score standardization employed in the experiments, cosine similarity and Pearson correlation result in the same value, hence the results of cosine similarity are omitted.

3.3 *Experimental study*

To deal with classification tasks, discriminant analysis and support vector machine (SVM) have been widely used in natural language processing, image recognition, data mining and machine learning. Accuracy is commonly employed to evaluate the effectiveness of each measure. To compute accuracy, a confusion matrix that contains all the possible classes is used.

A confusion matrix is an $N \times N$ matrix made for evaluating the performance of a classification model, where N is the number of target classes. A confusion matrix compares the true condition with a predicted

condition made by the machine learning model. It supports scholars in having a comprehensive view of how well the classification model performs and what kinds of errors it makes. For a binary classification problem, a 2×2 matrix is shown as an example in Table 1. It is extremely useful for measuring recall, precision, specificity, accuracy, and an AUC-ROC curve. In Table 1, TP indicates a true positive which means the classifier predicts positivity and the prediction is true; TN indicates a true negative which means the classifier predicts negativity and the prediction is negative; FP indicates a false positive which means the classifier predicts positivity but the prediction is false; FN a indicates false negative which means the classifier predicts negativity but the prediction is true. The accuracy is computed as dividing the sum of the elements on the leading diagonal by the sum of all the elements in the confusion matrix which provides the fraction of correct responses.

Table 1. Confusion matrix of the classification result for a binary classification problem.

		Predicted condition	
		Positive	Negative
True condition	Positive	TP	FN
	Negative	FP	TN

As an integral part of the model development process, model evaluation is helpful to find a suitable model and reveal how well the chosen model works. In the experiments, cross-validation was adopted to evaluate the effectiveness of the models. Cross-validation splits a data set into k groups, making one group work as the testing set and the rest of the groups work as the training sets, and repeats this process k times until each group has been used as the testing data. It is a resampling procedure for an evaluation of machine learning models with limited data samples. It stands out in the area of classification and prediction because of the simplicity of interpretation. Moreover, it generally leads to a less biased or less optimistic estimation of the model effectiveness than other methods. Cross-validation is a well-established technique for experiments that is capable of being used to determine optimal values for a set of unknown model parameters. Additionally, in view of the fact that the general idea of cross-validation is to divide the data sample into a number of randomly drawn, disjoint subsamples, this methodology is often used to minimize the bias associated with random sampling of the training and testing samples in comparing the predictive accuracy of two or more methods. It is also named k number of folds. For each potential value of k, the model is used to make predictions of the k^{th} fold while using the $k-1$ folds as examples to learn from. This process of testing each fold against the remaining examples repeats k times. If k is equal to the sample size, it is called leave-one-out cross-validation (LOOCV) which allows the size of a training set to be the largest. The cross-validation estimates the overall accuracy of a model by averaging the k individual accuracies. In the experiments, all the dependence measures were transformed into a dissimilarity version, and a 10-fold cross-validation was adopted to evaluate the performance of distance-based discriminant analysis and SVM with different measures; then the best results were recorded. Accuracy was used as the criterion to evaluate the results, and the standard errors of the 10-fold cross-validation results were also recorded to examine the stability of the experimental results. The processing time for calculating the distance and correlation matrixes was noted to assess the computational cost.

SVM is a popular supervised machine learning approach. Supervised learning develops a predictive model based on both input and output data. With a set of inputs along with the corresponding correct outputs, supervised learning builds models to predict the values of labels of additional unlabeled data. The classification problems consist of taking input vectors and deciding

which classes they belong to, based on learning the instances of each class. In these experiments, the outputs of models are discrete, that means each example belongs to precisely one class, and the set of classes covers the entire possible output space.

The SVM maps the data onto a higher-dimensional input space and constructs an optimal separated hyperplane with the maximum margin (i.e., the maximum distance between data points of every pair of classes) in this space. In this way, future data points will be mapped onto the same space and the prediction shows their outputs based on which side of the gap the examples fall on. Furthermore, SVM with kernel trick is able to fully perform a nonlinear classification that will consequently map their inputs onto high-dimensional feature spaces.

The kernel in SVM is customized, and the Gaussian kernel is an example of a radial basis function, which is the most frequently used kernel function:

$$K(x_i, x_j) = e^{-\left(|x_i - x_j|^2/\sigma^2\right)} \quad (3)$$

Amari and Wu (1999) proposed a new method to modify a kernel based on an information-geometric consideration of the structure of the Riemannian geometry induced by the kernel[10]. Song *et al.* (2008) proposed a novel kernel using a convex combination of the characteristics of a polynomial kernel and a Gaussian radial basis function kernel[11]. Abe (2005) proposed a model selection method for Mahalanobis kernels calculating the covariance matrix, which replaced the Euclidean distance portion of a Gaussian radial basis function kernel with the Mahalanobis distance form, and the proposed method has been verified to have comparable performance with a Gaussian radial basis function kernel optimization[12]. Furthermore, Mu and Zhou (2009) suggested a kernel function that evaluated by replacing Euclidean distance in the Gaussian kernel with a more generalized Minkovsky's distance, which resulted in better prediction accuracy[13].

