Comments on Supervised Feature Selection by Clustering Using Conditional Mutual Information-based Distances

Nguyen X. Vinh[a,*], James Bailey[b],

[a] *The University of New South Wales, Sydney, Australia*
[b] *The University of Melbourne, Melbourne, Australia*

Abstract

Supervised feature selection is an important problem in pattern recognition. Of the many methods introduced, those based on the mutual information and conditional mutual information measures are among the most widely adopted approaches. In this article, we re-analyze an interesting paper on this topic recently published by Sotoca and Pla (Pattern Recognition, Vol. 43 Issue 6, June, 2010, p. 2068-2081). In that work, a method for supervised feature selection based on clustering the features into groups is proposed, using a conditional mutual information based distance measure. The clustering procedure minimizes the objective function named the *minimal relevant redundancy*—mRR criterion. It is proposed that this objective function is the upper bound of the information loss when the full set of features is replaced by a smaller subset. We have found that their proof for this proposition is based on certain erroneous assumptions, and that the proposition itself is not true in general. In order to remedy the reported work, we characterize the specific conditions under which the assumptions used in the proof, and hence the proposition, hold true. It is our finding that there is a reasonable condition, namely when all features are independent given the class variable (as assumed by the popular naive Bayes classifier), under which the assumptions as required by Sotoca and Pla's framework hold true.

Keywords: Feature selection, conditional mutual information, mutual

*Corresponding author
Email addresses: n.x.vinh@unsw.edu.au (Nguyen X. Vinh), baileyj@unimelb.edu.au (James Bailey)

information properties, clustering, classification, naive Bayes classifier.

1. Introduction

Feature selection plays an important role in building efficient pattern recognition systems. When done properly, feature selection can greatly improve prediction accuracy whilst reducing computational cost, as well as potentially providing insights into the underlying data generating process. In this paper, we comment on a recent work by Sotoca and Pla [1], which proposes an approach for feature selection using mutual information. We show that the two key results from this work do not hold true in general. In order to remedy these issues, we investigate whether there exist special conditions under which the reported theoretical results hold true. We discuss several such conditions in this paper.

1.1. Background

The feature selection method proposed by Sotoca and Pla [1] falls into the *filter paradigm*, for which one needs to specify a measure of dependency between the features and the class variable. This dependency measure is then used to rank the feature subsets, and choose the one most relevant to the class variable. Of the many dependency measures available, those based on information theoretic concepts are very popular, not only for the feature selection problem [2, 1, 3], but across many topics in pattern recognition and data mining [4]. This popularity can be explained due to the strong theoretical foundation that exists within information theory. Given two random variables X and Y with domains $\mathcal{X}$ and $\mathcal{Y}$ respectively, their mutual information (MI) is defined as:

$$I(X;Y) = \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p_{XY}(x,y) \log \left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \right). \tag{1}$$

The MI measures the information shared between random variables, and is general enough to detect any kind of dependency, be it linear or non-linear. Also of interest is the conditional mutual information (CMI) between X and Y given another random variable Z, defined as:

$$I(X;Y|Z) = \sum_{z\in\mathcal{Z}} p_Z(z) \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p_{X,Y|Z}(x,y|z) \log \frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}. \tag{2}$$

The CMI can be interpreted as the information that X can predict about Y (and vice-versa) that Z cannot. We refer interested readers to [5] for comprehensive discussion on these measures and their basic properties.

The task of feature selection using the MI can be described as follows. Suppose we have an input dataset with N samples, M features $\mathbf{X} = \{X_1, \ldots, X_M\}$, and a target classification variable C. The goal of feature selection is to select the optimal feature subset $\widetilde{\mathbf{X}} = \{\widetilde{X}_1, \ldots, \widetilde{X}_m\}$ of m (normally $\ll M$) features that shares the maximal mutual information with C:

$$\widetilde{\mathbf{X}}^* = \arg\max_{\widetilde{\mathbf{X}} \subset \mathbf{X}} I(\widetilde{\mathbf{X}}; C). \tag{3}$$

