Mutual Information and Conditional Independent Testing for Causal Feature Selection

RAKKRIT DUANGSOITHONG†1 ZHAO YUYING†1

Abstract: Causal feature selection algorithms can discover causal relationship; however, the redundant features are difficult to defined in the causal graph. To overcome redundancy analysis in this causal feature selection problem, this paper proposes mutual information and conditional independent testing for causal feature selection algorithm (MICI). According to the results, MICI can remove both irrelevant and redundant features while also discover the causality graph. The average accuracy of MICI slightly improves compared to original data and other feature selection methods.

Keywords: Causal feature selection, redundant features, dimensionality reduction.

1. Introduction

High-dimensional data suffers from curse of dimensionality and overfitting problems. Traditional classification methods have poor classification or recognition performance [1]. Normally, there are two methods to reduce the dimension: feature extraction and feature selection. Feature extraction transforms the original data into new dimensions through the data projection while the feature selection algorithm selects a subset of features by removing irrelevant features and redundant features to reduce the complexity and improve the classification accuracy of the classifier [2].

Recently, the study of causality has become more active in the research fields. The research on causal discovery algorithm aims to reveal the causal relationship between feature and feature and between feature and class. Basically, the causality of the dataset can be discovered by using Bayesian networks (BNs). It uses directed acyclic graphs (DAG) and conditional probability distribution to reveal the causal graph which can be defined from the relation between DAG structure and condition independencies [3].

This paper presents MICI algorithm that uses BNs to reveal the causal relationship in the graph and remove both irrelevant and redundant features by using mutual information and conditional independent testing, respectively.

2. Proposed MICI algorithm

The proposed MICI algorithm consists of two parts: relevance analysis and redundancy analysis. MICI first uses the mutual information (MI) between features and classes to rank the feature and remove the irrelevant features that have the MI less than the threshold value. In relevancy analysis, the conditional independent testing (CI Testing) is used to identified and remove the redundant features. The schematic diagram of the proposed MICI algorithm is shown in figure 1.

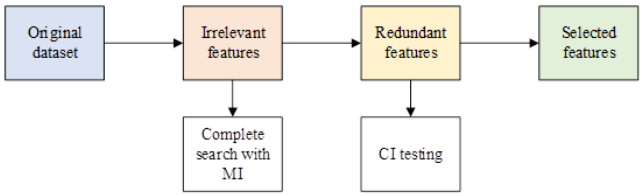

Fig. 1 Diagram of the proposed MICI Algorithm

2.1 Relevance analysis

In this research, Mutual information (MI) is used to measure feature relevance [4]. The formula of MI is as follows:

Given F, *S* and *C* represent the original feature set, selected feature subset and class, respectively. The set of classes $C = \{c_1, c_2, \dots, c_m\}$. The uncertainty in *C* or the entropy $H(C)$ is shown in equation 1.

$$H(C) = -\sum_{i=1}^{m} P(c_i) log P(c_i) \qquad (1)$$

where $P(c_i)$, $i = 1, \dots, m$, is the probability for the classes. Given a feature F and class C, $H(C \mid F)$ is the average uncertainty in *C* as shown in equation 2.

$$H(C \mid F) = -\sum_{f \in F} P(f) \sum_{c \in C} P(c|f) log P(c|f) \qquad (2)$$

where *P(f)* represents the probability for individual features in *F*, and *P(c|f)* is the conditional probability for class *c* given feature *f*.

The mutual information (MI) between feature and class is

$$MI(C;F) = MI(F;C) = H(C) - H(C \mid F) \qquad (3)$$

†1 Department of Electrical Engineering, Faculty of Engineering, Prince of Songkla University, Thailand, Songkhla, 90110, Thailand

2.2 Redundancy analysis

Generally, it is difficult to define the redundant in the causal graph. In this research, we proposed the algorithm to define the redundant features by using the conditional independent testing as follows.

Proposed definition of redundant feature: Two features F_i *and* F_j are the features from class C.

Given subset U, feature F_i is a redundant feature with feature F_j to the class C, if and only if

$MI(F_j, C|U) = MI(\mathrm{F}_i, C|U)$and

$CI(F_i, C|U)$or $CI(\mathrm{F}_j, C|U)$ is not independent.

Conditional independent Testing (CI Testing): Let F_i, $F_j$are the features of the dataset, U is the subset features. Given U, do a CI Testing between F_i *and* F_j. If $F_i \perp F_j \mid U$, it means under the condition of U, F_i and F_j are independent. Therefore, under the proposed definition of redundant feature, given the value of U, there is no additional information about F_i, (or F_j,) by the value of F_j, (or F_i).