These studies suggest that the replacement of distance is a feasible method to improve the performance of classifiers. The kernelization of the MIC techniques may lead to some new possibilities in classification tasks. To evaluate the classification performance of different measures, Euclidean distance in a Gaussian radial basis function was replaced by the transformed dependence measures and the kernels were applied for SVM. Besides the original Gaussian radial basis function, the other forms of Gaussian radial basis function for experiments were as follows.

The Pearson correlation radial basis kernel function is:

$$K_{RBFCOR}(x, x') = \exp\left(\frac{r-1}{\sigma^2}\right) \quad (4)$$

Spearman's rank correlation radial basis kernel function is:

$$K_{RBFSPERMAN}(x, x') = \exp\left(\frac{\rho-1}{\sigma^2}\right) \quad (5)$$

The MIC radial basis kernel function is:

$$K_{RBFMIC}(x, x') = \exp\left(\frac{MIC-1}{\sigma^2}\right) \quad (6)$$

To tune the hyperparameter σ for each kernel, a grid search for three parameter levels at 0.01, 0.1 and 1.0 was adopted, which controls how hard or soft the margin is. Finally, the constant parameter was set as the default value 1. All the experiments were processed under R (version 3.6.1) using the package "kernlab" to process the SVM model with a C-svc classification type.

In Table 2 and Table 3, only the numbers of cases with the best experimental results are shown. The results with the same highest values were all counted.

According to the results of distance-based discriminant analysis in Table 2, the MIC fails to show superior performance in most cases. For the data group with less than 100 attributes, the MIC performs best on only one data set, while Pearson correlation (cosine similarity) and Spearman's rank correlation perform best on five data sets and three data sets, respectively. On the other hand, Euclidean distance performs best on ten data sets, whose classification accuracies were substantially better than the other measures. For the data group with 100 or more attributes, the MIC also performs best on only one data set, while Pearson

correlation (cosine similarity) performs best on six data sets and Spearman's rank correlation performs best on three data sets. Euclidean distance performs best on seven data sets. Even if the discussion was limited to the correlations, the MIC only performs better on one data set. In most cases, the accuracies of the MICs are the lowest and are far lower than the other measures.

Table 2. Summary of distance-based discriminant analysis with the best accuracy cases.

Metric for distance-based discriminant analysis		Number of cases with the best accuracy	
ED	Attributes<100	10	
	Attributes≥100	7	17
$r_{Pearson}$	Attributes<100	5	
	Attributes≥100	6	11
$\rho_{Spearman}$	Attributes<100	3	
	Attributes≥100	3	6
MIC	Attributes<100	1	
	Attributes≥100	1	2

According to the results of SVM in Table 3, the MIC still fails to show a superior performance. For both data groups, the MIC performs well on none of the data sets, while Pearson correlation (cosine similarity) and Spearman's rank correlation perform best on six data sets and five data sets, respectively. Taking the result of discriminant analysis above into account, the MIC performs worst. Euclidean distance performs best on nineteen data sets, almost accounting for two thirds of cases. Additionally, the accuracies of SVM with an original Gaussian radial basis function kernel using Euclidean distance were significantly higher than the previous experiments with Euclidean distance. The result indicates that the original Gaussian radial basis function is a reliable choice in most cases. Regarding the MIC, when compared with the other dependence measures, its merit still has not been clearly displayed. Due to the page limit, the tables do not provide detail results of each data set, but it is worthy to mention that when the discussion is limited to the classification accuracies of correlation measures, the MIC performs not worse than the other correlations on four data sets. Regarding the standard errors of accuracies, the results of kernelized the MICs are relatively stable compared with the other measures.

Table 3. Summary of SVM with the best accuracy cases.

Metric for SVM		Number of cases with the best accuracy	
ED	Attributes<100	15	
	Attributes≥100	4	19
$r_{Pearson}$	Attributes<100	2	
	Attributes≥100	4	6
$\rho_{Spearman}$	Attributes<100	1	
	Attributes≥100	4	5
MIC	Attributes<100	0	
	Attributes≥100	0	0

In regard to the computational costs, the time cost of the MIC is higher than other measures. Since the time costs of computation of Euclidean distance, Pearson correlation (cosine similarity) and Spearman's rank correlation are too short to precisely record, the time cost results of these three measures in the less than 100 attributes data set group and a few in the over 100 attributes group rounded to 0. On the other hand, the computation speed of the MIC was significantly slower than other measures in the experiments. The longest computation time cost reached more than 120,000 seconds, which is equivalent to about a day and a half. On the contrary, among the other measures, the most time consuming computation did not exceed 2 seconds. Furthermore, as the data size increases, the computational cost of the MIC increases exponentially. This indicates that computing the MIC is more inefficient for large data sets than the other conventional measures. This conclusion is consistent with the study

of Posnett *et al.* (2012), which pointed out the slow computation of the MIC[5]. Additionally, considering the classification accuracies of the MIC in the experiments above, the MIC displays even worse performance in the larger data sets.