It is difficult to estimate the high dimensional MI, since in practice we only have a limited number of samples. Therefore, many works have approximated (3) with lower order MI. An example is the well known minimum Redundancy Maximum Relevance (mRMR) criterion [2], which maximizes the pairwise MI between the features and the class variable, i.e., relevancy, while penalizing the pairwise MI between the features, i.e., redundancy:

$$\widetilde{\mathbf{X}}^* = \arg\max_{\widetilde{\mathbf{X}} \subset \mathbf{X}} \left(\frac{1}{|\widetilde{\mathbf{X}}|} \sum_{\widetilde{X}_i \in \widetilde{\mathbf{X}}} I(\widetilde{X}_i; C) - \frac{1}{|\widetilde{\mathbf{X}}|^2} \sum_{\widetilde{X}_i, \widetilde{X}_j \in \widetilde{\mathbf{X}}} I(\widetilde{X}_i; \widetilde{X}_j) \right). \tag{4}$$

1.2. The minimal relevant redundancy (mRR) criterion

In this section, we briefly review the *minimal relevant redundancy* (mRR) criterion introduced by Sotoca and Pla [1]. They observe that maximizing $I(\widetilde{\mathbf{X}}; C)$ is in fact also equivalent to minimizing the information loss about C, i.e., $I(\mathbf{X}; C) - I(\widetilde{\mathbf{X}}; C) = I(\mathbf{X}; C|\widetilde{\mathbf{X}})$, when the full set of features $\mathbf{X}$ is replaced by a subset $\widetilde{\mathbf{X}}$. This is clear, since we have $I(\mathbf{X}; C) = \textit{constant}$ for a fixed data set. Thus, the aim is to solve the following optimization problem:

$$\widetilde{\mathbf{X}}^* = \arg\min_{\widetilde{\mathbf{X}} \subset \mathbf{X}} I(\mathbf{X}; C) - I(\widetilde{\mathbf{X}}; C) = \arg\min_{\widetilde{\mathbf{X}} \subset \mathbf{X}} I(\mathbf{X}; C|\widetilde{\mathbf{X}}). \tag{5}$$

However it can still be difficult to estimate the high order CMI $I(\mathbf{X}; C|\widetilde{\mathbf{X}})$, and also it is not clear as what optimization procedure can be employed to solve (5). Sotoca and Pla's subsequent development relies on the following key result:

Proposition 1. *[Sotoca and Pla [1]] Let* $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_m)$ *be a subset of* m *random variables from the original set of* M *random variables* $\boldsymbol{X} = (X_1, \ldots, X_M)$*, that is,* $\widetilde{\boldsymbol{X}} \subset \boldsymbol{X}$*, then, the decrease of mutual information of the original and the reduced set with respect to a relevant variable* C *is upper bounded by:*

$$I(\boldsymbol{X}; C) - I(\widetilde{\boldsymbol{X}}; C) = I(\boldsymbol{X}; C|\widetilde{\boldsymbol{X}}) \leq \frac{1}{m} \sum_{i=1}^{M} \sum_{j=1}^{m} I(X_i; C|\widetilde{X}_j) \tag{6}$$

This looks to be an attractive result, since it allows one to replace the loss function with its upperbound, which is based solely on low order CMIs. More specifically, this result offers a theoretical basis for replacing the single high order CMI term with a sum of triple-wise CMIs, which are much more amenable to numerical estimation with limited sample size. Also, Sotoca and Pla note that this bound resembles the objective of a K-means clustering process, i.e.,

$$F(\widetilde{X}_1, \ldots, \widetilde{X}_m) = \frac{1}{m} \sum_{i=1}^{M} \sum_{j=1}^{m} I(X_i; C|\widetilde{X}_j) = \sum_{i=1}^{M} \sum_{j=1}^{m} p(\widetilde{X}_j|X_i) I(X_i; C|\widetilde{X}_j) \tag{7}$$