3. Experiment

In this research, the proposed MICI algorithm is compared with non-causal and causal feature selection algorithms. The non-causal algorithms used in the experiment are fast correlation-based filter (FCBF) and ReliefF algorithms and compared with four causal feature selection algorithms: MMPC, IAMB, FBED, and MMMB algorithms. The average accuracy of 5 well-known classifiers with 10 datasets used in experiment collected from UCI machine learning repository and OpenML [5-6] are evaluated. The comparison of the complexity in terms of Big O notation will also be considered in the experiment.

4. Results

The average percent of classification accuracy for each classifier and feature selection are shown in table 1.

Table 1: The average accuracy from 10 datasets

Algorithm	kNN	NB	DT	SVM	MLP	Average
Original	77.38	73.28	**79.63**	77.40	**84.42**	78.42
FCBF	77.69	75.55	75.30	77.75	83.17	77.89
ReliefF	77.77	74.88	78.58	77.77	82.31	78.26
MMPC	76.48	76.52	76.41	76.50	82.34	77.65
IAMB	77.82	75.54	76.58	77.78	84.25	78.39
FBED	76.14	75.47	74.62	76.18	84.03	77.29
MMMB	77.69	74.93	73.60	77.68	80.65	76.91
Proposed MICI	**79.85**	**77.88**	77.05	**79.82**	83.05	**79.53**
Average	77.60	75.51	76.47	77.61	**83.03**	

Comparison between feature selection algorithms, the proposed MICI has slightly higher accuracy than using original features and other feature selection methods. The feature selection does not perform better than original feature when using DT and MLP. MLP provides the highest average classification accuracy compared with other classifiers. Figure 2 presents the example of causal graph from LUCAS dataset using proposed MICI algorithm. The output causal graph is compared with Ground truth graph of the dataset.

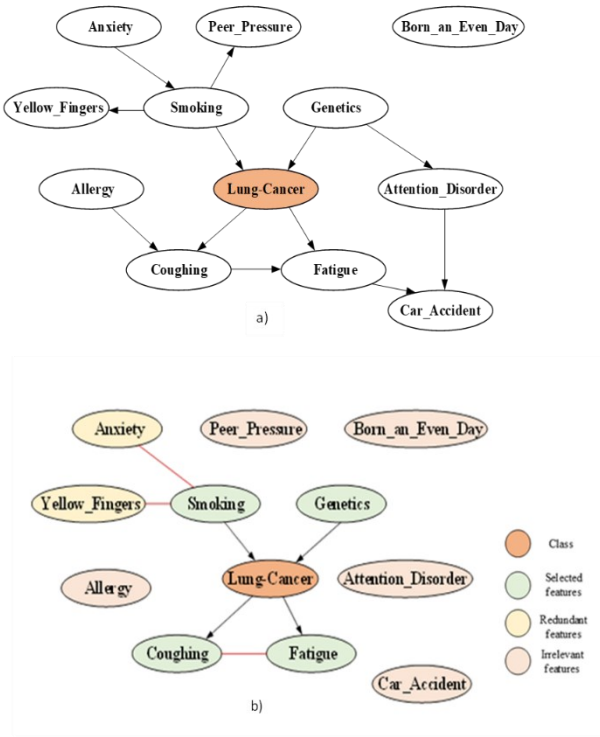

Fig.2: (a) Ground truth graph of Lucas dataset; (b) Causal graph of Lucas dataset using proposed MICI algorithm.

Reference

[1] Liu, H. and Motoda, H. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, USA, 1998.

[2] Yu, L and Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03), pp. 856-863, 2003.

[3] Guyon, I., Aliferis, I. and Elisseeff, C. Causal Feature Selection. In Computational Methods of Feature Selection, Liu, H. and Motoda, H. editors. Chapman and Hall, 2007.

[4] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. *Feature selection: A data perspective*. ACM Computing Surveys, 50(6), 2017.

[5] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2019.

[6] Vanschoren, J., N. van Rijn, J., Bischl, B. and Torgo, L. OpenML: networked science in machine learning. SIGKDD Explorations 15(2), pp 49-60, 2013.

Acknowledgments This research was supported by Thailand Science Research and Innovation (TSRI). The fundamental fund 2564, ref: ENG6405014S.