4. Conclusion and Future Work

The MIC has been seen as important for its ability to measure non-linear relationships. This study compared the performance of the MIC with Euclidean distance, Pearson correlation (cosine similarity) and Spearman's rank correlation in classification tasks using 30 real-world data sets through discriminant analysis and SVM. Although some previous research suggests that the performance of the MIC is acceptable with sufficient data, according to the current results, the MIC fails to show superior performance in real-world classification tasks. The MIC also fails to meet the expectation of performing better on larger data sets. Owing to the high computational cost, the MIC does not seem to be a good choice for complex real-world cases. To sum up, the experiments suggest that the performance of the MIC for classification is not desirable enough for a wider usage to deal with real-world tasks, and before employing the MIC to deal with real-world classification tasks, analysts should cautiously use it as a prospective measure. At least in practice, it is currently not capable of being a measure that effectively replaces the conventional correlation coefficients.

According to the results of this study, it is hard to clarify the relationship between the characteristics of a data set and the performance of the MIC. However, some features were captured through our experiments, considering that the result of classification can be impacted by many aspects including data characteristics and the adopted method, which are discussed in the following paragraphs.

Regarding the real-world dependence analysis in previous research, the performance of the MIC was not as poor as this study found. The reason may be due to the definition of the MIC, that is, the MIC is a measure derived from mutual information, which is expert in some specific fields such as natural language processing, medical imagining and communication engineering. In these fields, the MIC is assumed to have better performance. Therefore, in the future it is necessary to discover any applicable fields of the MIC especially in mutual information's specialized range of applications such as text mining.

In addition, due to the rapid increase in the amount of data, the requirement for dimension reduction processing is increasing. Since the MIC is supposed to retain a broader correlation structure, it is hoped that in the future it can be applied to dimension reduction methods including principal component analysis as well as factor analysis.

However, the main purpose of dimension reduction is to improve computational efficiency, where there is another challenge: even if the MIC worked in dimension reduction efficiently, regarding the low computational efficiency of the MIC, the requirement of optimizing the computational complexity becomes urgent. Although the asymptotic algorithm of the MIC with better time-accuracy tradeoffs has been discussed to save computational costs[6], empirical research to support the rationality and universality of this algorithm is still in short supply. Moreover, no one certain criterion of the parameter settings for computing the MIC has been made up to now. Therefore, to further explore the possibility of the MIC in practical use, the first obstacle is to make a thorough inquiry into how to improve the computational efficiency of the MIC, and the second obstacle is to prove effectiveness through empirical research. To achieve this goal, it is suggested to enhance the MIC from two aspects: (1) optimizing the computational steps for the MIC in order to keep up with the demands from the corresponding fields; (2) looking for a proper criterion of parameter tuning. If the above challenges are able to be successfully achieved, it is possible that the MIC would become the undisputed correlation coefficient for the 21st century.

References

1) D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, P. C. Sabeti, "Detecting Novel Associations in Large Data Sets", *Science*, **334**[6062], 1518-1524 (2011).
2) T. Speed, "A Correlation for the 21st Century", *Science*, **334**[6062], 1502-1503 (2011).
3) C. Pang, A. Goel, S. Li, and M. Wilkins, "A Multidimensional Matrix for Systems Biology Research and its Application to Interaction Networks", *Journal of Proteome Research*, **11**[11], 5204-5220 (2012).
4) R. S. de Souza, U. Maio, V. Biff, and B. Ciardi, "Robust PCA and MIC Statistics of Baryons in Early Mini-haloes", *Monthly Notices of the Royal Astronomical Society*, **440**[1], 240-248 (2014).
5) D. Posnett, P. Devanbu, and V. Filkov, "MIC Check: A Correlation Tactic for ESE Data", *In 2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*, 22-31 (2012).
6) D. Reshef, Y. Reshef, M. Mitzenmacher, and P. Sabeti, "Equitability Analysis of the Maximal Information Coefficient, with Comparisons", *arXiv:1308.6009* (2013).
7) "UCI Machine Learning Repository", *Center for Machine Learning and Intelligent Systems*, http://archive.ics.uci.edu (20210820)
8) J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework", *Journal of Multiple-Valued Logic and Soft Computing*, **17**[2-3], 255-287 (2011).
9) J. Vanschoren, J. N.V. Rijn, B. Bischl, and L. Torgo, "OpenML: Networked Science in Machine Learning", *ACM SIGKDD Explorations Newsletter,* **15**[2], 49-60 (2014).
10) S. I. Amari, and S. Wu, "Improving Support Vector Machine Classifiers by Modifying Kernel Functions", *Neural Networks*, **12**[6], 783-789 (1999).
11) H. Song, Z. Ding, C. Guo, Z. Li, and H. Xia, "Research on Combination Kernel Function of Support Vector Machine", *International Conference on Computer Science and Software Engineering, IEEE*, **1**, 838-841 (2008).
12) S. Abe, "Training of Support Vector Machines with Mahalanobis Kernels", *International Conference on Artificial Neural Networks*, 571-576 (2005).
13) X. Mu, and Y. Zhou, "A Novel Gaussian Kernel Function for Minimax Probability Machine", *2009 WRI Global Congress on Intelligent Systems, IEEE*, **3**, 491-494 (2009).