where the conditional posteriors $p(\widetilde{X}_j|X_i) = 1/m$ correspond to a uniform distribution, instead of the delta distribution (i.e., $p(\widetilde{X}_j|X_i) = 1$ if X_i is in the cluster centered by $\widetilde{X}_j$ and 0 otherwise) as in the usual K-means algorithm. This observation suggests that the features can be clustered into groups, then one or a small number of representative features from each group may be chosen to form the selected feature subset. The procedure looks to be both theoretically and practically appealing, since it addresses both feature diversity (via the clustering process—the features in different clusters are deemed to be as different as possible) and joint-optimality, via the bound in (6), while admitting low sample complexity (the use of low order CMI).

Clustering the features using a K-means like procedure does encounter some practical difficulties, as it is not clear as how to define the mean feature (centroid) for a cluster. Sotoca and Pla therefore instead use a hierarchical clustering approach. For that purpose, they propose a conditional mutual information based distance, which they argue is a true metric in the feature space:

Proposition 2. *[Sotoca and Pla [1]] The following conditional mutual information distance:*

$$D(X_i; \widetilde{X}_j) = I(X_i; C|\widetilde{X}_j) + I(\widetilde{X}_j; C|X_i) \tag{8}$$

satisfies the properties of a true metric, i.e., non-negativity, identity of indiscernibles, symmetry, and triangle inequality.

This also looks to be an appealing result. A true metric not only conforms well with one's intuition about distance, but also, working in a true metric space can potentially provide important theoretical and algorithmic advantages, since many useful theoretical results and efficient algorithms already exist for metric spaces.

2. Theoretical problems with the mRR framework

Although the mRR framework appears appealing, we have found that it contains some theoretical shortcomings. In this section, we reanalyze the development of the two key results in the mRR framework, namely Proposition 1 and 2, and point out the gaps in the analysis.

2.1. Erroneous assumptions made in Proposition 1 proof

The proof offered for this proposition relies on the following properties of the mutual information and condition mutual information where it is assumed:

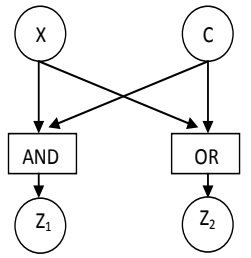
Assumption 1. *Conditioning on a third feature always reduces the mutual information, i.e.,*

$$I(X; C|Z) \leq I(X; C) \tag{9}$$

Assumption 2. *Increasing the conditioning set always reduces the conditional mutual information, i.e.,*

$$I(X, C|Z_1, \ldots, Z_k) \leq I(X; C|Z_i), \quad \forall i = 1, \ldots, k \tag{10}$$

Unfortunately, we will show here, via a simple counter example, that Assumptions 1 and 2 do not hold true in general. Let X and C be two independent random binary variables, each with equal probability for the '0' and '1' states. Let $Z_1 = X \wedge C$ and $Z_2 = X \vee C$. Thus Z_1 and Z_2 are also two other random variables. From these definitions, it is straightforward to write



P(X,C)

X\C	0	1
0	0.25	0.25
1	0.25	0.25

P(XC,Z1)

XC\Z1	0	1
00	0.25	0
01	0.25	0
10	0.25	0
11	0	0.25

P(XC,Z1Z2)

XC\ Z1Z2	00	01	10	11
00	0.25	0	0	0
01	0	0.25	0	0
10	0	0.25	0	0
11	0	0	0	0.25

P(XC,Z2)

XC\Z2	0	1
00	0.25	0
01	0	0.25
10	0	0.25
11	0	0.25

P(XC|Z1Z2)

XC\ Z1Z2	00	01	10	11
00	1	0	0	0
01	0	0.5	0	0
10	0	0.5	0	0
11	0	0	0	1

Figure 1: An example where Assumptions 1 and 2 fail to hold

down the joint and conditional probability distributions $P(X, C)$, $P(XC, Z_1)$, $P(XC, Z_2)$, $P(XC, Z_1Z_2)$ and $P(XC|Z_1Z_2)$ as in Fig. 1. From these distributions, it is easy to verify that Assumption 1 is violated here. More specifically, we have that $I(X; C|Z_1) = I(X; C|Z_2) = 0.187$ bit $> I(X; C) = 0$, that is, conditioning on a third variable *can increase* the MI. In this example, X and C are marginally independent, but become dependent conditioning on any variable Z_1 or Z_2. Within the Bayesian network literature, this effect is termed "*explaining away*". It is also easy to show that Assumption 2 also does not hold true in this example, as $I(X; C|Z_1, Z_2) = 0.5$ bit $> I(X; C|Z_1) = 0.187$ bit. Thus, given both Z_1 and Z_2, then X and C share significantly more information than if only Z_1 or Z_2 was given individually.

2.2. Proposition 1 does not hold true in general

We have shown that the proof for Proposition 1 is erroneous. Does there exist a correct proof for it? Unfortunately, such a proof does not exist, as

Proposition 1 itself does not hold true in general. Indeed, using the above example, if we take $\mathbf{X} = (X, Z_1, Z_2)$, and $\widetilde{\mathbf{X}} = (Z_1, Z_2)$, then according to Proposition 1 we must have:

$$I(\mathbf{X}; C|\widetilde{\mathbf{X}}) = I(X; C|Z_1, Z_2) \leq \frac{1}{2}\{I(X; C|Z_1) + I(X; C|Z_2)\} \tag{11}$$

but this is not true, given that in the above example, $I(X; C|Z_1, Z_2)$ is larger than $I(X; C|Z_1) + I(X; C|Z_2)$.

2.3. Erroneous proof for Proposition 2

The proof of the metricity property of the CMI distance is also erroneous. The following derivation step was used in the proof (for the triangle inequality property):

$$I(X_i; C|X_j) + I(X_j, C|X_k) \geq I(X_i; C|X_j, X_k) + I(X_j; C|X_k) \tag{12}$$

which is in fact equivalent to the erroneous Assumption 2.

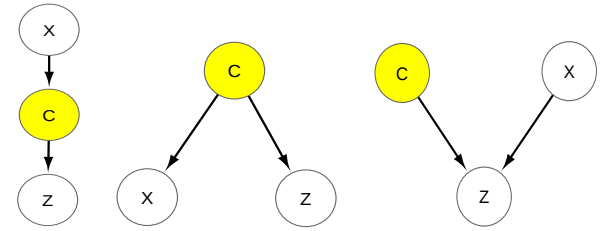
3. Conditions under which mRR is applicable

Since Sotoca and Pla's mRR framework contains several interesting and useful ideas, in this section we investigate the conditions under which Assumptions 1 and 2, and hence their two key Propositions 1 and 2, hold true. More importantly, we assess whether such conditions, if they exist, are reasonable and plausible in pattern recognition applications. We have identified [6], in which Renner and Maurer characterize a necessary and sufficient condition for Assumption 1 to hold, while an anonymous reviewer pointed us to [5, Theorem 2.8.1], where a sufficient condition for Assumption 1 is given. We discuss these conditions below:

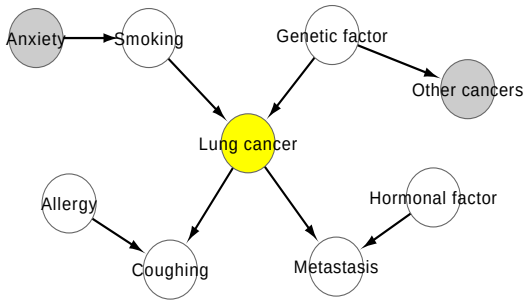
- From a communication theory point of view, Renner and Maurer [6] report a necessary and sufficient condition for Assumption 1 as follows. Suppose X and C are two random variables, and Z is an output variable from a communication channel that takes X and C as inputs, then a necessary and sufficient condition for Assumption 1 to hold, i.e., $I(X; C) \geq I(X, C|Z)$, is that the conditional distribution $P(Z|X, C)$ that characterizes the channel can be decomposed as $R(X, Z) \cdot S(Z, C)$ where the two functions R and S depend only on (Z, X) and (Z, C) respectively.

- In [5], it is proven that a sufficient condition for Assumption 1 is that X, C and Z form a Markov chain $X \rightarrow C \rightarrow Z$, where Z is conditionally independent of X given C, i.e., $P(Z, X|C) = P(X|C)P(Z|C)$.

For ease of analysis, we next adopt the graphical notation of Bayesian network (BN) [7]. A BN is defined by a graphical structure and a family of probabilistic distribution, which together allow efficient and accurate representation of the joint probability distributions over a set of random variables (RVs) of interest. The graphical part of a static BN is a directed acyclic graph (DAG), with nodes representing RVs and edges representing their conditional (in)dependence relations. In a BN, a node is conditionally independent of all its non-descendants, given its parents. In addition, it is often assumed that parent-child in a BN admits a direct cause-consequence relationship.



(a) Markov chain 1 (b) Markov chain 2 (c) Communication channel



(d) An example real Bayesian network, adapted from [8]

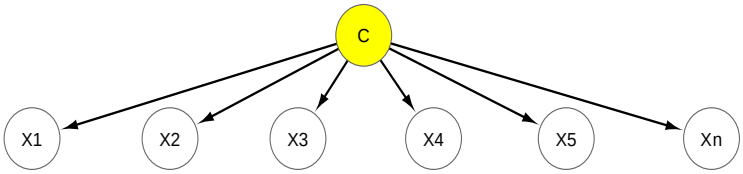
Figure 2: Bayesian network depiction for conditions where Assumption 1 holds.

The two possible network configurations for the Markov chain conditions are presented in Fig. 2(a,b), while the network configuration for the 'communication channel' scenario is presented in Fig. 2(c). The Markov chain

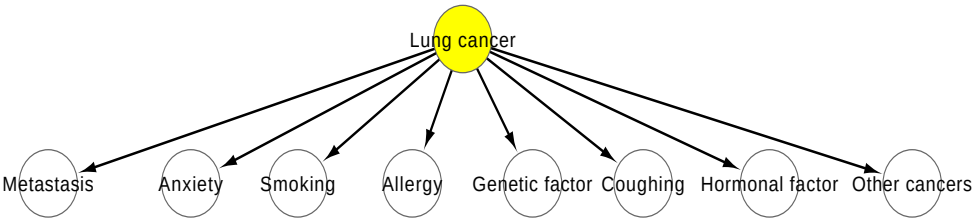
condition requires a feature to be parent while another be child of C, or both features to be child of C. In order to assess the feasibility of this scenario, we present an example real Bayesian network for lung cancer, adapted from [8]. Here, the target class variable C is lung cancer. It can be seen that for this network, there are triplets that form a Markov chain (e.g., Smoking $\rightarrow$ Lung cancer $\rightarrow$ Coughing), and triplets that do not form a Markov chain (e.g., Lung cancer, Coughing, and Allergy). Thus in general, the Markov chain condition does not hold. The 'communication channel' scenario requires a feature to be child of C, another feature to be the co-parent with C, and that each parent acts independently on the child, i.e., $P(Z|X,C) = R(X,Z) \cdot S(Z,C)$. Again from Fig. 2c, it is seen that not all triplets conform to this structure, e.g., {Smoking, Lung cancer, Genetic factor}, let alone the condition that each parent acts independently on the child. Thus, it can be argued that the two conditions under which Assumption 1 holds that we have investigated are not likely to always be reasonable in general pattern recognition applications.

Until now, we have just investigated the conditions under which Assumption 1 holds. We are yet to find out the requirements for Assumption 2 to hold (which might be different from those for Assumption 1). As we do not know how many such conditions exist, we therefore change our approach. Instead of asking what are the conditions under which Assumptions 1 and 2 hold, then checking to see whether such conditions are plausible for real pattern recognition applications, we now ask: on which commonly applied Bayesian network structures Assumption 1 and 2 both hold true. Our obvious target is the naive Bayesian network structure, as assumed by the naive Bayesian network classifier, in which the features are assumed to be conditionally independent given the class variable, i.e., $I(X_i, X_j|C) = 0, \forall i, j$ as in Fig. 3a. The naive Bayes network structure for the lung cancer problem is illustrated in Fig. 3b. Note that in this network structure, every triplet $\{X_i, C, X_j\}$ forms a Markov chain $X_i \leftrightarrow C \leftrightarrow X_j$. The naive Bayes network structure is actually an extension of the Markov condition in Fig. 2(b).

We should clarify here that the naive Bayes network is not a realistic Bayesian network that reflects the actual relationships between variables, but rather it is an assumption made by a learning algorithm, in this case the naive Bayes classifier, in order to simplify the learning process. Albeit this simplistic assumption, naive Bayes classifiers have been reported to perform remarkably well, on par with state of the art modern classifiers on many learning problems [9]. Text classification is a particular example where naive Bayes classifiers were very successful [10]. There have been numerous empir-



(a) Naive Bayes network structure



(b) An example naive Bayes network

Figure 3: Naive Bayes network

ical works showing that the naive Bayes classifier predicts equally well as the decision tree algorithm C4.5, as well as several theoretical investigations that try to explain the surprisingly good performance of naive Bayes [11]. For this network structure, Assumption 1 holds true as per the Markov chain condition. In the following, we prove that Assumption 2 also holds true on the naive Bayes network structure.

Lemma 1. *Given* $I(X; Z_i|C) = 0, \forall i$ *we have:*

$$I(X; C|Z_1, \ldots, Z_k) \leq I(X; C|Z_i), \forall i$$

Proof. Let $\mathbf{Z} = \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_k\}$, we prove $I(X; C|Z_i, \mathbf{Z}) \leq I(X; C|Z_i)$. We have $I(X; C, Z_i, \mathbf{Z})$ admitting the following decomposition:

$$I(X; C, Z_i, \mathbf{Z}) = I(X; Z_i) + I(X, C|Z_i) + I(X, \mathbf{Z}|C, Z_i)$$

We now prove that $I(X; \mathbf{Z}|CZ_i) = 0$ as follows:

$$\begin{aligned} I(X; \mathbf{Z}|C, Z_i) &= I(X; Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_k|C, Z_i) \\ &= I(X; Z_1|C, Z_i, Z_1) + \ldots + I(X; Z_k|C, Z_i, Z_1, \ldots, Z_{k-1}) \end{aligned}$$

It can be proven that $I(X; Z_i|C, \widetilde{\mathbf{Z}}) = 0, \forall i$ where $\widetilde{\mathbf{Z}}$ is an arbitrary feature set [12]. Thus $I(X; \mathbf{Z}|C, Z_i) = 0$, and so

$$I(X; C, Z_i, \mathbf{Z}) = I(X; Z_i) + I(X, C|Z_i) \tag{13}$$

On the other hand, $I(X; C, Z_i, \mathbf{Z})$ can also be decomposed as:

$$I(X; C, Z_i, \mathbf{Z}) = I(X; Z_i, \mathbf{Z}) + I(X; C|Z_i, \mathbf{Z}) \tag{14}$$

We also have

$$I(X; Z_i, \mathbf{Z}) = I(X; Z_i) + I(X; Z_i|\mathbf{Z}) \geq I(X; Z_i) \tag{15}$$

From (13)-(15) we have $I(X; C|Z_i, \mathbf{Z}) \leq I(X; C|Z_i)$. □

4. Discussion and conclusion

One of the best known principles of information theory states that conditioning reduces entropy, i.e., $H(X|Y) < H(X)$ [5]. Unfortunately, this principle does not carry over to the mutual information. Conditioning can either increase or decrease the mutual information. We note that Sotoca and Pla are not the only one to have made this assumption for mutual information. In a recent work [13], Guo and Nixon, in their mutual information based feature selection work for gait recognition, asserted: "It is known that given variables A, B and C, $I(A, B|C) \leq I(A, B)$". Yet in both [1] and [13], the authors reported reasonably good feature selection and classification results with their proposed methods.

In the current paper, we have investigated the conditions under which the assumptions that *conditioning reduces mutual information* and *increasing the conditioning set reduces mutual information*, as required by Sotoca and Pla's framework, hold true. It is our finding that under the condition of features conditionally independent given the class variable, as assumed by the popular naive Bayes classifier, then these assumptions both hold true. The conditionally-independent-features assumption is clearly a naive one, yet naive Bayes classifiers have been reported to perform well on a wide range of classification problems, and remain popular today, where they often serve as baseline for evaluating more sophisticated classifiers. It is our expectation that feature selection procedures, such as the ones proposed by Sotoca and Pla [1] and Guo and Nixon [13], will perform well in problems where the naive Bayes classifier also delivers good performance. It is likely that in these problems, the conditionally-independent-features assumption hold true, either exactly or approximately. Under this condition, the framework as proposed by Sotoca and Pla is appealing, as both feature diversity (via

clustering) and joint-optimality (via the information loss bound) can be ensured, while admitting low sample complexity at the same time.

On classification problems where the conditionally-independent-features assumption does not hold true, we have found that Proposition 1 will fail to hold true in general, and hence the clustering process as proposed by Sotoca and Pla can no longer be interpreted as minimizing the information loss when the full set of features is replaced with a smaller subset. Thus, while we can still perform feature clustering, the resulting chosen features may only provide a good coverage of the feature space, but the joint quality of these selected features, measured in terms of the joint MI, is not assured. Nevertheless, in this case, feature diversity achieved via the feature clustering process might still be useful. In this respect, Proposition 2 would be an interesting result, as working in a proper metric space can provide certain theoretical and algorithmic advantages. Unfortunately, the current proof, based on Assumption 2 which does not hold true in general, is erroneous. As we haven't been able to prove or disprove this proposition in general, we thus leave it it as an open problem for interested readers:

Open problem 1. *Prove or disprove the metricity properties of the CMI distance* $D(X_i; \widetilde{X}_j) = I(X_i; C|\widetilde{X}_j) + I(\widetilde{X}_j; C|X_i)$.

References

[1] J. Martínez Sotoca, F. Pla, Supervised feature selection by clustering using conditional mutual information-based distances, Pattern Recognition 43 (2010) 2068–2081.

[2] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (2005) 1226–1238.

[3] P. L. Carmona, J. M. Sotoca, F. Pla, F. K. H. Phoa, J. B. Dias, Feature selection in regression tasks using conditional mutual information., in: Proceedings of the 5th Iberian conference on Pattern recognition and image analysis, pp. 224–231.

[4] Y. Y. Yao, Information-theoretic measures for knowledge discovery and data mining, in: Entropy Measures, Maximum Entropy Principle and Emerging Applications, Karmeshu (ed.), Springer, 2003, pp. 115–136.

[5] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley-Interscience, 2nd edition, 2006.

[6] R. Renner, U. Maurer, About the mutual (conditional) information, in: Information Theory, 2002. Proceedings. 2002 IEEE International Symposium on, p. 364.

[7] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, The MIT Press, 2009.

[8] I. Guyon, C. Aliferis, A. Elisseeff, Computational Methods of Feature Selection, chapter Causal Feature Selection, Chapman and Hall, 2007.

[9] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (1997) 131–163.

[10] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification, in: in AAAI-98 workshop on learning for text categorization, AAAI Press, 1998, pp. 41–48.

[11] H. Zhang, The optimality of naive bayes, in: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA.

[12] G. Van Dijck, M. M. Van Hulle, Increasing and decreasing returns and losses in mutual information feature subset selection, Entropy 12 (2010) 2144–2170.

[13] B. Guo, M. S. Nixon, Gait feature subset selection by mutual information, Trans. Sys. Man Cyber. Part A 39 (2009) 36–46.